MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН

KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ УНИВЕРСИТЕТІ — INTERNATIONAL UNIVERSITY — SINCE 2009 — МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Published since 2020.
Volume 7. 1 (25). 2026
January–March

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

2020 жылдан бері шығарылады
Том 7. 1 (25). 2026
Қаңтар-Наурыз

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Издается с 2020 г.
Том 7. 1 (25). 2026
Январь-Март

Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан № KZ82VPY00020475, выданное от 20.02.2020 г.

Зарегистрировано в Международном центре регистрации серийных изданий ISSN (ЮНЕСКО, Париж, Франция). ISSN 2708–2032 (print), ISSN 2708–2040 (online)

Журнал входит в Перечень научных изданий, рекомендуемых КОКНВО МНВО РК для публикации основных результатов научной деятельности.

EDITOR-IN-CHIEF:
Kateryna Kolesnikova — Doctor of Technical Sciences, professor, Vice-Rector for Research, International Information Technology University (Kazakhstan)

DEPUTY EDITOR-IN-CHIEF:
Madina Ipalakova — Candidate of Technical Sciences, associate professor, Director of the Research Department, International Information Technology University (Kazakhstan)

EDITORIAL BOARD:
Abdul Razak — PhD, professor, Department of Cybersecurity, International Information Technology University (Kazakhstan)
Lucio Tommaso De Paolis — Director of the R&D Department of the AVR Laboratory, Department of Engineering for Innovation, University of Salento (Italy)
Liz Bacon — Professor, Deputy Vice-Chancellor, Abertay University (United Kingdom)
Michele Pagano — PhD, Professor, University of Pisa (Italy)
Mukhtarbay Otelbayev — Doctor of Physical and Mathematical Sciences, professor, academician of the National Academy of Sciences of the Republic of Kazakhstan, professor of the Department of Mathematical and Computer Modeling, International Information Technology University (Kazakhstan)
Bolatbek Rysbaiuly — Doctor of Physical and Mathematical Sciences, professor, professor of the Department of Computing and Data Science, Astana IT University (Kazakhstan)
Yevgeniya Daineko — PhD, research professor, Department of Information Systems, International Information Technology University (Kazakhstan)
Nurzhan Duzbayev — PhD, associate professor, Vice-Rector for Digitalization and Innovation, International Information Technology University (Kazakhstan)
Bakhtgerei Sinchev — Doctor of Technical Sciences, professor, Department of Information Systems, International Information Technology University (Kazakhstan)
Nurgul Seilova — Candidate of Technical Sciences, Dean of the Faculty of Computer Technologies and Cybersecurity, International Information Technology University (Kazakhstan)
Ardak Mukhamediyeva — Candidate of Economic Sciences, Dean of the Faculty of Business, Media and Management, International Information Technology University (Kazakhstan)
Zamira Abdikalikova — PhD, associate professor, Head of the Department of Mathematical and Computer Modeling, International Information Technology University (Kazakhstan)
Yerlan Shildibekov — PhD, associate professor, Head of the Department of Economics and Business, International Information Technology University (Kazakhstan)
Damilya Yeskendirova — Candidate of Technical Sciences, associate professor, Head of the Department of Cybersecurity, International Information Technology University (Kazakhstan)
Aigul Niyazgulova — Candidate of Philological Sciences, Professor, Head of the Department of Media Communications and History of Kazakhstan, International Information Technology University (Kazakhstan)
Altai Aitmagambetov — Candidate of Technical Sciences, Professor, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University (Kazakhstan)
Yelena Bakhtiyarova — Candidate of Technical Sciences, associate professor, Head of the Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University (Kazakhstan)
Kanibek Sansyzbay — PhD, research professor, Department of Cybersecurity, International Information Technology University (Kazakhstan)
Sakhybay Tynymbayev — Candidate of Technical Sciences, Professor, Research Professor, Department of Computer Engineering, International Information Technology University (Kazakhstan)
Ali Abd Almisreb — PhD, associate professor, Department of Cybersecurity, International Information Technology University (Kazakhstan)
Mohamed Ahmed Hamada — PhD, associate professor, Department of Information Systems, International Information Technology University (Kazakhstan)
Yang Im Chu — PhD, Professor, Gachon University (South Korea)
Tadeusz Wallas — PhD, Vice-Rector, Adam Mickiewicz University (Poland)
Orken Mamyrbayev — PhD, Deputy Director for Science, RSE Institute of Information and Computational Technologies, Committee for Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Kazakhstan)
Sergey Bushuyev — Doctor of Technical Sciences, professor, Director of the Ukrainian Project Management Association “UKRNET,” Head of the Department of Project Management, Kyiv National University of Construction and Architecture (Ukraine)
Svetlana Beloshitskaya — Doctor of Technical Sciences, professor, Department of Computing and Data Science, Astana IT University (Kazakhstan)

MANAGING EDITOR
Raushan Mrzabayeva — Master of Science, editor, International Information Technology University (Kazakhstan)

International Journal of Information and Communication Technologies
Periodicity: 4 times a year.
Languages: Kazakh, Russian, English
DOI prefix: 10.54309
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Thematic focus: "Information technology"; "Digital technologies in the development of socio-economic systems"; "Information security and communication technologies".
Distribution: Materials are distributed under the Creative Commons Attribution 4.0
Journal website: https://journal.iitu.edu.kz
Owner: International Information Technology University JSC (Almaty).
Copyright: © International Journal of Information and Communication Technologies, 2026

РЕДАКЦИЯ

БАС РЕДАКТОР:

Колесникова Катерина Викторовна — техника ғылымдарының докторы, профессор, Халықаралық ақпараттық технологиялар университетінің ғылыми-зерттеу қызметі жөніндегі проректор (Қазақстан)

БАС РЕДАКТОРДЫҢ ОРЫНБАСАРЫ:

Ипалакова Мадина Тулегеновна — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университетінің ғылыми-зерттеу қызметі жөніндегі департамент директоры (Қазақстан)

РЕДАКЦИЯЛЫҚ АЛҚА:

Разак Абдул — PhD, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының профессоры (Қазақстан)

Лучио Томмазо де Паолис — Саленто Университеті (Италия) инновация және технологиялық инжиниринг департаменті AVR зертханасының зерттеу және әзірлеу бөлімінің директоры

Лиз Бэкон — профессор, Абертей Университеті (Ұлыбритания) вице-канцлерінің орынбасары

Микеле Пагано — PhD, Пиза Университетінің (Италия) профессоры

Өтелбаев Мухтарбай Өтелбайұлы — физика-математика ғылымдарының докторы, профессор, ҚР ҰҒА академигі, Халықаралық ақпараттық технологиялар университеті математика және компьютерлік модельдеу кафедрасының профессоры (Қазақстан)

Рысбайулы Болатбек — физика-математика ғылымдарының докторы, профессор, Есептеу және деректер ғылымдары департаментінің профессоры, Astana IT University (Казахстан)

Дайнеко Евгения Александровна — PhD, Халықаралық ақпараттық технологиялар университеті ақпараттық жүйелер кафедрасының профессор-зерттеушісі (Қазақстан)

Дузбаев Нуржан Токкужаевич — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті цифрландыру және инновациялар жөніндегі проректор (Қазақстан)

Синчев Бахтгерей Куспанович — техника ғылымдарының докторы, профессор, Халықаралық ақпараттық технологиялар университеті ақпараттық жүйелер кафедрасының профессоры (Қазақстан)

Сейлова Нургуль Абадуллаевна — техника ғылымдарының докторы, Халықаралық ақпараттық технологиялар университеті компьютерлік технологиялар және киберқауіпсіздік факультетінің деканы (Қазақстан)

Мухамедиева Ардак Габитовна — экономика ғылымдарының кандидаты, Халықаралық ақпараттық технологиялар университеті бизнеса медиа және басқару факультетінің деканы (Қазақстан)

Абдикаликова Замира Турсынбаевна — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті математика және компьютерлік модельдеу кафедрасының меңгерушісі (Қазақстан)

Шильдибеков Ерлан Жаржанович — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті экономика және бизнес кафедрасының меңгерушісі (Қазақстан)

Дамеля Максутовна Ескендирова — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының меңгерушісі (Қазақстан)

Ниязгулова Айгуль Аскарбековна — филология ғылымдарының кандидаты, доцент, профессор, Халықаралық ақпараттық технологиялар университеті медиакоммуникация және Қазақстан тарихы кафедрасының меңгерушісі (Қазақстан)

Айтмагамбетов Алтай Зуфарович — техника ғылымдарының кандидаты, Халықаралық ақпараттық технологиялар университеті радиотехника, электроника және телекоммуникация кафедрасының профессоры (Қазақстан)

Бахтиярова Елена Ажибековна — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті радиотехника, электроника және телекоммуникация кафедрасының меңгерушісі (Қазақстан)

Канибек Сансызбай — PhD, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының профессор-зерттеушісі (Қазақстан)

Тынымбаев Сахибай — техника ғылымдарының кандидаты, профессор, Халықаралық ақпараттық технологиялар университеті компьютерлік инженерия кафедрасының профессор-зерттеушісі (Қазақстан)

Алмисреб Али Абд — PhD, Халықаралық ақпараттық технологиялар университеті киберқауіпсіздік кафедрасының қауымдастырылған профессоры (Қазақстан)

Мохамед Ахмед Хамада — PhD, Халықаралық ақпараттық технологиялар университеті ақпараттық жүйелер кафедрасының қауымдастырылған профессоры (Қазақстан)

Янг Им Чу — PhD, Гачон университетінің профессоры (Оңтүстік Корея)

Тадеуш Валлас – PhD, Адам Мицкевич атындағы (Польша) университеттің проректоры

Мамырбаев Оркен Жумажанович — PhD, ҚР ҒЖБМ Ғылым комитеті ақпараттық және есептеу технологиялары институты ӨМК директорының ғылым жөніндегі орынбасары (Қазақстан)

Бушуев Сергей Дмитриевич — техника ғылымдарының докторы, профессор, Украинаның "УКРНЕТ" жобаларды басқару қауымдастығының директоры, Киев ұлттық құрылыс және сәулет университеті жобаларды басқару кафедрасының меңгерушісі (Украина)

Белощицкая Светлана Васильевна — техника ғылымдарының докторы, доцент, Astana IT University есептеу және деректер ғылымы кафедрасының профессоры (Қазақстан)

ЖАУАПТЫ РЕДАКТОР:

Мрзабаева Раушан Жалиевна — магистр, Халықаралық ақпараттық технологиялар университетінің редакторы (Қазақстан)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Префикс DOI: 10.54309
Мерзімділігі: жылына 4 рет.
Басылым тілі: қазақ, орыс, ағылшын.
Тақырып бағыты: "Ақпараттық технологиялар"; "Ақпараттық қауіпсіздік және коммуникациялық технологиялар"; "Әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технология".
Журнал сайты: https://journal.iitu.edu.kz
Тарату: материалдар Creative Commons Attribution 4.0 лицензиясы бойынша таратылады
Меншік иесі: АҚ «Халықаралық ақпараттық технологиялар университеті» (Алматы қ.).
Авторлық құқық: © Халықаралық ақпараттық және коммуникациялық технологиялар журналы, 2026

РЕДАКЦИЯ

ГЛАВНЫЙ РЕДАКТОР:

Колесникова Катерина Викторовна — доктор технических наук, профессор, проректор по научно-исследовательской деятельности Международного университета информационных технологий (Казахстан)

ЗАМЕСТИТЕЛЬ ГЛАВНОГО РЕДАКТОРА:

Ипалакова Мадина Тулегеновна — кандидат технических наук, ассоциированный профессор, директор департамента по научно-исследовательской деятельности Международного университета информационных технологий (Казахстан)

РЕДАКЦИОННАЯ КОЛЛЕГИЯ:

Разак Абдул — PhD, профессор кафедры кибербезопасности Международного университета информационных технологий (Казахстан)

Лучио Томмазо де Паолис — директор отдела исследований и разработок лаборатории AVR департамента инноваций и технологического инжиниринга Университета Саленто (Италия)

Лиз Бэкон — профессор, заместитель вице-канцлера Университета Абертей (Великобритания)

Микеле Пагано — PhD, профессор Университета Пизы (Италия)

Отелбаев Мухтарбай Отелбайулы — доктор физико-математических наук, профессор, академик НАН РК, профессор кафедры математического и компьютерного моделирования Международного университета информационных технологий (Казахстан)

Рысбайулы Болатбек — доктор физико-математических наук, профессор, профессор Astana IT University (Казахстан)

Дайнеко Евгения Александровна — PhD, профессор-исследователь кафедры информационных систем Международного университета информационных технологий (Казахстан)

Дузбаев Нуржан Токкужаевич — PhD, ассоциированный профессор, проректор по цифровизации и инновациям Международного университета информационных технологий (Казахстан)

Синчев Бахтгерей Куспанович — доктор технических наук, профессор, профессор кафедры информационных систем Международного университета информационных технологий (Казахстан)

Сейлова Нургуль Абадуллаевна — кандидат технических наук, декан факультета компьютерных технологий и кибербезопасности Международного университета информационных технологий (Казахстан)

Мухамедиева Ардак Габитовна — кандидат экономических наук, декан факультета бизнеса медиа и управления Международного университета информационных технологий (Казахстан)

Абдикаликова Замира Турсынбаевна — PhD, ассоциированный профессор, заведующая кафедрой математического и компьютерного моделирования Международного университета информационных технологий (Казахстан)

Шильдибеков Ерлан Жаржанович — PhD, ассоциированный профессор, заведующий кафедрой экономики и бизнеса Международного университета информационных технологий (Казахстан)

Дамеля Максутовна Ескендирова — кандидат технических наук, ассоциированный профессор, заведующая кафедрой кибербезопасности Международного университета информационных технологий (Казахстан)

Ниязгулова Айгуль Аскарбековна — кандидат филологических наук, доцент, профессор, заведующая кафедрой медиакоммуникаций и истории Казахстана Международного университета информационных технологий (Казахстан)

Айтмагамбетов Алтай Зуфарович — кандидат технических наук, профессор кафедры радиотехники, электроники и телекоммуникаций Международного университета информационных технологий (Казахстан)

Бахтиярова Елена Ажибековна — кандидат технических наук, ассоциированный профессор, заведующая кафедрой радиотехники, электроники и телекоммуникаций Международного университета информационных технологий (Казахстан)

Канибек Сансызбай – PhD, ассоциированный профессор, профессор-исследователь кафедры кибербезопасности, Международного университета информационных технологий (Казахстан)

Тынымбаев Сахибай – кандидат технических наук, профессор, профессор-исследователь кафедры компьютерной инженерии, Международного университета информационных технологий (Казахстан)

Алмисреб Али Абд — PhD, ассоциированный профессор кафедры кибербезопасности Международного университета информационных технологий (Казахстан)

Мохамед Ахмед Хамада — PhD, ассоциированный профессор кафедры информационных систем Международного университета информационных технологий (Казахстан)

Янг Им Чу — PhD, профессор университета Гачон (Южная Корея)

Тадеуш Валлас – PhD, проректор университета имен Адама Мицкевича (Польша)

Мамырбаев Оркен Жумажанович — PhD, заместитель директора по науке РГП Института информационных и вычислительных технологий Комитета науки МНВО РК (Казахстан)

Бушуев Сергей Дмитриевич — доктор технических наук, профессор, директор Украинской ассоциации управления проектами «УКРНЕТ», заведующий кафедрой управления проектами Киевского национального университета строительства и архитектуры (Украина)

Белощицкая Светлана Васильевна — доктор технических наук, доцент, профессор кафедры вычислений и науки о данных Astana IT University (Казахстан)

ОТВЕТСТВЕННЫЙ РЕДАКТОР:

Мрзабаева Раушан Жалиевна — магистр, редактор Международного университета информационных технологий (Казахстан)

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Префикс DOI: 10.54309
Периодичность: 4 выпусков в год.
Язык издания: казахский, русский, английский.
Тематическая направленность: "Информационные технологии"; "Информационная безопасность и коммуникационные технологии"; "Цифровые технологии в развитии социо-экономических систем".
Сайт журнала: https://journal.iitu.edu.kz
Распространение: материалы распространяются по лицензии Creative Commons Attribution 4.0
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Авторские права: © Международный журнал информационных и коммуникационных технологий, 2026

CONTENTS

DIGITAL TECHNOLOGIES IN THE DEVELOPMENT OF SOCIOECONOMIC SYSTEMS

A.B. Zhalgas, Y.N. Kalpakov, B.Ye. Amirgaliyev
MACHINE LEARNING-DRIVEN OPTIMIZATION OF LOGISTICS IN SMART CITIES: A CASE STUDY OF ASTANA 11
L. Kurmangaziyeva, Sh. Kodanova, M. Urazgaliyeva, O. Findik, S. Iskakova
INTEGRATING FUZZY LOGIC AND ARTIFICIAL INTELLIGENCE IN OPTIMIZING BUSINESS PROCESS AUTOMATION DECISIONS 26
Y. Mailybayev, U. Adilbayeva, R. Amanova
ORGANIZATION OF AN ONLINE SURVEY OF PARTICIPANTS IN THE EDUCATIONAL PROCESS AND ANALYSIS OF THE RESULTS BASED ON THE MODIFIED DELPHI METHOD 48
V.A. Takizhanov, A.Z. Ibragimov, A. Shalakhmetov
SIMULATION-BASED ROBUSTNESS ASSESSMENT OF ASTANA'S BUS NETWORK UNDER RANDOM AND TARGETED FAILURES 63
G. Hamada, M. Hamada
LEVERAGING MACHINE LEARNING MODELS TO BOOST THE MANIPULATION OF UNSTRUCTURED DECISIONS AND UNSTRUCTURED DATA IN DIGITAL BUSINESS FIRMS 78

INFORMATION TECHNOLOGY

M. Zh. Aitimov, G. K. Muratova, Zh. K. Bissenbayeva, I.M. Bapiyev, M. Kassim
SEMANTIC COMPLETENESS IN KAZAKH-LANGUAGE EXTRACTIVE QA THROUGH ONTOLOGY AND RETRIEVAL MECHANISMS 96
A.Aitim, A. Auyezova
ENCODING NP-COMPLETE PROBLEM INSTANCES INTO NATURAL LANGUAGE FOR LLM PROCESSING 109
O.N. Akylbekov, Y.T. Dauletbek, A.N. Moldagulova, G.S. Zakariya, D.A. Gura
MACHINE LEARNING METHODS FOR ANALYSING THREE-DIMENSIONAL SPATIAL DATA IN KAZAKHSTAN'S LAND USE PLANNING 126
S.Zh. Aliaskarov, R.K. Uskenbayeva, A. Razaque, A.B. Kassymova, A.M. Anartayeva
TOWARDS EFFICIENT BIG DATA ANALYTICS IN REGIONAL SYSTEMS: PRACTICAL INSIGHTS FROM HYBRID ARCHITECTURE DEPLOYMENT 146
Zh. Bilyalova, N. Shazhdekeyeva, K. Uteuliyeva, T.Akhmurzina, L. Nurgaliyeva
A MATHEMATICAL MODEL FOR DECISION-MAKING UNDER UNCERTAINTY HYBRID BASED ON FUZZY LOGIC AND GENERATIVE ARTIFICIAL INTELLIGENCE 165
A. Ismailova, G. Yessenbayeva, K. Kadyrkulov, R. Moldasheva, A. Amangeldi
DEVELOPMENT OF A HYBRID DEEP LEARNING MODEL FOR MULTICLASS CLASSIFICATION OF MICROSCOPIC IMAGES OF BACTERIA 186
G. Kalman, J. Kultan, A.N. Ismukamova, N.M. Ausilova, Y.V. Makhatova
A DOMAIN-KNOWLEDGE-BASED MODEL FOR REFERENCE RESOLUTION IN LOW-RESOURCE LANGUAGES....199
Y. Kamen, Zh. Yessendauletova, L. Fazylova, M. Rakhimzhanova, A.M. Nedzved
USING NEURAL NETWORKS FOR OBJECTIVE ASSESSMENT OF ATTENTION IN CHILDREN BASED ON EEG DATA 216
A.Ye. Kulakayeva, Ye.A. Bakhtiyarova, G.T. Jakanova, Sh. Nursultan
COMPARATIVE ANALYSIS OF VARIOUS RADIO WAVE PROPAGATION MODELS FOR MOBILE NETWORK COVERAGE PREDICTION 231
A. Kalykulova, K. Bakirov, A. Shoman, A. Nugumanova
FORMATION OF A UNIFIED DATA SET FOR ANALYZING THE RELATIONSHIP BETWEEN DOG BREEDS AND DISEASES 247
D. Kantayeva, A. Seisenbekkyzy, S. Biloshchytska
COMPARATIVE EVALUATION MODEL BASED ON EXPERT SYSTEMS FOR ABSTRACTIVE SUMMARIZATION OF SCIENTIFIC TEXTS 269
N. Mordvintsev, Zh. Oralbekova, A. Bakiyeva
DEEP LEARNING-BASED VEHICLE MAKE AND MODEL RECOGNITION: DATASET CHALLENGES

AND INFLUENCING FACTORS IN URBAN TRAFFIC 269
M.B. Nurpeissova, Sh.K. Aitkazinova, A.M. Abenov, N.S. Donenbayeva
METHODOLOGY FOR TRANSFORMING SATELLITE COORDINATES INTO A TOPOCENTRIC RECTANGULAR COORDINATE SYSTEM 315
L. Naizabayeva, G. Turken, R.A. Druba
EFFECTIVE METHOD FOR OPTIMAL CLUSTERING IN UNSUPERVISED LEARNING 328
S. Nurgaliyeva, Zh. Seitkaliyev, A. Sakhipov, M. Kali, R. Hrmo
YIELD FORECASTING FOR FRUITS AND VEGETABLES USING RANDOM FOREST ON LARGE-SCALE AGROCLIMATIC DATA 339
A. Nurgalykov, S. Sapakova, D. Zhalmukhanbetov, G. Nurgazy, A. Myrzabayev
LANGUAGE ASYMMETRY IN LLMS: THE IMPACT OF NON-ENGLISH DATA ON ACCURACY AND RELIABILITY 355
A. Ospanov, P. Alonso-Jordá, A. Zhumadillayeva
BLOCKCHAIN-ENABLED ERP WAREHOUSE INTEGRATION WITH IOT DIMENSIONERS AND MACHINE LEARNING–OPTIMIZED DIMENSIONAL WEIGHT RECONCILIATION 371
A.A. Sakhipov, R.B. Seiitbek
EVENT-DRIVEN MICROSERVICES FOR INCIDENT DETECTION AND RESPONSE IN INTELLIGENT TRAFFIC SYSTEM 387
A. Tursynova, B. Omarov
EVOLUTION OF THE EFFECTIVENESS OF CONDITIONAL GAN FOR THE SYNTHESIS OF CT IMAGES OF THE BRAIN IN THE DIAGNOSIS OF STROKE 400
G. Yusupova, K.S. Shadinova, D. Ussipbekova, Zh.Zh. Azhibekova, P. Schmidt
DETERMINATION OF SOIL PROFILE STRATIFICATION AT 0–200 CM DEPTH USING A MULTILEVEL STACKING MODEL 414

INFORMATION SECURITY AND COMMUNICATION TECHNOLOGIES

S.A. Adilzhanova, M.Zh. Sakypbekova, L.Sh. Cherikbaeva, G.A. Tyulepberdinova, G.T. Zhubanysheva
SYSTEMATIC ANALYSIS OF RISK ASSESSMENT METHODS AND MODELS IN INFORMATION SECURITY....427
T. K. Zhukabayeva, D.B. Baumuratova, E. Benkhelifa , N.A. Niyetbayeva
EDGE COMPUTING-BASED TECHNIQUE FOR CONSTRUCTION OF ATTACK DETECTION MEANS IN CYBER-PHYSICAL SYSTEMS OF INDUSTRIAL INTERNET-OF-THINGS 453
N.E. Karabayev, S.K. Serikbayeva, Y.M. Mardenov, B. Tassuov, M. Fajkus
DETECTION OF CYBER ATTACKS IN TRANSPORT NETWORKS BASED ON MACHINE LEARNING METHODS 475
B.A. Kumalakov, A.O. Dargulova
A HYBRID FRAMEWORK FOR RESUME-JOB MATCHING SYSTEM 494
V. Makhatova, B. Dzhugembayeva, A. Gabdulova, L. Nurgaliyeva, A. Abdigaliyeva
MATHEMATICAL MODEL FOR OPTIMAL SENSOR SELECTION IN SIEM SYSTEMS USING THE ANALYTIC HIERARCHY PROCESS 509

МАЗМҰНЫ
ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР

А.Б. Жалғас, Е.Н. Калпаков, Б.Е. Амиргалиев
АҚЫЛДЫ ҚАЛАЛАРДАҒЫ ЛОГИСТИКАНЫ МАШИНАЛЫҚ ОҚЫТУҒА НЕГІЗДЕЛГЕН ОҢТАЙЛАНДЫРУ: АСТАНАНЫҢ ЖАҒДАЙЫН ЗЕРТТЕУ....11
Л.Курманғазиева, Ш. Қоданова, М. Уразғалиева, O. Findik, С. Искакова
ЖАСАНДЫ ИНТЕЛЛЕКТ ПЕН АЙҚЫН ЕМЕС ЛОГИКАНЫ БІРІКТІРУ АРҚЫЛЫ БИЗНЕС-ПРОЦЕСТЕРДІ АВТОМАТТАНДЫРУ ШЕШІМДЕРІН ОҢТАЙЛАНДЫРУ 26
Е. Майлыбаев, У. Адилбаева, Р. Аманова
ҰЙЫМДАСТЫРЫЛҒАН ОНЛАЙН САУАЛНАМА АРҚЫЛЫ БІЛІМ БЕРУ ПРОЦЕСІНЕ ҚАТЫСУШЫЛАРДЫҢ ПІКІРЛЕРІН ЖИНАУ ЖӘНЕ НӘТИЖЕЛЕРІН МОДИФИКАЦИЯЛАНҒАН ДЕЛЬФИ ӘДІСІ НЕГІЗІНДЕ ТАЛДАУ 48
В.А. Такижанов, А.Ж. Ибрагимов, А. Шалахметов

МОДЕЛЬДЕУ НЕГІЗІНДЕ АСТАНАНЫҢ АВТОБУС ЖЕЛІСІНІҢ ТҰРАҚТЫЛЫҒЫН БАҒАЛАУ: КЕЗДЕЙСОҚ ЖӘНЕ МАҚСАТТЫ ІСТЕН ШЫҒУЛАР ЖАҒДАЙЫНДА 63
Г. Хамада, М. Хамада
ЦИФРЛЫҚ БИЗНЕС-ҰЙЫМДАРЫНДА ҚҰРЫЛЫМДАЛМАҒАН ШЕШІМДЕР МЕН ҚҰРЫЛЫМДАЛМАҒАН ДЕРЕКТЕРМЕН ЖҰМЫС ІСТЕУ ТИІМДІЛІГІН АРТТЫРУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН ПАЙДАЛАНУ 78

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР

М.Ж. Айтимов, Г.К. Муратова, Ж.К. Бисенбаева, И.М. Бапиев, М. Кассим
ОНТОЛОГИЯ ЖӘНЕ ІЗДЕУ МЕХАНИЗМДЕРІ АРҚЫЛЫ ҚАЗАҚ ТІЛІНДЕГІ ЭКСТРАКЦИЯЛЫҚ QA-ДАҒЫ СЕМАНТИКАЛЫҚ ТОЛЫҚТЫҚ 96
Ә. Әйтім, Ә. Әуезова
LLM ӨҢДЕУІ ҮШІН NP-ТОЛЫҚ ЕСЕПТЕРДІҢ ЭКЗЕМПЛЯРЛАРЫН ТАБИҒИ ТІЛДЕ КОДТАУ 109
О.Н. Акылбеков, Е.Т. Даулетбек, А.Н. Молдагулова, Г.С. Закария, Д.А. Гура
ҚАЗАҚСТАННЫҢ АУМАҚТЫҚ ЖОСПАРЛАУЫНДАҒЫ ҮШ ӨЛШЕМДІ КЕҢІСТІКТІК МӘЛІМЕТТЕРДІ ТАЛДАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ 126
С.Ж. Алиаскаров, Р.К. Ускенбаева, А. Разак, А.Б. Касымова, А.М. Анартаева
АЙМАҚТЫҚ ЖҮЙЕЛЕРДЕГІ ҮЛКЕН ДЕРЕКТЕРДІ ТИІМДІ ТАЛДАУҒА ҚАРАЙ: ГИБРИДТІ АРХИТЕКТУРАНЫ ЕНГІЗУДІҢ ПРАКТИКАЛЫҚ ТҮСІНІКТЕР.....146
Ж. Билялова, Н. Шаждекеева, К. Өтеулиева, Т. Ахмурзина, Л. Нұрғалиева
АЙҚЫН ЕМЕС ЛОГИКА ЖӘНЕ ГЕНЕРАТИВТІ ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ БЕЛГІСІЗДІК ЖАҒДАЙЫНДА ШЕШІМ ҚАБЫЛДАУДЫҢ ГИБРИДТІ МАТЕМАТИКАЛЫҚ МОДЕЛІ 165
А.А. Исмаилова, Г.Р. Есенбаева, К.К. Кадиркулов, Р.Н. Молдашева, А. Амангелді
РОСКОПИЯЛЫҚ БЕЙНЕЛЕРІН КӨПКЛАССТЫ ЖІКТЕУГЕ АРНАЛҒАН ГИБРИДТІ ТЕРЕҢ ОҚЫТУ МОДЕЛІН ӘЗІРЛЕУ 186
Г. Қалман, К. Ярослав, А.Н. Исмуканова, Н.М. Аусилова, В.Е. Махатова
ПӘНДІК САЛА БІЛІМ НЕГІЗІНДЕ РЕУСРСТАРЫ АЗ ТІЛДЕРДЕГІ РЕФЕРЕНЦИЯНЫ ШЕШУДІҢ МОДЕЛІ.....199
Е.Г. Кәмен, Ж.Т. Есендаулетова, Л.С. Фазылова, М.Б. Рахимжанова, А.М. Недзьведь
ЭЭГ ДЕРЕКТЕРІ БОЙЫНША БАЛАЛАРДЫҢ ЗЕЙІНІН ОБЪЕКТИВТІ БАҒАЛАУ ҮШІН НЕЙРОНДЫҚ ЖЕЛІЛЕРДІ ҚОЛДАНУ 216
А.Е. Кулакаева, Е.А. Бахтиярова, Г.Т. Джаканова, Ш. Нурсултан
ҰЯЛЫ БАЙЛАНЫС ЖЕЛІЛЕРІНІҢ ҚАМТУ АЙМАҒЫН БОЛЖАУҒА АРНАЛҒАН ӘРТҮРЛІ РАДИОТОЛҚЫН ТАРАЛУ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ 231
А. Калыкулова, К. Бакиров, А. Шоман, А. Нугуманова
ИТ ТҰҚЫМДАРЫ МЕН АУРУЛАРЫ АРАСЫНДАҒЫ БАЙЛАНЫСТЫ ТАЛДАУҒА АРНАЛҒАН БІРЫҢҒАЙ ДЕРЕКТЕР ЖИЫНТЫҒЫН ҚАЛЫПТАСТЫРУ 247
Д. Кантаева, А. Сейсенбеккызы, С. Билощицкая
ҒЫЛЫМИ МӘТІНДЕРДІ АБСТРАКТИВТІ РЕФЕРАТТАУҒА АРНАЛҒАН ЭКСПЕРТТІК ЖҮЙЕЛЕРГЕ НЕГІЗДЕЛГЕН САЛЫСТЫРМАЛЫ БАҒАЛАУ МОДЕЛІ 269
Н. Мордвинцев, Ж. Оралбекова, А. Бакиева
ТЕРЕҢ ОҚЫТУҒА НЕГІЗДЕЛГЕН КӨЛІК ҚҰРАЛДАРЫНЫҢ МАРКАСЫ МЕН МОДЕЛІН ТАНУ: ҚАЛАЛЫҚ КӨЛІК ҚОЗҒАЛЫСЫНДАҒЫ ДЕРЕКТЕР ЖИЫНТЫҒЫНЫҢ МӘСЕЛЕЛЕРІ ЖӘНЕ ЫҚПАЛ ЕТУШІ ФАКТОРЛАР 285
М.Б. Нұрпейісова, Ш.Қ. Айтқазинова, А.М. Абенов, Н.С. Дөненбаева
СПУТНИКТІК КООРДИНАТТАРДЫ ТОПОЦЕНТРЛІК ТІК БҰРЫШТЫ КООРДИНАТТАР ЖҮЙЕСІНЕ ТҮРЛЕНДІРУДІҢ ӘДІСТЕМЕСІ 315
Л. Найзабаева, Г. Түркен, Р.А. Друба
БАҚЫЛАНБАЙТЫН ОҚЫТУДАҒЫ ОҢТАЙЛЫ КЛАСТЕРЛЕУДІҢ ТИІМДІ ӘДІСІ 328
С. Нургалиева, Ж. Сеиткалиев, А. Сахипов, М. Кали, Р. Грмо
ҮЛКЕН КӨЛЕМДЕГІ АГРОКЛИМАТТЫҚ ДЕРЕКТЕР НЕГІЗІНДЕ RANDOM FOREST ӘДІСІН ПАЙДАЛАНА ОТЫРЫП, ЖЕМІС-КӨКӨНІС ДАҚЫЛДАРЫНЫҢ ӨНІМДІЛІГІН БОЛЖАУ 339
А. Нұрғалықов, С. Сапақова, Д. Жалмұханбетов, Г. Нұрғазы, А. Мырзабаев
ҮЛКЕН ТІЛДІК МОДЕЛЬДЕРДЕГІ ТІЛДІК АСИММЕТРИЯ: АҒЫЛШЫН ТІЛІНЕН БАСҚА ДЕРЕКТЕРДІҢ ДӘЛДІК ПЕН СЕНІМДІЛІККЕ ӘСЕРІ 355
А. Оспанов, П. Алонсо-Хорда, А. Жұмаділлауева

БЛОКЧЕЙН-ТЕХНОЛОГИЯСЫМЕН ЫҚПАЛДАС ERP ҚОЙМА ЖҮЙЕСІН IOT ДИМЕНСИОНЕРЛЕР ЖӘНЕ МАШИНАЛЫҚ ОҚЫТУ АРҚЫЛЫ ОПТИМИЗАЦИЯЛАНҒАН ӨЛШЕМДІ САЛМАҚ ЕСЕПТЕУМЕН ИНТЕГРАЦИЯЛАУ 371
А.А. Сахипов, Р.Б. Сейітбек
ОҚИҒАҒА БАҒДАРЛАНҒАН МИКРОҚЫЗМЕТТЕР ЖҮЙЕСІ АРҚЫЛЫ АҚЫЛДЫ ТРАФИК ЖҮЙЕЛЕРІНДЕ ОҚИҒАЛАРДЫ АНЫҚТАУ ЖӘНЕ ШАРАЛАР ҚОЛДАНУ 387
А. Турсынова, Б. Омаров
ИНСУЛЬТТІ ДИАГНОСТИКАЛАУ КЕЗІНДЕ МИ КТ-БЕЙНЕЛЕРІН СИНТЕЗДЕУГЕ АРНАЛҒАН ШАРТТЫ GAN ТИІМДІЛІГІНІҢ ЭВОЛЮЦИЯСЫ 400
Г.М. Юсупова, К.С. Шадинова, Д.И. Усипбекова, Ж.Ж. Ажибекова, P. Schmidt
ТОПЫРАҚ ПРОФИЛІНІҢ 0–200 СМ ТЕРЕҢДІКТЕГІ СТРАТИФИКАЦИЯСЫН КӨПДЕҢГЕЙЛІ СТЕКИНГ-МОДЕЛІМЕН АНЫҚТАУ.....414

АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАРҒА АРНАЛҒАН

С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева
АҚПАРАТТЫҚ ҚАУІПСІЗДІКТЕ ТӘУЕКЕЛДЕРДІ БАҒАЛАУ ӘДІСТЕРІ МЕН МОДЕЛЬДЕРІН ЖҮЙЕЛІ ТАЛДАУ.....427
Т.К. Жукабаева, Д . Б Баумуратова, Е. Бенкхелифа, Н.А. Ниетбаева
ШЕКАРАЛЫҚ ЕСЕПТЕУЛЕРДІ ҚОЛДАНА ОТЫРЫП, ЗАТТАРДЫҢ ӨНЕРКӘСІПТІК ИНТЕРНЕТІНІҢ КИБЕРФИЗИКАЛЫҚ ЖҮЙЕЛЕРІНДЕГІ ШАБУЫЛДАРДЫ АНЫҚТАУ ҚҰРАЛДАРЫН ҚҰРУ ӘДІСТЕМЕСІ.....453
Н.Е. Қарабаев, С.К. Серикбаева, Е.М. Марденов, Б. Тасуов, М. Файкус
МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІНЕ НЕГІЗДЕЛГЕН КӨЛІК ЖЕЛІЛЕРІНДЕГІ КИБЕРШАБУЫЛДАРДЫ АНЫҚТАУ 475
Б.А. Кумалаков, А.О. Даргулова
ТҮЙІНДЕМЕЛЕР МЕН ВАКАНСИЯЛАРДЫ АВТОМАТТАНДЫРЫЛҒАН СӘЙКЕСТЕНДІРУГЕ НЕГІЗДЕЛГЕН ГИБРИДТІ ҮМІТКЕРЛЕРДІ ІРІКТЕУ ЖҮЙЕСІ 494
В. Махатова, Б. Джугембаева, А. Габдулова, Л. Нургалиева, А. Абдигалиева
ИЕРАРХИЯЛАРДЫ ТАЛДАУ ӘДІСІ НЕГІЗІНДЕ SIEM ЖҮЙЕЛЕРІНДЕ ОҢТАЙЛЫ СЕНСОРДЫ ТАҢДАУДЫҢ МАТЕМАТИКАЛЫҚ МОДЕЛІ 509

СОДЕРЖАНИЕ
ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИСОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

А.Б. Жалгас, Е.Н. Калпаков, Б.Е. Амиргалиев
ОПТИМИЗАЦИЯ ЛОГИСТИКИ В УМНЫХ ГОРОДАХ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ: НА ПРИМЕРЕ АСТАНЫ 11
Л. Курмангазиева, Ш. Коданова, М. Уразгалиева, О. Финдик, С. Искакова
ИНТЕГРАЦИЯ НЕЧЕТКОЙ ЛОГИКИ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ ОПТИМИЗАЦИИ РЕШЕНИЙ ПО АВТОМАТИЗАЦИИ БИЗНЕС-ПРОЦЕССОВ 26
Е. Майлыбаев, У. Адилбаева, Р. Аманова
СБОР МНЕНИЙ УЧАСТНИКОВ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПОСРЕДСТВОМ ОРГАНИЗОВАННОГО ОНЛАЙН-АНКЕТИРОВАНИЯ И АНАЛИЗ РЕЗУЛЬТАТОВ НА ОСНОВЕ МОДИФИЦИРОВАННОГО МЕТОДА ДЕЛЬФИ 48
В.А. Такижанов, А.Ж. Ибрагимов, А. Шалахметов
ОЦЕНКА УСТОЙЧИВОСТИ АВТОБУСНОЙ СЕТИ АСТАНЫ НА ОСНОВЕ МОДЕЛИРОВАНИЯ ПРИ СЛУЧАЙНЫХ И ЦЕЛЕНАПРАВЛЕННЫХ ОТКАЗАХ 63
Г. Хамада, М. Хамада
ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ РАБОТЫ С НЕСТРУКТУРИРОВАННЫМИ РЕШЕНИЯМИ И НЕСТРУКТУРИРОВАННЫМИ ДАННЫМИ В ЦИФРОВЫХ БИЗНЕС-ОРГАНИЗАЦИЯХ 78

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

М.Ж. Айтимов, Г.К. Муратова, Ж.К. Бисенбаева, И.М. Бапиев, М. Кассим

СЕМАНТИЧЕСКАЯ ПОЛНОТА В КАЗАХСКОЯЗЫЧНОМ EXTRACTIVE QA ЧЕРЕЗ ОНТОЛОГИЮ И RETRIEVAL-МЕХАНИЗМЫ 96
А. Айтим, А. Ауезова
КОДИРОВАНИЕ ЭКЗЕМПЛЯРОВ NP-ПОЛНЫХ ЗАДАЧ В ЕСТЕСТВЕННОМ ЯЗЫКЕ ДЛЯ ОБРАБОТКИ LLM 109
О.Н. Акылбеков, Е.Т. Даулетбек, А.Н. Молдагулова, Г.С. Закария, Д.А. Гура
МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТРЁХМЕРНЫХ ПРОСТРАНСТВЕННЫХ ДАННЫХ В ТЕРРИТОРИАЛЬНОМ ПЛАНИРОВАНИИ КАЗАХСТАНА 126
С.Ж. Алиаскаров, Р.К. Ускенбаева, А. Разак, А.Б. Касымова, А.М. Анартаева
НА ПУТИ К ЭФФЕКТИВНОЙ АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ В РЕГИОНАЛЬНЫХ СИСТЕМАХ: ПРАКТИЧЕСКИЕ ВЫВОДЫ ИЗ ВНЕДРЕНИЯ ГИБРИДНОЙ АРХИТЕКТУРЫ 146
Ж. Билялова, Н. Шаждекеева, К. Утеулиева, Т. Ахмурзина, Л. Нургалиева
ГИБРИДНАЯ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПРИНЯТИЯ РЕШЕНИЙ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ И ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА 165
А.А. Исмаилова, Г.Р. Есенбаева, К.К. Кадиркулов, Р.Н. Молдашева, А. Амангелды
РАЗРАБОТКА ГИБРИДНОЙ МОДЕЛИ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ МИКРОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ БАКТЕРИЙ 186
Г. Калман, К. Ярослав, А.Н. Исмуканова, Н.М. Аусилова, В.Е. Махатова
МОДЕЛЬ НА ОСНОВЕ ЗНАНИЙ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ РАЗРЕШЕНИЯ КОРЕФЕРЕНЦИИ В МАЛОРЕСУРСНЫХ ЯЗЫКАХ 199
Е.Г. Камен, Ж.Т. Есендаулетова, Л.С. Фазылова, М.Б. Рахимжанова, А.М. Недзьведь
ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБЪЕКТИВНОЙ ОЦЕНКИ ВНИМАНИЯ У ДЕТЕЙ ПО ДАННЫМ ЭЭГ 216
А.Е. Кулакаева, Е.А. Бахтиярова , Г.Т. Джаканова , Ш. Нурсултан
СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАЗЛИЧНЫХ МОДЕЛЕЙ РАСПРОСТРАНЕНИЯ РАДИОВОЛН ДЛЯ ПРОГНОЗИРОВАНИЯ ПОКРЫТИЯ СЕТЕЙ МОБИЛЬНОЙ СВЯЗИ 231
А. Калыкулова, К. Бакиров, А. Шоман, А. Нугуманова
ФОРМИРОВАНИЕ ЕДИНОГО НАБОРА ДАННЫХ ДЛЯ АНАЛИЗА ВЗАИМОСВЯЗИ МЕЖДУ ПОРОДАМИ СОБАК И ЗАБОЛЕВАНИЯМИ 247
Д. Кантаева, А. Сейсенбеккызы, С. Билощицкая
СРАВНИТЕЛЬНАЯ МОДЕЛЬ ОЦЕНКИ НА ОСНОВЕ ЭКСПЕРТНЫХ СИСТЕМ ДЛЯ АБСТРАКТИВНОГО РЕФЕРИРОВАНИЯ НАУЧНЫХ ТЕКСТОВ 269
Н. Мордвинцев, Ж. Оралбекова, А. Бакиева
РАСПОЗНАВАНИЕ МАРКИ И МОДЕЛИ ТРАНСПОРТНЫХ СРЕДСТВ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ: ПРОБЛЕМЫ НАБОРОВ ДАННЫХ И ФАКТОРЫ, ВЛИЯЮЩИЕ НА РАБОТУ СИСТЕМЫ В УСЛОВИЯХ ГОРОДСКОГО ДВИЖЕНИЯ 285
М.Б. Нурпеисова, Ш.К. Айтказинова, А.М. Абенов, Н.С. Доненбаева
МЕТОДИКА ПРЕОБРАЗОВАНИЯ СПУТНИКОВЫХ КООРДИНАТ В ТОПОЦЕНТРИЧЕСКУЮ ПРЯМОУГОЛЬНУЮ СИСТЕМУ КООРДИНАТ 315
Л. Найзабаева, Г. Туркен, Р.А. Друба
ЭФФЕКТИВНЫЙ МЕТОД ОПТИМАЛЬНОЙ КЛАСТЕРИЗАЦИИ В ОБУЧЕНИИ БЕЗ УЧИТЕЛЯ 328
С. Нургалиева, Ж. Сеиткалиев, А. Сахипов, М. Кали, Р. Грмо
ПРОГНОЗИРОВАНИЕ УРОЖАЙНОСТИ ФРУКТОВ И ОВОЩЕЙ С ИСПОЛЬЗОВАНИЕМ RANDOM FOREST НА ОСНОВЕ КРУПНОМАСШТАБНЫХ АГРОКЛИМАТИЧЕСКИХ ДАННЫХ 339
А. Нургалыков, С. Сапакова, Д. Жалмуханбетов, Г. Нургазы, А. Мырзабаев
ЯЗЫКОВАЯ АСИММЕТРИЯ В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ: ВЛИЯНИЕ ДАННЫХ НА НЕАНГЛИЙСКИХ ЯЗЫКАХ НА ТОЧНОСТЬ И НАДЁЖНОСТЬ 355
А. Оспанов, П. Алонсо-Хорда, А. Жумадиллаева
ИНТЕГРАЦИЯ СКЛАДСКИХ МОДУЛЕЙ ERP-СИСТЕМ С ИСПОЛЬЗОВАНИЕМ БЛОКЧЕЙНА, IOT-ДИМЕНСИОНЕРОВ И ОПТИМИЗИРОВАННОГО МАШИННЫМ ОБУЧЕНИЕМ РАСЧЁТА ГАБАРИТНОГО ВЕСА 371
А.А. Сахипов, Р.Б. Сейитбек
СОБЫТИЯ-ОРИЕНТИРОВАННЫЕ МИКРОСЕРВИСЫ ДЛЯ ОБНАРУЖЕНИЯ И РЕАГИРОВАНИЯ НА ИНЦИДЕНТЫ В ИНТЕЛЛЕКТУАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМАХ 387
А. Турсынова, Б. Омаров
ЭВОЛЮЦИЯ ЭФФЕКТИВНОСТИ УСЛОВНОГО GAN ДЛЯ СИНТЕЗА КТ-ИЗОБРАЖЕНИЙ ГОЛОВНО-

ГО МОЗГА ПРИ ДИАГНОСТИКЕ ИНСУЛЬТА 400
Г.М. Юсупова, К.С. Шадинова, Д.И. Усипбекова, Ж.Ж. Ажибекова, П. Шмидт
ОПРЕДЕЛЕНИЕ СТРАТИФИКАЦИИ ПОЧВЕННОГО ПРОФИЛЯ НА ГЛУБИНЕ 0–200 СМ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ МНОГОУРОВНЕВОГО НАЛОЖЕНИЯ 414

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева
СИСТЕМАТИЧЕСКИЙ АНАЛИЗ МЕТОДОВ И МОДЕЛЕЙ ОЦЕНКИ РИСКОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ 427
Т.К. Жукабаева, Д Б . Баумуратова, Е. Бенкхелифа, Н.А. Ниетбаева
МЕТОДИКА ПОСТРОЕНИЯ СРЕДСТВ ОБНАРУЖЕНИЯ АТАК В КИБЕРФИЗИЧЕСКИХ СИСТЕМАХ ПРОМЫШЛЕННОГО ИНТЕРНЕТА ВЕЩЕЙ С ИСПОЛЬЗОВАНИЕМ ГРАНИЧНЫХ ВЫЧИСЛЕНИЙ 453
Н.Е. Карабаев, С.К. Серикбаева, Е.М. Марденов, Б. Тасуов, М. Файкус
ОБНАРУЖЕНИЕ КИБЕРАТАК В ТРАНСПОРТНЫХ СЕТЯХ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ 475
Б.А. Кумалаков, А.О. Даргулова
ГИБРИДНЫЙ ПОДХОД К АВТОМАТИЗИРОВАННОМУ ПОДБОРУ КАНДИДАТОВ НА ОСНОВЕ СОПОСТАВЛЕНИЯ РЕЗЮМЕ И ВАКАНСИЙ 494
В. Махатова, Б. Джугембаева, А. Габдулова, Л. Нургалиева, А. Абдигалиева
МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ВЫБОРА ОПТИМАЛЬНОГО СЕНСОРА В SIEM-СИСТЕМАХ СРЕДСТВАМИ МЕТОДА АНАЛИЗА ИЕРАРХИЙ 509

DIGITAL TECHNOLOGIES IN THE DEVELOPMENT OF SOCIO-ECONOMIC SYSTEMS
ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР
ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИСОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 11–25
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.001

MACHINE LEARNING-DRIVEN OPTIMIZATION OF LOGISTICS IN SMART CITIES: A CASE STUDY OF ASTANA

A.B. Zhalgas*, Y.N. Kalpakov, B.Ye. Amirgaliyev
Astana IT University, Astana, Kazakhstan.
E-mail: aidana.zhalgas@astanait.edu.kz

Aidana B. Zhalgas — PhD Student, Vice Dean for Student Affairs, Astana IT University
E-mail: aidana.zhalgas@astanait.edu.kz, https://orcid.org/0000-0003-1091-8483;
Yerbolat N. Kalpakov — PhD, Vice Dean for Academic Affairs, Astana IT University
https://orcid.org/0000-0002-8898-7190;
Beibut Y. Amirgaliyev — Candidate of Technical Sciences, Professor, Department of Computer Engineering, Astana IT University
https://orcid.org/0000-0003-0355-5856.

© A.B. Zhalgas, Y.N. Kalpakov, B.Ye. Amirgaliyev

Abstract. This study explores the development and application of machine learning (ML) techniques to enhance the efficiency of logistics operations in the context of smart cities. Focusing on clustering methods and shortest path algorithms, the research aims to optimize courier delivery logistics in Astana, Kazakhstan. The study integrates K-means clustering and Dijkstra's algorithm to address challenges in urban logistics, providing actionable insights into route optimization, resource allocation, and operational scalability. By leveraging geospatial data and advanced clustering methodologies, this research offers a framework for improving logistics operations while contributing to broader smart city goals. Experimental outcomes prove the efficiency of the suggested method: K-means clustering and Dijkstra algorithm worked on the routes enablement and decreased the operation time. In particular, K-means took 1.12 seconds and Fuzzy C-means indicated a higher speed in clustering of 0.023 seconds. Dynamic

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

visualization, optimization of delivery routes within logistic clusters was the couple use of the K-means and the Dijkstra routing that was used to make decision-making and minimize the travel distances.

Keywords: machine learning, logistics optimization, smart city, K-means clustering, Dijkstra's algorithm, route efficiency, geographic information systems (GIS), urban planning, transportation management

For citation: A.B. Zhalgas, Y.N. Kalpakov, B.Ye. Amirgaliyev (2026). Machine learning-driven optimization of logistics in smart cities: a case study of Astana // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 11-25. https://doi.org/10.54309/IJICT.2026.25.1.001. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

АҚЫЛДЫ ҚАЛАЛАРДАҒЫ ЛОГИСТИКАНЫ МАШИНАЛЫҚ ОҚЫТУҒА НЕГІЗДЕЛГЕН ОҢТАЙЛАНДЫРУ: АСТАНАНЫҢ ЖАҒДАЙЫН ЗЕРТТЕУ

А.Б. Жалғас*, Е.Н. Калпаков, Б.Е. Амиргалиев
Astana IT University, Астана, Қазақстан.
E-mail: aidana.zhalgas@astanait.edu.kz

Жалғас Айдана Бозқұланқызы — PhD студент, деканның тәрбие жұмысы бойынша орынбасары, Astana IT University
E-mail: aidana.zhalgas@astanait.edu.kz, https://orcid.org/0000-0003-1091-8483;
Калпаков Ерболат Нұрбергенович — PhD, деканның оқу ісі жөніндегі орынбасары, Astana IT University
https://orcid.org/0000-0002-8898-7190;
Амиргалиев Бейбут Едилханович — т.ғ.к., профессор, Компьютерлік инженерия департаменті, Astana IT University
https://orcid.org/0000-0003-0355-5856.

© А.Б. Жалғас, Е.Н. Калпаков, Б.Е. Амиргалиев

Аннотация. Бұл зерттеу "ақылды қалалар" контекстінде логистикалық операциялардың тиімділігін арттыру үшін машиналық оқыту (ML) әдістерін әзірлеу мен қолдануды зерттейді. Кластерлеу әдістеріне және ең қысқа жол алгоритмдеріне назар аудара отырып, зерттеу Қазақстанның Астана қаласында курьерлік жеткізу логистикасын оңтайландыруға бағытталған. Зерттеу маршруттарды оңтайландыру, ресурстарды бөлу және операциялық ауқымдылық туралы нақты түсінік бере отырып, қалалық логистикадағы мәселелерді шешу үшін K-means кластерлеуін және Дейкстр алгоритмін біріктіреді. Геокеңістіктік деректерді және кластерлеудің озық әдістемелерін пайдалана отырып, бұл зерттеу "ақылды қалалардың" кеңірек мақсаттарына қол жеткізуге ықпал ете отырып, логистикалық операцияларды жақсарту үшін негіз ұсынады. Эксперименттік

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

нәтижелер ұсынылған әдістің тиімділігін дәлелдейді: K-means кластерлеуді білдіреді және Дейкстр алгоритмі маршруттарды қосу кезінде айтарлықтай жұмыс істеді және жұмыс уақытын қысқартты. Атап айтқанда, K-means 1,12 секундты алды, ал Fuzzy C-means кластерлеудің жоғары жылдамдығын 0,023 секундты көрсетті. Динамикалық визуализация, логистикалық кластерлердегі жеткізу маршруттарын оңтайландыру шешім қабылдау және жол жүру қашықтығын азайту үшін пайдаланылған K-means мен Дейкстр маршруттауының жұптық қолданылуы болды.

Түйінді сөздер: машиналық оқыту, логистиканы оңтайландыру, ақылды қала, K-means кластерлеу, Дейкстр алгоритмі, маршруттардың тиімділігі, географиялық ақпараттық жүйелер (ГАЖ), қала құрылысы, көлікті басқару

Дәйексөздер үшін: А.Б. Жалғас, Е.Н. Калпаков, Б.Е. Амиргалиев (2026). Ақылды қалалардағы логистиканы машиналық оқытуға негізделген оңтайландыру: астананың жағдайын зерттеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. № 25. Б. 13–27. 11-25 бет. https://doi.org/10.54309/IJICT.2026.25.1.001. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ОПТИМИЗАЦИЯ ЛОГИСТИКИ В УМНЫХ ГОРОДАХ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ: НА ПРИМЕРЕ АСТАНЫ

А.Б. Жалгас*, Е.Н. Калпаков, Б.Е. Амиргалиев
Astana IT University, Астана, Казахстан.
E-mail: aidana.zhalgas@astanait.edu.kz

Жалгас Айдана Бозкуланковна — PhD студент, заместитель декана по воспитательной работе, Astana IT University
E-mail: aidana.zhalgas@astanait.edu.kz, https://orcid.org/0000-0003-1091-8483;
Калпаков Ерболат Нурбергенович — PhD, заместитель декана по академической работе, Astana IT University
https://orcid.org/0000-0002-8898-7190;
Амиргалиев Бейбут Едилханович — кандидат технических наук, профессор, департамент компьютерной инженерии, Astana IT University
https://orcid.org/0000-0003-0355-5856.

© А.Б. Жалгас, Е.Н. Калпаков, Б.Е. Амиргалиев

Аннотация. Это исследование посвящено разработке и применению методов машинного обучения (ML) для повышения эффективности логистических операций в контексте умных городов. Основное внимание в исследовании уделяется методам кластеризации и алгоритмам кратчайшего пути, а также оптимизации логистики курьерской доставки в Астане, Казахстан. Исследование

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

объединяет кластеризацию K-means и алгоритм Дейкстры для решения задач городской логистики, предоставляя полезную информацию об оптимизации маршрутов, распределении ресурсов и масштабируемости операций. Используя геопространственные данные и передовые методологии кластеризации, это исследование предлагает основу для улучшения логистических операций, способствуя при этом достижению более широких целей умного города. Результаты экспериментов подтверждают эффективность предложенного метода: кластеризация по K-means и алгоритм Дейкстры значительно улучшили работу с маршрутами и сократили время работы. В частности, K-means заняло 1,12 секунды, а Fuzzy C-means показало более высокую скорость кластеризации – 0,023 секунды. Динамическая визуализация и оптимизация маршрутов доставки в логистических кластерах – это совместное использование K-means и алгоритм Dijkstra, которые использовались для принятия решений и минимизации расстояний в пути.

Ключевые слова: машинное обучение, оптимизация логистики, умный город, кластеризация K-means, алгоритм Дейкстры, эффективность маршрутов, географические информационные системы (ГИС), городское планирование, управление транспортом

Для цитирования: А.Б. Жалгас, Е.Н. Калпаков, Б.Е. Амиргалиев (2026). Оптимизация логистики в умных городах на основе машинного обучения: на примере астаны // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 11-25. (На анг.). https://doi.org/10.54309/IJICT.2026.25.1.001.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

In today's globalized world, coordination systems play a key part in guaranteeing proficient transportation of merchandise. This segment may be a foundation for both residential and worldwide exchange, and its significance has expanded due to the issues that have emerged amid the COVID-19 widespread (Barabási et al., 2004). These challenges have highlighted the pressing have to make coordination frameworks that combine unwavering quality, adaptability and tall productivity. Such frameworks are critical for the quick conveyance of merchandise, keeping up financial soundness and expanding strength to worldwide stuns. The logistics systems in the modern environment are not limited to pre-conditioning the operational continuity only but should also help the decision-makers to act in a flexible and delays-free reaction to the changes. These frameworks are fundamental to the speedy transportation of items, maintains financial viability, and grows might to global shivers. As cities ended up "smart cities", errands and openings emerge for optimizing coordination forms, particularly through the development of ideal courses for coordination companies and dispatch administrations (Bezdek et al., 2020). This process involves the absorption of artificial intelligence,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

IoT and big data analytics into urban logistics to act in accordance with rising needs and infrastructural limitations. Astana, the capital of Kazakhstan, may be a prime illustration of such urban change, advancing into an energetic center of financial and innovative improvement. The urban demography and the expanding numbers of the city population necessitated by the ever-increasing business activity and functions requires a change in approach to the automatic control logistics modelling that is intelligent enough to adjust to the changing forces in urban environments.

The urban landscape of Astana is characterized by its rapid growth and increasing complexity, necessitating the adoption of innovative logistics and transportation strategies. The city has a serious logistic issue connected with traffic jam, space disequilibrium in service delivery, and absence of comprehensive digital infrastructuring. The city's trajectory towards becoming a smart city makes it an exemplary model for applying the unsupervised ML techniques to improve logistics operations' efficiency (Bradley et al., 1998). This further qualifies Astana as a research topic as well as a place to test scalable smart logistics technology which would subsequently be applied to the other urban settings of Kazakhstan and Central Asia.

The objectives of this research are carefully designed to address the intricacies of enhancing logistics efficiency in an urban context. These objectives include:

1. Extensive Analysis and Assessment of Routing Efficiency Techniques: This requires a careful investigation of current coordinations industry approaches pointed at moving forward directing viability. With the assistance of a wide extend of sources, this examination looks for to supply a strong hypothetical system for the consider.

2. Identification and Exploration of Routing Efficiency Challenges: The investigate will examine the challenges to accomplishing high directing productivity inside Astana's urban system. This examination will consider different components, counting activity clog, framework imperatives, and the spatial dissemination of conveyance focuses.

3. Formulation of Strategic Solutions to Enhance Routing Efficiency: Leveraging the knowledge from the explanatory stage, the research will propose imaginative techniques pointed at overcoming the distinguished challenges. This activity will include the vital application of K-means and C-means algorithms to plan directing arrangements customized to the special logistics and messenger benefit scene of Astana.

4. Adoption of Global Best Practices within the Local Context: The research will undertake comparative investigation of worldwide best practices in logistics routing, pointing to distinguish and adjust procedures that are significant and implementable in Astana. This comparison is vital to guaranteeing that the proposed arrangements are informed by successful practices.

5. Development of a Predictive Model for Efficient Routing: The result of this research will be the creation of a prescient demonstrate competent of precisely distinguishing the foremost productive courses for logistics and courier services in Astana. This model will coordinate different parameters, counting real-time activity information, vehicle capacities, and delivery plans, to powerfully optimize delivery routes.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

This paper centralizes the application of clustering algorithms, particularly K-means and C-means, as tools to tackle Astana's logistical challenges. These algorithms provide a sophisticated means of analyzing complex datasets, facilitating the identification of routing strategies that maximize efficiency and scalability (Bouhmala, 2016). The application of these algorithms marks a pivotal advancement in addressing urban logistics complexities, offering actionable insights to improve routing efficiency and operational effectiveness.

Materials and methods.

1. The Mechanics of K-means Clustering

The K-means clustering algorithm arranges a collection of objects so that those in the same group (or cluster) are more similar to one another than to those in other groups (Christopher et al.,). It is especially helpful in the field of logistics in smart cities, where it may assist in allocating resources and optimizing routes according to supply and demand trends.

Each data point is assigned to the nearest centroid, determined by the Euclidean distance, in the first stage of the K-means algorithm. After then, centroids are recalculated using the newly created clusters during the centroid update phase (Cohen et al., 2014). This cycle of assignment and update repeats until the cluster memberships stabilize or a set number of iterations are completed (Cormen et al., 2019).

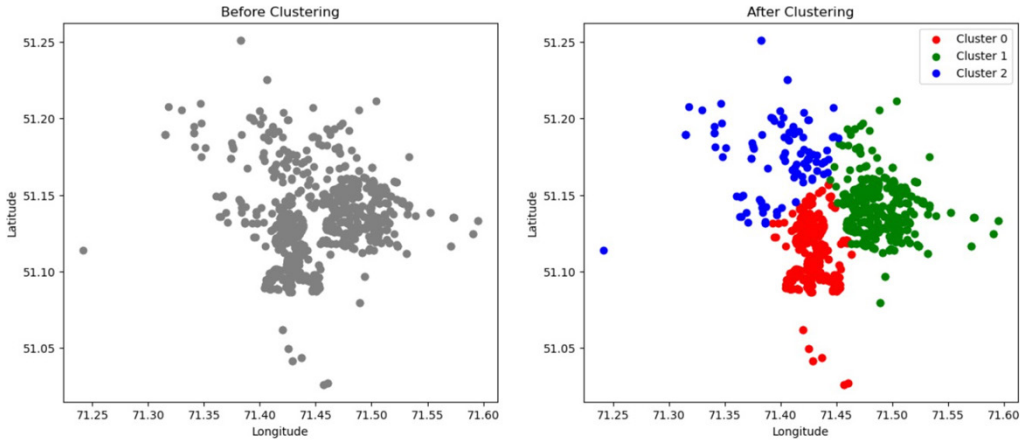


Fig. 1. Depiction of K-means Clustering

2. Fuzzy C-means Clustering

Fuzzy C-means clustering falls into the category of objective function algorithms, which are designed to minimize a specific error metric (Dijkstra, 1959). This method is distinguished using *'c'*, representing the number of clusters or features, and employs fuzzy logic, simplifying the approach to fuzzy. The technique utilizes a fuzzy membership function that assigns a degree of belonging to each data point across various clusters, like estimating pixel probabilities in image analysis (Dunn, 2019).

To start the Fuzzy C-means Clustering process, first define the number of clusters

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

'c', which can range from 2 to a predefined upper limit *'c_n'*. Choose a fuzziness parameter *'m'* to set the level of cluster fuzziness. Initialize the partition matrix *U(0)* with initial membership degrees and label the beginning of the sequential clustering operations (Glaeser, 2019).

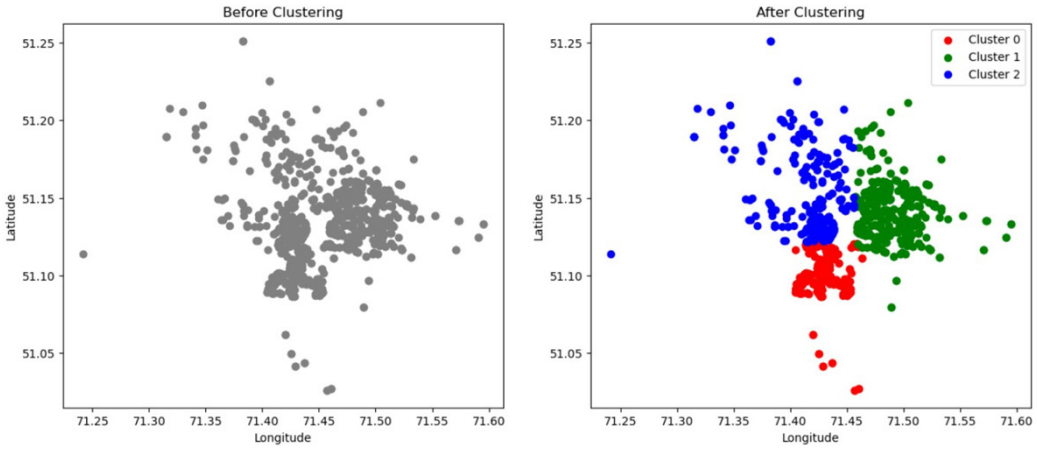


Fig. 2. Illustration of Clustering with the Fuzzy C-means

To calculate the center vector V_{ij} for each cluster, use the following formula:

$$V_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{n} (\mu_{ik})^m}$$

Here, *μ_ik* represents the membership degree of the *k*-th data point in the *i*-th cluster, *x_kj* is the *j*-th attribute of the *k*-th data point, and *m* is the fuzziness exponent that weights the influence of each data point's membership degree on the cluster center (Harary, 2018).

The distance compute matrix *D[c,n]* measures the Euclidean distances between each data point and the cluster centers:

$$D = \left(\sum_{j=1}^{m} (x_{kj} - v_{kj})^2\right)^{\frac{1}{2}}$$

3. Graph Theory

The study of graphs, which are abstract representations that depict relationships and interactions between different entities, is the main emphasis of graph theory, an important area of mathematics. A set of vertices *V* and a set of edges *E* make up a graph *G = (V, E)*. While edges signify the links or interconnections between these things, vertices indicate discrete entities (Jain, 2019). Graphs can be grouped according to their characteristics. Bidirectional interactions between vertices are indicated by undirected graphs (Lee, 2006), which feature edges without direction. On the other hand, directed graphs, also known as digraphs, have edges that point in particular directions, signifying

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

one-way links. In contrast to unweighted graphs, which treat all edges equally, weighted graphs provide edges numerical values, such as costs, distances, or other metrics (Miyamoto, 1997).

4. Dijkstra's Algorithm

Dijkstra's algorithm could be an essential method in graph hypothesis, broadly utilized to decide the foremost proficient hubs in a chart, such as on street systems (Pal, 1995). The algorithm operates on a graph $G = (V, E)$, where V represents the vertices and E the edges. The process begins by initializing the tentative distance for each vertex: the source vertex has zero, while all other vertices are initially assigned a distance of infinity. All vertices are marked as unvisited, and the source vertex is set as the current vertex. The algorithm then iteratively examines the unvisited neighbors of the current vertex, updating their tentative distances based on the shortest known path. Once all neighbors are processed, the vertex that is currently marked as visited will now be the one with the smallest tentative distance selected. This cycle continues until all vertices are visited or the smallest tentative distance among the unvisited vertices becomes infinite (Yu et al., 2024).

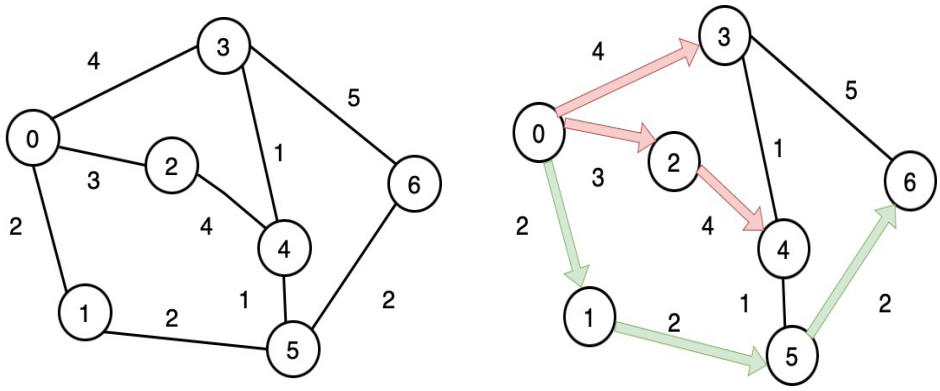


Fig. 3. Visualization of Dijkstra's algorithm

Results and discussion.

Data Collection

A comprehensive dataset compiled on various companies, which includes their precise geographical coordinates (i.e., longitude, latitude), business classifications, and district placements. These information focuses were sourced with fastidious care from 2GIS to guarantee their exactness and completeness. Essential objective in collecting this dataset is to utilize it as a foundational asset for creating ML calculations that can optimize logistics forms. The arranges (i.e., longitude, scope) are crucial for any geospatial investigations and for understanding assignments related to directing and geolocation. The district location provides additional context and allows for consideration of regional characteristics when making logistics decisions. The type of business allows

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

for the classification of companies by industry and activity type, which is necessary for analyzing industry trends and resource planning. Each entry in the dataset is intricately detailed, including business types for industry categorization, essential coordinates for geospatial analyses, and district locations for contextual understanding in logistic decision-making. This dataset spans a multitude of companies, providing a robust foundation for constructing and refining predictive models, streamlining delivery routes, and formulating strategies to elevate overall business process efficiency within the logistic cluster an interconnected network of enterprises offering a diverse range of logistics services.

Moreover, the data set is important in the deployment of spatial clustering algorithms because it gives high geolocation resolution. This has been made possible by the availability of accurate coordinates thus allowing accurate clustering of any logistics facility and location of services with real-world geography being the basis of the model outputs. Also, it is possible to complement the dataset with the temporal characteristics, including peak delivery time, seasonal demand fluctuations, and business working hours, which allows logistics planning to be more dynamic and responsive. Such a temporal-spatial combination is particularly necessary in the process of creating time-sensitive machine learning solutions that will be able to respond to varying circumstances in the city. The fullness of the dataset also would allow its use with multi-objective optimization problems, and trade-offs between distance, delivery time, and resource availability may be investigated. This would be very useful in city logistics operations where the decisions must compromise between being cost-effective or deliver good service. As an enhancement in the future this dataset can be tied with real time API feedbacks like traffic information, weather and schedule of events available to the public to make the system more contextually intelligent. Besides, it demonstrates the scalability of a similar template to other cities that desire to use their ML-based logistics optimization systems. Based on both extensibility and precision of this dataset, it will be useful both in the long-term operations and academics studies on urban informatics and smart city logistics.

Data Visualization

Following the meticulous collection and compilation of a dataset on various companies within Astana, transitioned to transform this dataset into a visually interpretable format. This step was crucial for both simplifying complex data and facilitating detailed geographic analyses. To achieve this, ‘folium’ library was utilized to create an interactive map that visually represents each company’s geographical positioning along with its business specifics. Clarity, interactivity, and analytical value principles were used during the visualization process. Graphical display is important in the conversion of raw coordinates to practical information in spatial data. Using the power of folium, the users can now examine the business environment of Astana interactively in a manner that cannot be achieved with either a table or a map. The mentioned maps could be used to support the decision making of logistic managers, urban planners as well as researchers, and help in defining the geographic clusters, geographies with compromised

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

service, and delivery coverage optimization.

The map was initiated with a central focus on Astana. This central positioning on the map is essential as it provides a geographical context, helping to interpret spatial relationships between the data points. Such a visual frame is instrumental for stakeholders who rely on geographic layouts for strategic decision-making.

For additional usability, the map could be set to contain zoom, tailoring the map tiles, and layer selectors, which allow the filtering by business type or a district. The features give a multi dimensional image of the dataset thereby facilitating more detailed studies. To illustrate, the user can segregate a group of logistic firms in particular districts or can study the white spaces on the peripheral regions. At the same time, since business metadata is integrated in marker tooltips, this provides immediate access to company data without leaving the visualization and makes the experience very smooth.

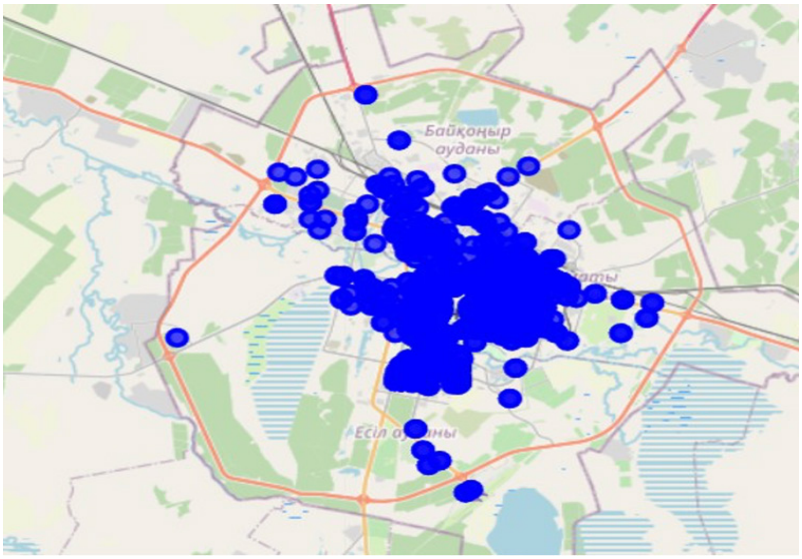


Fig. 4. Visualization of a folium map

'CircleMarker' tool within 'folium' was used to visually mark each company on the map, which allows for the creation of distinct, circular markers. Each data point is represented by a blue circular marker on the map, plotted based on its coordinates. These markers are designed to enhance visibility and uniformity. They also embed crucial information such as business type and district within their popup and tooltip, making the map not only informative but interactive. The resulting map is a dynamic visualization tool that displays the spatial distribution of businesses across Astana.

Spatial Data Analysis and Visualization Techniques

Data clustering using the K-means algorithm was implemented and visualized the results on a map. The data was divided into 3 clusters. Adding a column of clusters to the DataFrame (df). Next, we have a df containing Latitude and Longitude columns with the coordinates of the data points. The fit predict method of the K-means object is called. The fit predict method accomplishes two tasks:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

```
logistics field: ['почта' 'кафе/ресторан' 'лог-кампния' 'доставка']
           Название    Долгота     Широта           район  вид логистики
0          Казпочта  71.364747  51.135793      Нура район          почта
1                SF  71.421466  51.123062     Есиль район  кафе/ресторан
2                SF  71.444708  51.159920  Байконыр район  кафе/ресторан
3                SF  71.413177  51.100906     Есиль район  кафе/ресторан
4                SF  71.442174  51.093828     Есиль район  кафе/ресторан
...             ...        ...        ...             ...            ...
1300       Крендель  71.428763  51.130520     Есиль район  кафе/ресторан
1301     Лаундж-бар  71.426236  51.091852     Есиль район  кафе/ресторан
1302  Лепим и варим  71.408749  51.088617     Есиль район  кафе/ресторан
1303         Ми Ген  71.426090  51.149419     Есиль район  кафе/ресторан
1304     Наша пицца  71.426857  51.093043     Есиль район  кафе/ресторан

[1305 rows x 5 columns]
```

Fig. 5. DataFrame

Fits the model to the data.

Predicts the closest cluster each sample in the data belongs to.

The resulting DataFrame (df) will have a new column named Cluster that indicates the cluster to which each data point belongs.

Trains the K-means model on the latitude and longitude data from ‘df[’Latitude’, ’Longitude’]’.

Predicts clusters for each data point and stores the results in a new Cluster column in the same ‘df’.

The Cluster column will contain integers from 0 to 2, where each number corresponds to the cluster to which the data point belongs. The execution time of each algorithm was given in Table 1. Even though Fuzzy C-means demonstrated a significantly lower execution time compared to K-means, this result is primarily attributed to implementation-specific factors and dataset characteristics rather than algorithmic superiority. The dataset used in this study is low-dimensional and moderately sized, allowing Fuzzy C-means to converge rapidly with minimal iterations. Contrary, the K-means implementation includes additional computational overhead related to initialization strategies and convergence checks. From a practical perspective, the observed time difference does not have a significant impact on real-time logistics systems, as clustering is typically performed offline or periodically, while real-time performance is constrained by routing algorithms and traffic-aware graph processing.

Table 1 – Execution time

Algorithm	Execution time (in seconds)
K-means	1.119977397918701
Fuzzy C-means	0.022998571395874023

K-means with Dijkstra’s Algorithm

The K-means algorithm from the scikit-learn library with a few clusters of 3 was initialized. This number represents the groups into which logistic objects will be orga-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

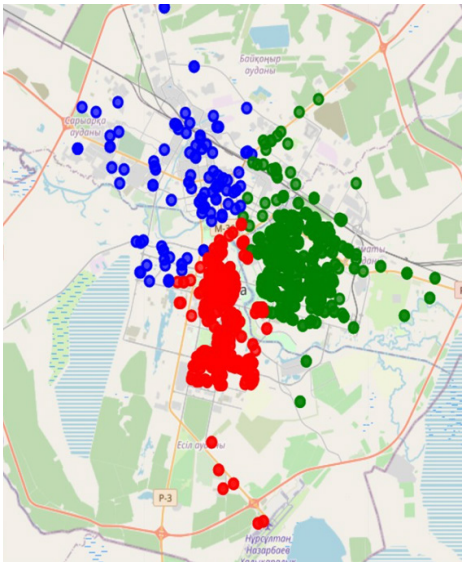
nized based on their geographic coordinates. For visualization, each logistics facility is marked on an interactive map with a circle colored according to its cluster assignment. This not only aids in visual differentiation but also enhances strategic decision-making about resource allocation. Dijkstra's algorithm was also used to optimize routes in the urban environment of Astana. Using the OSMnx library, a graph of the Astana road network was constructed and applied Dijkstra's algorithm to determine the shortest paths between different logistic nodes.

These two algorithms, K-means and Dijkstra, when combined form a synergistic effect, where a whole spatial analysis and optimization of the path could be performed properly. The advantage of the clustering process is that it allows the cluster of facilities within shorter distances to be found and long distance journeys to and fro are limited and route planning can be done at local levels within each cluster. This reduces the amount of fuel used and the time of operation as well as efficiency of the delivery cycles. Also, the dynamic aspect of the Dijkstra algorithm permits its application to the real-world inconsistencies in traffic and constraints which is important in the urban environment where there are fluctuations in road loading and road availability. The usage of OSMnx library extends the real-world applicability of the model with the ability to use real-world geographical and infrastructural information provided in the OpenStreetMap. This makes sure that the results of the calculation are valid to the real road network of Astana and can be applied in the direct case of real actions of logistics. Further, stakeholders can easily navigate through various clusters and visualise an optimal path in a familiar user interface with the interactive map that has been created in HTML format. The usage of this visual interface is especially helpful when working with the decision-makers who are often not technically trained in data science and need clean actionable insights. The map includes hover-based tooltips, and cluster specific legend, more insights into the spatial network and routing connections are possible. Future enhancements could be the integration of real-time traffic through API calls as well as vehicle specific constraints using payload, fuel type or emission type. Altogether, even such a simple application demonstrates the potential of machine learning and graph theory coming together to make logistics management an efficient, smart, and data-driven environment.

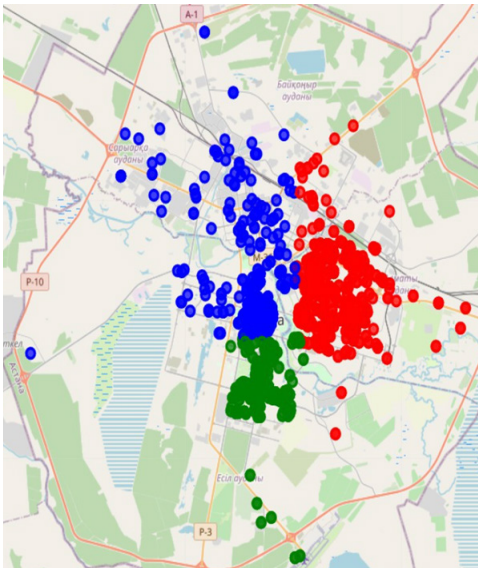
The code combines the K-means clustering results and the optimized routing paths on an interactive map, which is saved as combinedmap.html. This visualization tool assists stakeholders in making informed decisions by providing an interactive view of logistical data and routing information. The obtained clustering results (Figures 6a and 6b) reveal a meaningful spatial and functional segmentation of Astana's urban structure. The first cluster corresponds to the central business districts characterized by high object density and a prevalence of office and service-oriented enterprises, which require high-frequency, time-sensitive logistics. The second cluster represents residential and mixed-use areas with a dominance of retail and consumer services, reflecting moderate and predictable delivery demand. The third cluster is associated with peripheral industrial and logistics zones, where warehouses and large-scale facilities are concentrated, prioritizing distance and cost efficiency. The alignment between spatial clusters and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

business types confirms that the applied clustering methods capture the functional-economic geography of the city rather than performing purely geometric partitioning. Figure 6 (d) will provide a visual representation of data clustering using K-means with centroids and a route constructed using Dijkstra's algorithm on an interactive map.



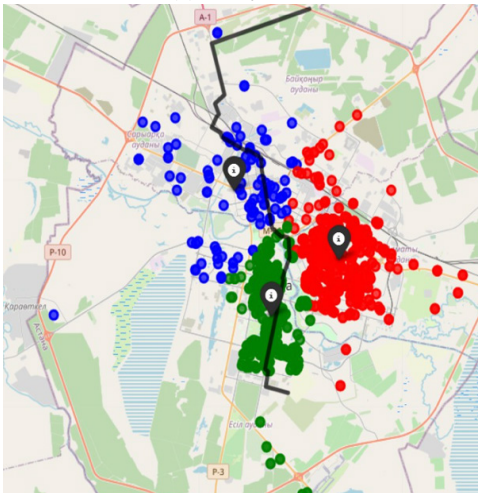
(a) K-means;



(b) Fuzzy C-means;



(c) Dijkstra's Algorithm;



(d) K-means and Dijkstra's algorithm.

Fig. 6. Algorithms on the Map

Conclusion.

The primary objective of the paper was to enhance the logistics efficiency in Astana by developing and implementing a ML model. This objective was accomplished through the application of K-means clustering and Dijkstra's calculation, which together optimized courses and progressed the administration of logistics. The inquiry start-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ed with a comprehensive audit of the vital part and affect of transportation logistics clusters in supply chain administration. This included an examination of the centrality of logistics clusters, their affect on industry flow and territorial advancement, and the challenges confronted in transport logistics operations. The part of mechanical clusters in cultivating little trade development and productivity, especially within the setting of Astana city. To address the distinguished challenges, different strategies for making transport foundation choices were assessed, counting the mechanics of K-means and Fuzzy C-means clustering calculations and the application of chart hypothesis and Dijkstra's calculation. The dataset for this work was collected, including geological arranges, commerce classifications, and area situations of different companies in Astana. This information shaped the establishment for creating the ML models. The comes about area point by point the information visualization procedures utilized, counting the utilize of the 'folium' library to make intuitively maps, and the usage of K-means and Fuzzy C-means clustering calculations to recognize ideal areas for logistics offices. Dijkstra's calculation was utilized to optimize courses inside the urban environment, progressing the productivity of logistics operations. The application of these calculations illustrated an outstanding enhancement in logistics operations, contributing to Astana's improvement as a smart city. The use of K-means clustering improved asset arrangement over the city, whereas Dijkstra's calculation made strides course arranging, subsequently decreasing travel times and operational costs. Along with a technical implementation, the research also concerned the real applicability of smart logistics structures in real-life urban environments. In the case of the developed models, it is possible to adjust them to different city sizes; also, they can be applied in other cities with similar logistical peculiarities. Such flexibility assures wider applicability and sustainability of the intended solutions. One more remarkable input of the study is related to its multidisciplinary contributions, as the theory is based on the intersection of urban planning, artificial intelligence, and transportation engineering solutions to the complex logistic problems. The outcome of the research can be used by stakeholders such as municipal officials, logistic firms, data analysts who want to achieve effective and green strategies in city-freight transportation. Another benefit of incorporating machine learning algorithms in logistics is its paving the way to environmentally-sound decision-making, i.e., lowering the level of carbon emissions generated by optimizing routes and clustering-based consolidation of the delivery services. Moreover, interactive maps created using geospatial visualization tools, e.g., Folium, do not only improve spatial distribution understanding but also facilitate adequate decisions by the city government. They allow the better use of the resources by enabling them to locate areas with a high demand, bottlenecks, and underutilized areas with the assistance of these maps. Lastly this study highlights the role of smart city infrastructure in the realisation of United Nations Sustainable Development Goals (SDGs) especially in the areas of sustainable cities and communities, innovation, and climate action. As the proposed methodology would address both the efficiency of operation and the sustainability of environment, this will make Astana an example of how the digitalization of urban logistic processes should be organized

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

REFERENCES

Barabási A.L. & Oltvai Z.N. (2004). Network biology: Understanding the cell's functional organization // Nature Reviews Genetic. — Vol. 5. — Pp. 101–113. https://doi.org/10.1038/nrg1272 [in Eng.].

Bezdek J.C., Ehrlich R. & Full W. (2020). FCM: The fuzzy c-means clustering.Computers & G eosciences. — Vol. 10. — Pp. 191–203. https://doi.org/10.1016/0098-3004(84)90020-7 [in Eng.].

Bradley P. S. & Fayyad U.M. (1998). Refining initial points for K-means clustering. — Proceedings of the 15th International Conference on Machine Learning. Pp. 91–99. https://doi.org/10.5555/645527.657466 [in Eng.].

Bouhmala N. (2016). How good is the Euclidean distance metric for the clustering problem. — 2016. 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). Pp. 312–315. https://doi.org/10.1109/II-AI-AAI.2016.26. [in Eng.].

Christopher M. & Peck H. (2020). Building the resilient supply chain // *International Journal of Logistics Management*. — Vol. 31. — Pp. 1–14. https://doi.org/10.1108/09574090410700275 [in Eng.].

Cohen S. & Kietzmann J. (2014). Ride on! Mobility business models for the sharing economy. — Organization & Environment. — Vol. 27. — Pp. 279–296. https://doi.org/10.1177/1086026614546199 [in Eng.].

Cormen T.H., Leiserson C.E., Rivest R.L. & Stein C. (2019). Introduction to algorithms. — MIT Press. [in Eng.].

Dijkstra E. W. (1959). A note on two problems in connexion with graphs. — Numerische Mathematik. — Vol. 1. — Pp. 269–271. https://doi.org/10.1007/BF01386390 [in Eng.].

Dunn J. C. (2019). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. — Journal of Cybernetics. — Vol. 3. — Pp. 32–57. https://doi.org/10.1080/01969727308546046 [in Eng.].

Glaeser E.L. & Kahn M.E. (2019). The greenness of cities: Carbon dioxide emissions and urban development. — Journal of Urban Economics. — Vol. 110. — Pp. 404–418. DOI: https://doi.org/10.1016/j.jue.2009.11.006 [in Eng.].

Harary F. (2018). Graph theory. — CRC Press. https://doi.org/10.1201/9780429493768 [in Eng.].

Jain A.K. (2019). Data clustering: 50 years beyond K-means // Pattern Recognition Letters. — Vol. 31. — Pp. 651–666. https://doi.org/10.1016/j.patrec.2009.09.011 [in Eng.].

Lee D.C. (2006). Proof of a modified Dijkstra's algorithm for computing shortest bundle delay in networks with deterministically time-varying links. — IEEE Communications Letters. — Vol. 10. — Pp. 734–736. https://doi.org/10.1109/LCOMM.2006.051982 [in Eng.].

Miyamoto S. & Mukaidono M. (1997). Fuzzy c-means as a regularization and maximum entropy approach. — In: Proceedings of the 7th Fuzzy System Association World Congress, Prague, Czech Republic. — Vol. 2. — Pp. 86–92 [in Eng.].

Pal N.R. & Bezdek J. C. (1995). On cluster validity for the fuzzy c-means model // IEEE Transactions on Fuzzy Systems. — Vol. 3. — Pp. 370–379. https://doi.org/10.1109/91.413225 [in Eng.].

Yu C. & Xu F. (2024). Research on travel route recommendation algorithm based on graph neural network // 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA). Pp. 855–859. https://doi.org/10.1109/ICIPCA61593.2024.10709327 [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

IINTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 26–47
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.002

INTEGRATING FUZZY LOGIC AND ARTIFICIAL INTELLIGENCE IN OPTIMIZING BUSINESS PROCESS AUTOMATION DECISIONS

L. Kurmangaziyeva[1], Sh. Kodanova[2*], M. Urazgaliyeva[3], O. Findik[4], S. Iskakova[2]
[1]Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[2]S. Utebayev Atyrau Oil and Gas University, Atyrau, Kazakhstan;
[3]West Kazakhstan Innovative and Technological University, Uralsk, Kazakhstan;
[4]Karabuk University, Karabuk, Turkiye.
E-mail: kodanova_s@mail.ru

Lyailya Kurmangaziyeva — Candidate of Technical Sciences, Professor, Department of Software Engineering, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan https://orcid.org/0000000306407306;
Shynar Kodanova — Candidate of Technical Sciences, Associate Professor, Facuty of Information Technology, Safi Utebayev Atyrau Oil and Gas University, Atyrau, Kazakhstan
E-mail: kodanova_s@mail.ru, https://orcid.org/0000-0002-1589-4268;
Meiramgul Urazgaliyeva — Master of Technical Sciences, West Kazakhstan Innovative and Technological University, Uralsk, Kazakhstan
https://orcid.org/0000-0003-0640-7306;
Oguz Findik — PhD, Department of Computer Engineering Karabuk University, Karabuk, Turkiye
https://orcid.org/0000-0001-5069-6470;
Sandugash Iskakova — Candidate of Technical Sciences, Professor, Department of Software Engineering, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-6589-854X.

© L. Kurmangaziyeva, Sh. Kodanova, O. Findik, S. Iskakova, M. Urazgaliyeva

Abstract. The paper examines the application of artificial intelligence methods for optimizing business decision-making processes in modern enterprises operating under conditions of uncertainty and digital transformation. The study focuses on the development of an intelligent fuzzy model designed to support the selection of a robotic process automation (RPA) platform for trade and manufacturing companies. The primary objective of the research is to enhance the efficiency, transparency, and justification of managerial decisions during the implementation of RPA technologies in complex

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and multi-criteria environments. The proposed approach is based on fuzzy logic theory and a fuzzy inference mechanism, enabling the formalization of expert knowledge and the integration of qualitative and quantitative evaluation criteria. The model has been implemented in the FuzzyTECH software environment and structured around three key groups of parameters characterizing RPA platforms: functionality, security, and accessibility. Each parameter is represented through linguistic variables and evaluated using a multi-level fuzzy scale. A comparative analysis of five widely used RPA platforms-PIX Robotics, Primo RPA, Robin, Sherpa RPA, and ROOMY bots-was conducted to validate the effectiveness of the developed model. The results demonstrate that the fuzzy model ensures a comprehensive and objective assessment of platform compliance with enterprise requirements while reducing subjectivity in expert evaluation. The study confirms that the proposed decision-support framework can be adapted to various industries and extended by incorporating additional economic or organizational criteria, thus contributing to the advancement of intelligent business process optimization methodologies.

Keywords: artificial intelligence, fuzzy logic, robotic process automation, FuzzyTECH, decision support, business process optimization

For citation: L. Kurmangaziyeva, Sh. Kodanova, M. Urazgaliyeva, O. Findik, S. Iskakova (2026). Integrating fuzzy logic and artificial intelligence in optimizing business process automation decisions // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 26–47. https://doi.org/10.54309/IJICT.2026.25.1.002. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ЖАСАНДЫ ИНТЕЛЛЕКТ ПЕН АЙҚЫН ЕМЕС ЛОГИКАНЫ БІРІКТІРУ АРҚЫЛЫ БИЗНЕС-ПРОЦЕСТЕРДІ АВТОМАТТАНДЫРУ ШЕШІМДЕРІН ОҢТАЙЛАНДЫРУ

Л. Курмангазиева[1], Ш. Қоданова[2]*, М. Уразгалиева[3], О. Findik[4], С. Искакова[2]

[1]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;

[2]Атырау мұнай және газ университеті, Атырау, Қазақстан;

[3]Батыс Қазақстан инновациялық-технологиялық университеті Орал, Қазақстан;

[4]Қарабүк университеті, Қарабүк, Түркия.

E-mail: kodanova_s@mail.ru

Курмангазиева Ляйля — техника ғылымдарының кандидаты, Х. Досмұхамедов атындағы Атырау университетінің «Бағдарламалық инженерия» кафедрасының профессоры

https://orcid.org/0000000306407306;

Коданова Шынар — техника ғылымдарының кандидаты, қауымд. профессор, Ақпараттық технологиялар факультеті, Сафи Өтебаев атындағы Атырау мұнай және газ университеті

E-mail: kodanova_s@mail.ru, https://orcid.org/0000-0002-1589-4268;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Уразгалиева Мейрамгүл — техника ғылымдарының магистрі, Батыс Қазақстан инновациялық-технологиялық университеті
https://orcid.org/0000-0003-0640-7306;
Финдик Оғуз — философия докторы (PhD), Қарабүк университетінің Компьютерлік инженерия кафедрасы
https://orcid.org/0000-0001-5069-6470;
Искакова Сандуғаш — техника ғылымдарының кандидаты, қауымд. профессор, Ақпараттық технологиялар факультеті, Сафи Өтебаев атындағы Атырау мұнай және газ университеті
https://orcid.org/0000-0002-6589-854X.

© Л. Курманғазиева, Ш. Коданова, М. Уразгалиева, О. Финдик, С. Искакова

Аннотация. Бұл жұмыста белгісіздік және цифрлық трансформация жағдайында қызмет ететін заманауи кәсіпорындарда басқарушылық шешімдерді оңтайландыру мақсатында жасанды интеллект әдістерін қолдану қарастырылады. Зерттеу сауда және өндірістік компаниялар үшін роботтандырылған бизнес-процестерді автоматтандыру (RPA) платформасын таңдауды қолдауға арналған интеллектуалды айқын емес модельді әзірлеуге бағытталған. Зерттеудің негізгі мақсаты – күрделі және көпкритерийлі ортада RPA технологияларын енгізу кезінде басқарушылық шешімдердің тиімділігін, ашықтығын және негізділігін арттыру. Ұсынылған тәсіл айқын емес логика теориясына және айқын емес логикалық қорытынды механизміне негізделген, бұл сараптамалық білімді формализациялауға және бағалаудың сапалық және сандық критерийлерін біріктіруге мүмкіндік береді. Модель FuzzyTECH бағдарламалық ортасында жүзеге асырылып, RPA-платформаларды сипаттайтын үш негізгі параметрлер тобы бойынша құрылымдалған: функционалдылық, қауіпсіздік және қолжетімділік. Әрбір параметр лингвистикалық айнымалылар түрінде ұсынылып, көпдеңгейлі айқын емес шкала арқылы бағаланады. Әзірленген модельдің тиімділігін тексеру мақсатында кеңінен қолданылатын бес RPA-платформаға — PIX Robotics, Primo RPA, Robin, Sherpa RPA және ROOMY bots - салыстырмалы талдау жүргізілді. Нәтижелер айқын емес модельдің кәсіпорын талаптарына платформалардың сәйкестігін жан-жақты әрі объективті бағалауға мүмкіндік беретінін және сараптамалық бағалаудағы субъективтілікті төмендететінін көрсетті. Зерттеу нәтижелері ұсынылған шешім қабылдауды қолдау жүйесінің түрлі салаларға бейімделе алатынын және экономикалық немесе ұйымдастырушылық қосымша критерийлерді енгізу арқылы кеңейтілуі мүмкін екенін дәлелдейді, бұл бизнес-процестерді интеллектуалды оңтайландыру әдіснамаларын дамытуға ықпал етеді.

Түйін сөздер: жасанды интеллект, бұлдыр логика, роботтандырылған бизнес-процестерді автоматтандыру, FuzzyTECH, шешім қабылдауды қолдау, бизнес-процестерді оңтайландыру

Дәйексөздер үшін: Л. Курманғазиева, Ш. Коданова, М. Уразгалиева, О. Финдик, С. Искакова (2026). Жасанды интеллект пен айқын емес логиканы

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

біріктіру арқылы бизнес-процестерді автоматтандыру шешімдерін оңтайландыру // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 25. 26–47 бет. https://doi.org/10.54309/IJICT.2026.25.1.002 (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ИНТЕГРАЦИЯ НЕЧЕТКОЙ ЛОГИКИ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ ОПТИМИЗАЦИИ РЕШЕНИЙ ПО АВТОМАТИЗАЦИИ БИЗНЕС-ПРОЦЕССОВ

Л. Курмангазиева[1], Ш. Коданова[2]*, М. Уразгалиева[3], О. Финдик [4], С. Искакова[2]

[1]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан;
[2]Атырауский университет нефти и газа имени С. Утебаева, Атырау, Казахстан;
[3]Западно-Казахстанский инновационно-технологический университет, Уральск, Казахстан;
[4]Карабюкский университет, Карабюк, Турция.
E-mail: kodanova_s@mail.ru

Курмангазиева Ляйля — кандидат технических наук, профессор кафедры «Программная инженерия» Атырауского университета имени Х. Досмухамедова https://orcid.org/0000000306407306;
Коданова Шынар — кандидат технических наук, ассоциированный профессор факультета информационных технологий Атырауского университета нефти и газа имени С. Утебаева, Атырау, Казахстан
E-mail: kodanova_s@mail.ru, https://orcid.org/0000-0002-1589-4268;
Уразгалиева Мейрамгүл — магистр технических наук, Западно-Казахстанский инновационно-технологический университет, Уральск, Казахстан
https://orcid.org/0000-0003-0640-7306;
Финдик Огуз — доктор философии (PhD), кафедра компьютерной инженерии Карабюкского университета, Карабюк, Турция
https://orcid.org/0000-0001-5069-6470;
Искакова Сандугаш — кандидат технических наук, ассоциированный профессор факультета информационных технологий Атырауского университета нефти и газа имени С. Утебаева, Атырау, Казахстан
https://orcid.org/0000-0002-6589-854X.

© Л. Курмангазиева, Ш. Коданова, М. Уразгалиева, О. Финдик, С. Искакова

Аннотация. В работе рассматривается применение методов искусственного интеллекта для оптимизации процессов принятия управленческих решений на современных предприятиях, функционирующих в условиях неопределённости и цифровой трансформации. Исследование сосредоточено на разработке интеллектуальной нечеткой модели, предназначенной для поддержки выбора

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

платформы роботизированной автоматизации бизнес-процессов (RPA) для торговых и производственных компаний. Основная цель исследования заключается в повышении эффективности, прозрачности и обоснованности управленческих решений при внедрении технологий RPA в сложной многокритериальной среде. Предложенный подход основан на теории нечеткой логики и механизме нечеткого логического вывода, что позволяет формализовать экспертные знания и интегрировать качественные и количественные критерии оценки. Модель реализована в программной среде FuzzyTECH и структурирована по трём ключевым группам параметров, характеризующих RPA-платформы: функциональность, безопасность и доступность. Каждый параметр представлен в виде лингвистических переменных и оценивается с использованием многоуровневой нечеткой шкалы. Для проверки эффективности разработанной модели проведён сравнительный анализ пяти широко используемых RPA-платформ- PIX Robotics, Primo RPA, Robin, Sherpa RPA и ROOMY bots. Результаты демонстрируют, что нечеткая модель обеспечивает комплексную и объективную оценку соответствия платформ требованиям предприятия, снижая субъективность экспертных оценок. Исследование подтверждает, что предложенная система поддержки принятия решений может быть адаптирована для различных отраслей и расширена за счёт включения дополнительных экономических и организационных критериев, что способствует развитию интеллектуальных методологий оптимизации бизнес-процессов.

Ключевые слова: искусственный интеллект, нечеткая логика, роботизированная автоматизация бизнес-процессов, FuzzyTECH, принятие решений, оптимизация бизнес-процессов

Для цитирования: Л. Курманғазиева, Ш. Коданова, М. Уразгалиева, О. Финдик, С. Искакова (2026). Интеграция нечеткой логики и искусственного интеллекта при оптимизации решений по автоматизации бизнес-процессов // Международный журнал информационных и коммуникационных технологий. Vol. 7. No. 25. Pp. 26–47. https://doi.org/10.54309/IJICT.2026.25.1.002. (In Eng.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The modern era is characterized by the rapid development of digital technologies, leading to significant changes in the economy, industry, and management. As enterprises digitalize and transition to smart manufacturing, the use of artificial intelligence (AI) and robotic process automation (RPA) technologies is becoming crucial. These technologies enable companies to optimize operations, reduce costs, improve the quality of products and services, and accelerate data-driven management decision-making.

The relevance of this topic stems from the need to improve the efficiency of business processes in the face of increasing competition, external instability, and the need for flexible responses to market changes. Traditional approaches to enterprise management are becoming insufficient, as they do not allow for the rapid processing of large

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

volumes of information or the forecasting of possible development scenarios. Artificial intelligence methods provide tools for analyzing, modeling, and optimizing processes, facilitating the transition of enterprises to a qualitatively new level of digital management.

Robotic automation technologies, which can mimic human actions when working with information systems, are particularly important. Software robots perform routine operations-data processing, reporting, and customer interaction-without human intervention, ensuring accuracy, stability, and high productivity. The use of RPA is especially relevant for retail and manufacturing companies, where the number of repetitive operations is high and the demands on information processing speed are constantly increasing (Zadeh, 1965).

However, when implementing RPA, a key task arises - choosing the optimal platform for robotic automation of business processes. There is a wide range of domestic and foreign solutions that differ in functionality, architecture, cost and degree of integration with existing systems (Mamdani, E.H et al., 1975). Choosing a suitable platform requires a comprehensive assessment of many factors, which is associated with the uncertainty and subjectivity of expert assessments. In these conditions, the use of fuzzy logic and intelligent modeling methods becomes a rational tool (Softline Corporation, 2022).

The use of fuzzy models allows for the consideration of uncertainty, fuzziness, and incompleteness of source data. Such models provide a more flexible and realistic representation of expert knowledge than classical deterministic approaches. Fuzzy modeling makes it possible to formalize expert judgments and translate them into quantitative assessments, which is especially important when analyzing multi-criteria problems, such as selecting an RPA platform (IKS). (Media, 2021).

The aim of this study is to develop an intelligent fuzzy model for selecting a platform for robotization of processes in a trade and manufacturing enterprise (Zimmermann et al., 2001).

FuzzyTECH software environment, which provides extensive capabilities for constructing and analyzing decision support systems, was chosen as the implementation tool.

The scientific novelty of the work lies in the development and testing of an intelligent fuzzy model that provides a quantitative assessment of the platform's compliance with enterprise requirements based on expert and statistical data (Van der Aalst et.al., 2018).

The practical significance of the study lies in the possibility of using the created model to select the optimal RPA platform, as well as to adapt the methodology to the needs of other industries that require decision-making under conditions of uncertainty (Delen, D et.al., 2018).

Thus, the development of an intelligent fuzzy model for selecting a platform for robotic business processes is a relevant scientific and practical task aimed at improving management methods and increasing the efficiency of modern enterprises (Syed, et al., 2020).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Materials and Methods.

Description of a fuzzy model for selecting a platform for robotic automation of processes in a trade and manufacturing enterprise.

We describe the development of a fuzzy model for selecting a platform for robotic process automation at a retail and manufacturing enterprise. Before implementing the model in the FuzzTECH software environment, it is necessary to define the input and output variables, as well as their relationships and sets of fuzzy rules (Van der Aalst et.al., 2018).

Our model will include three sets of variables. The first will assess the platform's compliance with the customer's functional requirements, the second will assess the platform's compliance with security requirements, and the third will assess the platform's availability for robotics.

Let's describe all the variables. The input parameters of the first intermediate variable Y_1, reflecting the platform's functionality, will be three linguistic variables:

X_1 - the ability of software robots to interact not only with basic web and desktop applications, but also with external business systems, the availability of tools for building complex processes from various robots.

X_2 - operating systems supported by the platform (multisystem), supported programming languages, supported DBMS.

X_3 - functionality available: Optical Character Recognition (OCR), speech synthesis and recognition, availability of Low - Code and No - Code programming.

The input parameters of the second intermediate variable Y_2, which evaluates compliance with the safety requirements of the robotic platform, will be 2 linguistic variables:

X_4 - the presence of protection against unauthorized access, methods of verification and control of changes in accordance with the role model, the ability to manage rights for robots, workstations, users, actions, the presence of an audit of user actions, the presence of a password storage;

X_5 - presence in the Register of domestic software, compliance with Law "On personal data".

The input parameters of the third intermediate variable Y_3, responsible for accessibility, will be the following linguistic variables:

X_6 - availability of technical support, community, training, trial period, demo version of the platform;

X_7 - number of clients, number of implemented robots, presence of awards and prizes from thematic competitions.

the term sets of input linguistic variables X_1-X_7 as sets T_1-T_7= {« Low » (low), « Medium » (medium), « High » (high)}.

The variable term sets will be scored as follows. $X1_{will}$, $X2_{will}$, $X3_{will}$, $X4_{will}$, $X5_{will}$, $X6_{will}$, $X7_{will}$ be scored from 0 to 10, where 0 to 3 is poor performance, 4 to 7 is average performance, and 8 to 10 is excellent performance.

The terms of the intermediate and output variables will have a rating from 0 to 10,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

where values from 0 to 3 indicate low compliance with customer requirements, from 3 to 6 indicate moderate compliance with customer requirements, and from 7 to 10 indicate high compliance with customer requirements.

Let us compile a list of heuristic rules for intermediate variables Y_1, Y_2, Y_3, with the help of which the final output variable Y will be calculated.

For the variable $Y_{1,}$ 27 rules were compiled, presented in Figure 1.

Номер правила	IF X1	AND X2	AND X3	THEN Y1
1.	LOW	LOW	LOW	LOW
2.	LOW	LOW	MEDIUM	LOW
3.	LOW	LOW	HIGH	MEDIUM
4.	LOW	MEDIUM	LOW	LOW
5.	LOW	MEDIUM	MEDIUM	MEDIUM
6.	LOW	MEDIUM	HIGH	MEDIUM
7.	LOW	HIGH	LOW	MEDIUM
8.	LOW	HIGH	MEDIUM	MEDIUM
9.	LOW	HIGH	HIGH	MEDIUM
10.	MEDIUM	LOW	LOW	LOW
11.	MEDIUM	LOW	MEDIUM	MEDIUM
12.	MEDIUM	LOW	HIGH	MEDIUM
13.	MEDIUM	MEDIUM	LOW	MEDIUM
14.	MEDIUM	MEDIUM	MEDIUM	MEDIUM
15.	MEDIUM	MEDIUM	HIGH	HIGH
16.	MEDIUM	HIGH	LOW	MEDIUM
17.	MEDIUM	HIGH	MEDIUM	MEDIUM
18.	MEDIUM	HIGH	HIGH	HIGH
19.	HIGH	LOW	LOW	MEDIUM
20.	HIGH	LOW	MEDIUM	MEDIUM
21.	HIGH	LOW	HIGH	MEDIUM
22.	HIGH	MEDIUM	LOW	MEDIUM
23.	HIGH	MEDIUM	MEDIUM	MEDIUM
24.	HIGH	MEDIUM	HIGH	HIGH
25.	HIGH	HIGH	LOW	MEDIUM
26.	HIGH	HIGH	MEDIUM	HIGH
27.	HIGH	HIGH	HIGH	HIGH

Fig. 1. Rules for variable Y 1

For the variable Y_2, nine rules have been created, presented in Figure 2

Номер правила	IF X4	AND X5	THEN Y2
1.	LOW	LOW	LOW
2.	LOW	MEDIUM	MEDIUM
3.	LOW	HIGH	MEDIUM
4.	MEDIUM	LOW	LOW
5.	MEDIUM	MEDIUM	MEDIUM
6.	MEDIUM	HIGH	HIGH
7.	HIGH	LOW	LOW
8.	HIGH	MEDIUM	MEDIUM
9.	HIGH	HIGH	HIGH

Fig. 2. Rules for variable Y 1

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

For variable Y_3 the following rules will apply, as shown in Figure 3.

Номер правила	IF X6	AND X7	THEN Y3
1.	LOW	LOW	LOW
2.	LOW	MEDIUM	MEDIUM
3.	LOW	HIGH	MEDIUM
4.	MEDIUM	LOW	LOW
5.	MEDIUM	MEDIUM	MEDIUM
6.	MEDIUM	HIGH	HIGH
7.	HIGH	LOW	MEDIUM
8.	HIGH	MEDIUM	HIGH
9.	HIGH	HIGH	HIGH

Fig. 3. Rules for variable Y_3

For the variable Y, 27 rules were formulated, presented in Figure 4.

Номер правила	IF Y1	AND Y2	AND Y3	THEN Y
1.	LOW	LOW	LOW	LOW
2.	LOW	LOW	MEDIUM	LOW
3.	LOW	LOW	HIGH	LOW
4.	LOW	MEDIUM	LOW	LOW
5.	LOW	MEDIUM	MEDIUM	MEDIUM
6.	LOW	MEDIUM	HIGH	MEDIUM
7.	LOW	HIGH	LOW	MEDIUM
8.	LOW	HIGH	MEDIUM	MEDIUM
9.	LOW	HIGH	HIGH	MEDIUM
10.	MEDIUM	LOW	LOW	LOW
11.	MEDIUM	LOW	MEDIUM	LOW
12.	MEDIUM	LOW	HIGH	LOW
13.	MEDIUM	MEDIUM	LOW	MEDIUM
14.	MEDIUM	MEDIUM	MEDIUM	MEDIUM
15.	MEDIUM	MEDIUM	HIGH	MEDIUM
16.	MEDIUM	HIGH	LOW	MEDIUM
17.	MEDIUM	HIGH	MEDIUM	HIGH
18.	MEDIUM	HIGH	HIGH	HIGH
19.	HIGH	LOW	LOW	LOW
20.	HIGH	LOW	MEDIUM	LOW
21.	HIGH	LOW	HIGH	LOW
22.	HIGH	MEDIUM	LOW	MEDIUM
23.	HIGH	MEDIUM	MEDIUM	MEDIUM
24.	HIGH	MEDIUM	HIGH	MEDIUM
25.	HIGH	HIGH	LOW	MEDIUM
26.	HIGH	HIGH	MEDIUM	HIGH
27.	HIGH	HIGH	HIGH	HIGH

Fig. 4. Rules for variable Y

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Thus, blocks of rules were prescribed, all linguistic variables were described, and heuristic rules were compiled to create a fuzzy model for assessing the quality of software product implementation.

Implementation of a Fuzzy Model in the FUZZYTECH Environment

We'll walk through the step-by-step creation and configuration of a fuzzy model for robotic platform selection in FuzzyTECH. To do this, we'll create six input variables, three intermediate variables, and one output variable.

As a result of setting up the above actions, we obtain the final fuzzy model for selecting a robotics platform (Figure 5).

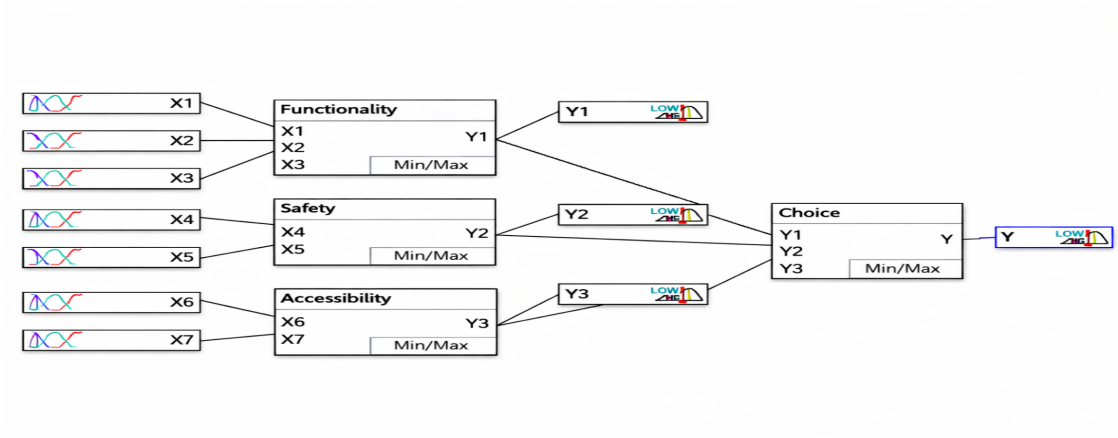


Fig. 5. Fuzzy model of the quality level of software product implementation

Next, we set up membership graphs for each variable of our fuzzy model as shown in Figures 33–39.

The next step is to set up rule blocks for intermediate and output variables. To do this, go to the rule editor for each block and use the rule block filling function. Afterwards, we check the correctness of the resulting rules and obtain the following rule blocks (see Figures 6–10).

Spreadsheet Rule Editor - Functionality

#	IF X1	X2	X3	THEN DoS	Y1
1	low	low	low	1.00	low
2	low	low	medium	1.00	low
3	low	medium	high	1.00	medium
4	low	high	low	1.00	medium
5	low	high	medium	1.00	low
6	medium	low	low	1.00	medium
7	medium	low	medium	1.00	medium
8	medium	medium	high	1.00	low
9	medium	medium	low	1.00	medium
10	medium	medium	medium	1.00	medium
11	medium	high	low	1.00	low
12	medium	low	medium	1.00	medium
13	medium	high	high	1.00	medium
14	medium	low	low	1.00	medium
15	medium	medium	medium	1.00	high
16	medium	high	high	1.00	medium
17	medium	low	low	1.00	high
18	medium	high	medium	1.00	medium
19	high	high	low	1.00	medium
20	high	low	medium	1.00	high
21	high	medium	high	1.00	medium
22	high	high	low	1.00	medium
23	high	high	medium	1.00	high
24	high	high	high	1.00	medium
25	high	high	low	1.00	medium
26	high	high	medium	1.00	high
27	high	high	high	1.00	high

Fig. 6. Block of rules for variable Y_1

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Spreadsheet Rule Editor - Time

#	IF X4	X5	THEN DoS	Y2
1	low	low	1.00	low
2	low	medium	1.00	medium
3	low	high	1.00	medium
4	medium	low	1.00	low
5	medium	medium	1.00	medium
6	medium	high	1.00	high
7	high	low	1.00	low
8	high	medium	1.00	medium
9	high	high	1.00	high

Fig. 7. Block of rules for variable Y_2

Spreadsheet Rule Editor - Cost

#	IF X6	X7	THEN DoS	Y3
1	low	low	1.00	low
2	low	medium	1.00	medium
3	low	high	1.00	medium
4	medium	low	1.00	low
5	medium	medium	1.00	medium
6	medium	high	1.00	high
7	high	low	1.00	medium
8	high	medium	1.00	high
9	high	high	1.00	high
9	high	high	1.00	high

Fig. 8. Block of rules for variable Y_3

Spreadsheet Rule Editor - Quality

#	IF [Delete All Rules] Y1	Y2	Y3	THEN DoS	Y
1	low	low	medium	1.00	low
2	low	low	medium	1.00	low
3	low	low	high	1.00	low
4	low	medium	low	1.00	medium
5	low	medium	medium	1.00	medium
6	low	high	high	1.00	medium
7	low	high	low	1.00	medium
8	low	high	medium	1.00	low
9	medium	low	low	1.00	low
10	medium	low	medium	1.00	low
11	medium	low	high	1.00	medium
12	medium	medium	low	1.00	medium
13	medium	high	medium	1.00	medium
14	medium	high	high	1.00	medium
15	high	high	low	1.00	low
16	high	low	medium	1.00	medium
17	high	low	high	1.00	medium
18	high	medium	low	1.00	high
19	high	high	medium	1.00	high
20	high	high	low	1.00	medium
21	high	low	medium	1.00	high
22	high	low	high	1.00	high
23	high	medium	high	1.00	medium
24	high	high	low	1.00	high
25	high	high	medium	1.00	high
26	high	low	high	1.00	high
27	high	low	high	1.00	high
27	high	low	high	1.00	high
28	high	medium	high	1.00	high
27	high	high	high	1.00	high

Fig. 9. Block of rules for variable Y

Next, we will switch to debug mode to check the functionality of our fuzzy model (Figure 5).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

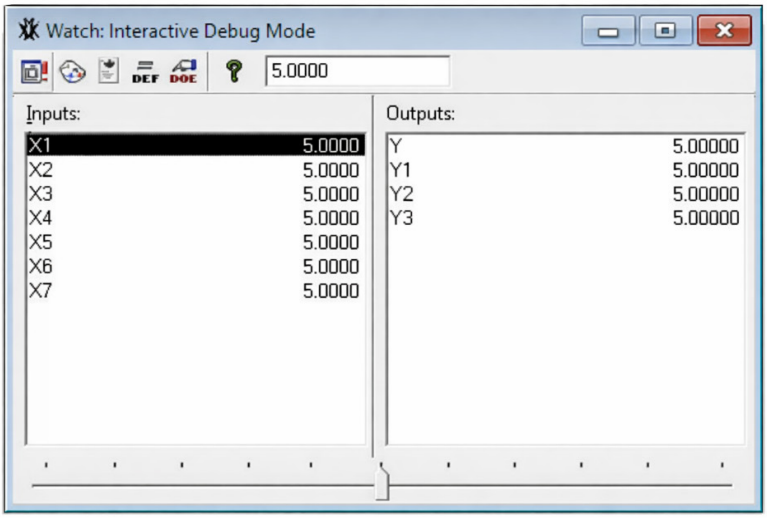


Fig. 10. Interactive debugging window

Next, we'll open all windows in the program's workspace to fully observe the dependencies of all variables. This creates a convenient workspace for configuring and testing the model (Figure 11).



Fig. 11. Fuzzy model in the FuzzyTECH environment

Thus, in this section, a fuzzy model for selecting a platform for robotics was im-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

plemented in the FuzzyTECH software environment.

Interpretation and analysis of the obtained results of the model for selecting a platform for robotization of processes in a trade and manufacturing enterprise, we will proceed to its use for fuzzy modeling (Leonenkov, 2024)

We will enter the PIX Robotics platform data based on the data on the official website and the platform presentation, and conduct an assessment of the level of compliance with customer requirements.

– the ability of software robots to interact not only with basic web and desktop applications, but also with external business systems, the availability of tools for building complex processes from various robots (X_1) has a value of 9 out of 10;

– operating systems supported by the platform (multi-system), supported programming languages, supported DBMS (X_2) has a value of 8 out of 10;

– availability of functionality: Optical Character Recognition (OCR), speech synthesis and recognition, the presence of Low - Code and No - Code programming (X_3) has a value of 9 out of 10;

– the presence of protection against unauthorized access, methods of verification and control of changes in accordance with the role model, the ability to manage rights for robots, workstations, users, actions, the presence of an audit of user actions, the presence of a password storage (X_4) has a value of 8 out of 10;

– presence in the Register of domestic software, compliance with Law “On personal data” (X_5) has a value of 10 out of 10;

– availability of technical support, community, training, trial period, demo version of the platform (X_6) has a value of 10 out of 10;

– the number of clients, the number of implemented robots, the presence of awards and prizes from thematic competitions (X_7) has a value of 9 out of 10.

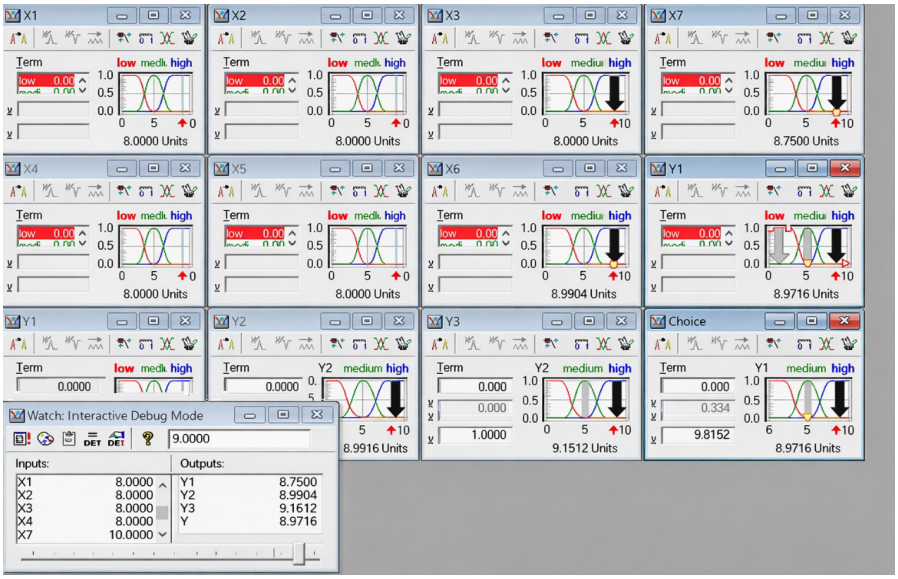


Fig. 12. Fuzzy model in the FuzzyTECH environment
Robotics Platform Customer Compliance Level

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

After entering all input variables into our model, we obtain the next level of compliance with customer requirements (see Figure 12).

For the PIX Robotics platform, the model produced a score of 8.9716.

Let's repeat the previously completed steps for the remaining robotization platforms.

For the Primo RPA platform:

X_1 has a value of 10 out of 10;
X_2 has a value of 10 out of 10;
X_3 has a value of 9 out of 10;
X_4 has a value of 10 out of 10;
X_5 has a value of 10 out of 10;
X_6 has a value of 10 out of 10;
X_7 has a value of 9 out of 10;

After entering all input variables into our model, we obtain the next level of compliance with customer requirements (see Figure 13).

Robin platform:

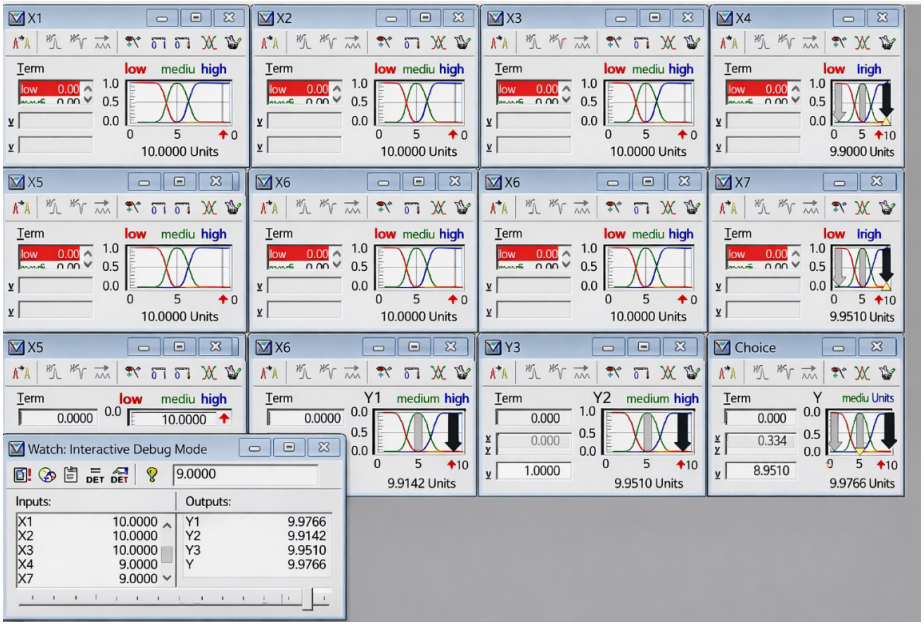


Fig. 13. Primo RPA Platform Customer Compliance Level

The model came up with a score of 9.9510.

X_1 has a value of 10 out of 10;
X_2 has a value of 9 out of 10;
X_3 has a value of 10 out of 10;
X_4 has a value of 9 out of 10;
X_5 has a value of 10 out of 10;
X_6 has a value of 9 out of 10;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

X_7 has a value of 10 out of 10;

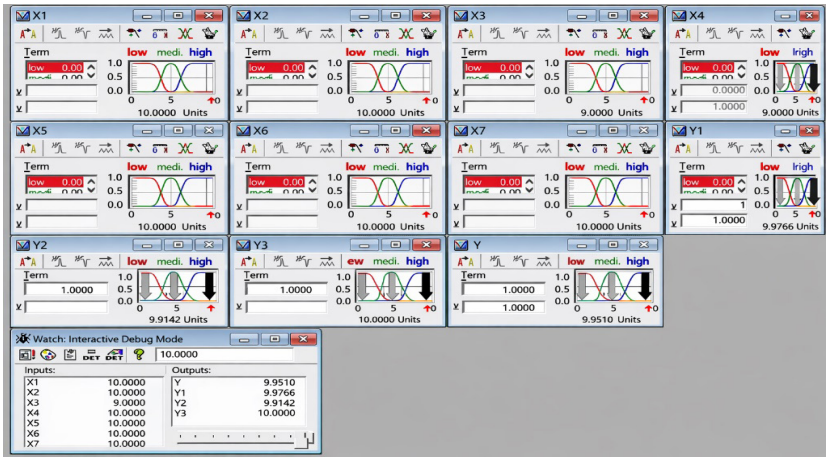


Fig. 14. Robin Platform Customer Compliance Level

After entering all input variables into our model, we obtain the next level of compliance with customer requirements (see Figure 14).

The model came up with a score of 9.9510.

For the Sherpa RPA platform:

X_1 has a value of 8 out of 10;

X_2 has a value of 8 out of 10;

X_3 has a value of 9 out of 10;

X_4 has a value of 9 out of 10;

X_5 has a value of 10 out of 10;

X_6 has a value of 10 out of 10;

X_7 has a value of 7 out of 10;



Fig. 15. Sherpa RPA Platform Customer Compliance Level

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

After entering all input variables into our model, we obtain the next level of compliance with customer requirements (see Figure 15).

The model came up with a score of 8.8718.

For the ROOMY bots platform:

X_1 has a value of 7 out of 10;

X_2 has a value of 7 out of 10;

X_3 has a value of 8 out of 10;

X_4 has a value of 8 out of 10;

X_5 has a value of 10 out of 10;

X_6 has a value of 10 out of 10;

X_7 has a value of 8 out of 10;



Fig. 16. Level of compliance with customer requirements of the ROOMY bots platform

After entering all input variables into our model, we obtain the next level of compliance with customer requirements (see Figure 12).

The model came up with a score of 8.6776.

As can be seen, the highest level of compliance with the customer's requirements by the model was 9.9510 for the Robin and Primo RPA platforms. The choice of one of the two selected vendors is recommended to be made through a tender.

Thus, fuzzy modeling was carried out when introducing input variables into the developed fuzzy model for selecting a platform for robotizing the processes of a trade and manufacturing enterprise.

Results and discussion

The study is based on the application of modern artificial intelligence and fuzzy modeling techniques aimed at optimizing the process of selecting a platform for robotic process automation in a retail and manufacturing enterprise. The main idea is to use intelligent technologies to create a tool that will enable objective evaluation and compari-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

son of various software solutions, taking into account the uncertainty and heterogeneity of the initial data. By Lotfi was used as the theoretical basis for the study. Zadeh. This theory allows for the description of complex processes in which quantitative assessment is impossible or difficult, and decisions are based on expert judgment and subjective criteria. Fuzzy modeling allows for the use of not only precise numerical values but also qualitative characteristics, expressed as words such as “low,” “medium,” and “high.” This makes the method particularly suitable for solving management problems involving uncertainty.

The subject of this study is the process of selecting a platform for robotic automation of enterprise business processes. This choice is always associated with a multitude of factors—from functionality and security to licensing costs, technical support, and compatibility with other systems. Traditional evaluation methods often use a strict quantitative scale that fails to capture all the nuances. Therefore, this paper utilizes a fuzzy logic approach, which allows for a more flexible and realistic approach to the problem.

The study analyzed leading robotic automation platforms: PIX Robotics, Primo RPA, Robin, Sherpa RPA, and ROOMY bots. For comparison, key criteria were identified, grouped into three key areas:

1. Functionality, which reflects the platform’s ability to perform a wide range of operations, integrate with other systems, and support modern technologies such as speech recognition, OCR, and Low -Code tools.

2. Security, including data protection measures, access rights management and compliance with personal data laws.

3. Accessibility, which characterizes the availability of technical support, training materials, demo versions, as well as the popularity and recognition of the platform in the market.

Each criterion was assessed on a linguistic scale, with the values “low,” “medium,” and “high” corresponding to a range from 0 to 10. This scale allowed experts to provide a more flexible assessment that reflected their actual perception of the platforms’ quality.

FuzzyTECH software environment, which is designed for creating and visualizing fuzzy inference systems, was used to build the model. This tool was chosen due to its user-friendly interface, support for multi-block structures, and the ability to clearly define membership functions, which is especially important when analyzing complex dependencies between variables (Büyüközkan, et al., 2018).

The model was developed in several stages. First, the main criteria were defined and linguistic variables were formed. Then, for each group of factors (functionality, security, accessibility), fuzzy inference rules were developed, linking the input variables to intermediate indicators. All intermediate assessments were then combined into a final variable reflecting the overall level of platform compliance with enterprise requirements.

Each design stage was accompanied by expert analysis and tuning of membership functions, which improved the accuracy and reliability of the model. Implementation in the FuzzyTECH environment allowed for visual observation of the results generation process and system debugging in an interactive mode.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The result of this stage was the development of an intelligent fuzzy model that enables quantitative and qualitative assessment of the level of compliance of RPA platforms with customer requirements. The model takes into account the subjectivity of human judgment, combines it with formal criteria, and enables a comprehensive comparative analysis of software solutions under conditions of uncertainty.

Thus, the developed methodology combines scientific rigor and practical applicability. It can be used not only for selecting robotic platforms but also for solving similar problems in other areas requiring decision-making based on multi-criteria assessment.

The practical part of the study focused on implementing an intelligent fuzzy model for selecting a platform for robotic process automation and evaluating its effectiveness in management decision-making. The model was built in the FuzzyTECH environment, which allows for the creation of fuzzy inference systems, visualization of relationships between variables, and interactive debugging.

In the first stage, a model structure was developed, comprising three conceptual blocks: functionality, security, and availability. Each represented a set of parameters characterizing various aspects of RPA platform operation. The functionality block included indicators such as the availability of integration tools with external systems, support for modern technologies, and the level of versatility of software solutions. The security block reflected data protection mechanisms, access control, and regulatory compliance. The availability block considered the availability of technical support, documentation, training courses, and open user communities.

Each parameter was assessed by experts on a ten-point scale, ranging from minimum to maximum compliance with customer requirements. This data was entered into the FuzzyTECH system, where each variable was assigned linguistic term values: “low,” “medium,” and “high.” Based on the expert assessments, heuristic rules were formed linking the input and intermediate variables.

After building the model, a series of experiments were conducted evaluating five modern RPA platforms: PIX Robotics, Primo RPA, Robin, Sherpa RPA, and ROOMY bots. Each was analyzed for functionality, reliability, scalability, and ease of implementation in a production environment.

The simulation results showed that the Robin and Primo RPA platforms demonstrated the best alignment with enterprise requirements, receiving equally high overall scores. Both platforms featured extensive integration capabilities with external business systems, low-code development support, and a well-developed technical support system. The PIX Robotics platform placed in the middle, demonstrating strong functionality but inferior accessibility and customization flexibility. The Sherpa RPA and ROOMY bots solutions demonstrated consistent but less pronounced results, primarily due to the limited number of implementations and lower level of automation of supporting processes.

The analysis revealed an important conclusion: the effectiveness of an RPA platform is determined not only by its technological capabilities but also by a combination of organizational factors, such as the availability of support, training resources, and an active professional community. Therefore, the choice of platform should be based on a

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

balance between functionality, security, and ease of implementation.

The results confirmed the validity of the developed model. The FuzzyTECH system accurately reflected the relationships between criteria and demonstrated stability when varying the initial parameters. The use of fuzzy modeling mitigated the subjectivity of expert assessments and produced quantifiable results that can be used in management decision-making.

A distinctive feature of the proposed model is its versatility and adaptability. If necessary, it can be supplemented with new criteria—for example, cost efficiency, payback period, or compatibility with corporate infrastructure. This makes the model a flexible tool that can be applied in retail, manufacturing, financial, or logistics organizations.

Thus, the study results demonstrated that fuzzy modeling is an effective tool for analyzing and selecting technological solutions under uncertainty. The model, built in the FuzzyTECH environment, demonstrated a high degree of reliability and practical applicability. It can serve as a basis for developing decision support systems for digital transformation of enterprises and business process optimization.

The study examined the theoretical and practical aspects of using artificial intelligence methods to optimize enterprise business decisions. The focus was on developing an intelligent fuzzy model for selecting a platform for robotic process automation in a retail and manufacturing enterprise.

The first part of the paper analyzes the main development trends in artificial intelligence and robotic process automation (RPA) technology. It demonstrates that the implementation of such technologies is an important element of the digital transformation of enterprises and can significantly improve management efficiency, reduce costs, and speed up routine operations (Business Architecture Institute, 2020).

Based on an analysis of existing solutions and scientific approaches, a methodology for constructing a fuzzy model was developed, enabling multi-criteria evaluation of RPA platforms under uncertainty. The use of fuzzy logic allowed for the formalization of expert knowledge and the consideration of subjective factors that are difficult to quantify.

FuzzyTECH software environment, an intelligent model was built, including three main blocks of variables: functionality, security, and availability. Each block combined key platform selection criteria and reflected various aspects of its application. Based on the entered expert data, an assessment was conducted of five modern RPA platforms - PIX Robotics, Primo RPA, Robin, Sherpa RPA, and ROOMY bots (Kumar et al., 2021).

Primo RPA and Robin platforms demonstrated the highest levels of compliance with customer requirements, receiving the highest overall scores. These solutions are characterized by flexibility, extensive integration capabilities, advanced support, and modern automation tools. The remaining platforms showed good, but somewhat more limited, results (Aguirre et.al., (2017).

The developed fuzzy model has proven its effectiveness as a decision support tool. It ensures transparent analysis, allows for the consideration of multiple, diverse factors, and generates objective recommendations for selecting the optimal software solution.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Furthermore, the model can be expanded by adding new criteria-economic, technological, and organizational-making it universal and adaptable to various industries.

The main findings of the study can be summarized as follows:

1. The use of artificial intelligence and fuzzy logic methods allows us to effectively solve problems of selecting software solutions under conditions of uncertainty.

2. FuzzyTECH environment ensures visual modeling and high accuracy of computational experiments.

3. The developed model for selecting an RPA platform has proven its practical applicability and can be used to substantiate management decisions during the digitalization of an enterprise.

The practical significance of the work lies in the fact that the proposed methodology can be implemented at enterprises in various industries to assess the quality of the implementation of digital technologies and select optimal solutions for process automation.

Thus, all the stated goals and objectives of the study were achieved, and the results confirmed the effectiveness of using fuzzy methods to support decision-making when selecting platforms for robotic automation of business processes.

Comparison of fuzzy modeling results with classical methods and model sensitivity analysis.

To verify the robustness and validity of the obtained results, the fuzzy modeling results were compared with those obtained using a classical multicriteria evaluation method. Classical decision-making methods in multicriteria choice problems are typically based on the additive aggregation of normalized indicators. Therefore, a classical additive criteria aggregation model was used to compare the results.

Since the developed fuzzy model includes three main criterion groups-functionality (X_1- X_3), safety (X4-X5), and availability (X_6- X_7)-the classical evaluation was conducted using a hierarchical scheme. First, partial scores were calculated for each criterion group, after which they were aggregated into a final indicator. The classical evaluation function has the following form:

$$Y_{class} = \frac{1}{3}\left(\frac{X_1 + X_2 + X_3}{3} + \frac{X_4 + X_5}{2} + \frac{X_6 + X_7}{2}\right)$$

where X_1 –X_7 are the evaluation criteria used in the developed model.

Based on the initial data used in the fuzzy modeling, final assessment values were calculated for the five RPA platforms studied. The comparison results are presented in the table1.

Table 1 – Comparison of results of fuzzy modeling and classical estimation

Platform	The result of the fuzzy model Y	Yclass rating
PIX Robotics	8,9716	9,0556
Primo RPA	9,9510	9,7222
Robin	9,9510	9,5556
Sherpa RPA	8,8718	8,7778
ROOMY bots	8,6776	

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The results show that the ranking of alternatives is consistent across both approaches. In both cases, the Primo RPA and Robin platforms demonstrate the highest level of compliance with enterprise requirements, followed by PIX Robotics, Sherpa RPA, and ROOMY bots. This confirms the validity and robustness of the developed fuzzy model.

At the same time, certain differences are observed in the numerical values of the final assessments. The fuzzy model provides a more pronounced differentiation of alternatives, as it takes into account nonlinear relationships between criteria and expert inference rules. Unlike the classical additive method, which assumes a linear relationship between indicators, fuzzy modeling allows for more flexible consideration of the uncertainty of the initial data and expert knowledge.

a sensitivity analysis of the model was conducted, the purpose of which was to determine the influence of individual input criteria on the final result Y. The analysis was carried out by assessing the change in the final value when varying the input parameters of the model.

The analysis results showed that the X_1 and X_2 criteria, which characterize the platform's functional and integration capabilities, have the greatest impact on the final score. These parameters reflect the ability of software robots to interact with various information systems, support various operating systems, programming languages, and database management systems, which is a key factor when choosing a robotic platform.

The X_7 criterion also has a significant impact, reflecting the platform's market penetration, the number of completed projects, and professional recognition. This indicator indirectly characterizes the reliability and maturity of the technological solution.

Criteria X_3 and X_4 have a moderate impact on the final result. They reflect the presence of advanced functionality (e.g., OCR technologies, speech recognition, and low - code tools) and the level of information security of the system.

Criterion X_6, related to the availability of technical support, training materials, and a user community, demonstrates relatively low sensitivity within the sample under study. This is explained by the fact that most of the platforms studied already have a high level of user support, which reduces the differences between alternatives.

Finally, the X_5 criterion, which characterizes compliance with regulatory requirements and the presence of a software product in the domestic software registry, has virtually no impact on the differentiation of results in this experiment. This is because this indicator has a uniformly high value for all platforms under consideration.

Based on the sensitivity analysis, the relative degree of influence of the criteria on the final result can be determined:

$$X_2 > X_1 > X_7 > X_4 > X_3 > X_6 > X_5$$

Thus, the comparison with classical methods and sensitivity analysis confirm the stability of the developed fuzzy model and its practical applicability for the tasks of selecting business process robotization platforms under conditions of multi-criteria evaluation and uncertainty of the initial information.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Conclusion.

This study presented the development and implementation of an intelligent fuzzy model for selecting a robotic process automation (RPA) platform in a retail and manufacturing enterprise. The proposed approach integrates artificial intelligence methods and fuzzy logic to support decision-making under conditions of uncertainty and multi-criteria evaluation. The model was implemented in the FuzzyTECH environment and tested on five modern RPA platforms.

The results demonstrated that the application of fuzzy modeling provides a transparent and quantitatively justified assessment of platform compliance with enterprise requirements. Primo RPA and Robin platforms achieved the highest overall compliance scores, confirming the effectiveness of the developed methodology. The proposed model reduces the subjectivity of expert judgments and enhances the reliability of managerial decisions.

The developed approach has practical value and can be adapted to other industries requiring digital transformation and process automation. Future research may focus on expanding the model by incorporating economic efficiency indicators, implementation costs, and long-term performance metrics to further improve decision support mechanisms.

REFERENCES

Aguirre S., Rodriguez A. (2017). Automation of a Business Process Using Robotic Process Automation (RPA): A Case Study. IEEE Latin America Transactions. IEEE. -Issue 12. — Vol. 15. — Pp. 2332–2339. https://doi.org/10.1109/TLA.2017.8127507 [In Eng.].

Büyüközkan, G., Göçer, F. (2018). Digital Supply Chain: Literature Review and a Proposed Framework for Future Research. Computers in Industry. - Elsevier. -Issue 97. — Vol. 97. — Pp. 157-177. https://doi.org/10.1016/j.compind.2018.02.010 [In Eng.].

Delen D., Zolbanin H.M. (2018). The Analytics Paradigm in Business Research // *Journal of Business Research.* Elsevier. Issue 90. — Vol. 90. — Pp. 186. https://doi.org/10.1016/j.jbusres.2018.05.013 [In Eng.].

Van der Aalst, W.M.P., Bichler, M., Heinzl, A. (2018). Robotic Process Automation. Business & Information Systems Engineering. Springer. Issue 4. — Vol. 60. — Pp. 269–272. https://doi.org/10.1007/s12599-018-0542-4 [In Eng.].

Kumar, A., Yadav, S., et al. (2021). Decision Support System Based on Fuzzy Logic for Industry Applications. Applied Soft Computing. - Elsevier. - Issue 108. — Vol. 108. — Article 107442. https://doi.org/10.1016/j.asoc.2021.107442 [In Eng.].

Lacity M., Willcocks L. (2016). Robotic Process Automation at Telefónica O2. MIS Quarterly Executive. MISQ. Issue 1. — Vol. 15. — Pp. 21–35. https://doi.org/10.17705/2msqe.00002 [In Eng.].

Mamdani, E.H., Assilian, S. (1975). An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller // *International Journal of Man-Machine Studies.* Elsevier. Issue 1. — Vol. 7. — Pp. 1–13. https://doi.org/10.1016/S0020-7373(75)80002-2 [In Eng.].

Syed R., Suriadi S., Adams M., Bandara W., et al. (2020). Robotic Process Automation: Contemporary Themes and Challenges. Computers in Industry. Elsevier. Issue 115. — Vol. 115. Article 103162. https://doi.org/10.1016/j.compind.2019.103162 [In Eng.].

Zadeh L.A. (1965). Fuzzy Sets. Information and Control. — Publisher: Elsevier. Issue 3. — Vol. 8. — Pp. 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X [In Eng.].

Zimmermann H.J. (2001). Fuzzy Set Theory-and Its Applications. 4th ed. — New York. - Springer. https://doi.org/10.1007/978-94-010-0646-0 [In Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 48–62
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.003

ORGANIZATION OF AN ONLINE SURVEY OF PARTICIPANTS IN THE EDUCATIONAL PROCESS AND ANALYSIS OF THE RESULTS BASED ON THE MODIFIED DELPHI METHOD

Y. Mailybayev[1*], U. Adilbayeva[2], R. Amanova[3]
[1]International University of Transportation and Humanities, Almaty, Kazakhstan;
[2]ALT University, Almaty, Kazakhstan;
[3]International Information Technology University, Almaty, Kazakhstan.
E-mail: ersaiyn.kurmanbaiuly@mtgu.edu.kz

Yersaiyn Mailybayev — PhD, Associate Professor, Department of Computer Technology and Telecommunications, International University of Transport and Humanities, Almaty, Kazakhstan
E-mail: ersaiyn.kurmanbaiuly@mtgu.edu.kz. https://orcid.org/0000-0002-1977-3690;
Ulzhalgas Adilbayeva — PhD, Associate Professor, Department of Language Education, ALT University, Almaty, Kazakhstan
https://orcid.org/0000-0003-4976-4178;
Raihan Amanova — PhD student, Lecturer, Department of Information Systems, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0000-8969-582X.

© Y. Mailybayev, U. Adilbayeva, R. Amanova

Abstract. In the context of the digital transformation of education, an urgent task is to search for effective tools for collecting and analyzing the opinions of participants in the educational process to make informed management decisions. Traditional survey methods often require significant time and material costs, and the data obtained may lack completeness and reliability. This study proposes an effective way to organize online surveys based on the modified Delphi method, integrated with decision support systems and cloud platforms. The aim of the work is to develop and test a model for collecting and harmonizing expert assessments to analyze the current state of the educational process. The methodological basis was the modified Delphi method, which differs from the classical version by using interval assessments, allowing for the consideration of uncertainty and the level of confidence of experts. The Google Forms platform was chosen for data collection due to its accessibility, ease of integration, and the ability to automatically export results. Data processing and calculation of statistical indicators

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

were performed automatically using developed mathematical and information software. The expert group included 12 specialists: teachers, methodologists, IT specialists, and independent experts in the field of educational quality assessment. The study was conducted in two rounds. In the first round, experts assessed five key indicators, which made it possible to identify the initial scatter of opinions. After familiarizing themselves with the aggregated results of the first round, the second round observed a significant narrowing of the assessment intervals and an increase in the consensus level above 0.75 for all indicators, confirming the effectiveness of feedback. The final integral assessment of the effectiveness of the educational process was 7.73 on a 10-point scale. The proposed model allows not only to quickly collect data but also to minimize subjectivity through a multi-stage coordination procedure. Automating data collection and processing through the integration of Google Forms and a decision support system significantly reduces the survey time and increases the clarity of results for decision-makers. The modified Delphi method in combination with online tools is an effective and accessible tool for monitoring, forecasting development, and optimizing management in educational institutions, and can also be adapted for other subject areas.

Keywords: online survey, modified Delphi method, expert assessment, decision support system, management of the educational process

For citation: Y. Mailybayev, U. Adilbayeva, R. Amanova (2026)**.** Organization of an online survey of participants in the educational process and analysis of the results based on the modified delphi method // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 48–62. https://doi.org/10.54309/IJICT.2026.25.1.003. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ҰЙЫМДАСТЫРЫЛҒАН ОНЛАЙН САУАЛНАМА АРҚЫЛЫ БІЛІМ БЕРУ ПРОЦЕСІНЕ ҚАТЫСУШЫЛАРДЫҢ ПІКІРЛЕРІН ЖИНАУ ЖӘНЕ НӘТИЖЕЛЕРІН МОДИФИКАЦИЯЛАНҒАН ДЕЛЬФИ ӘДІСІ НЕГІЗІНДЕ ТАЛДАУ

Е. Майлыбаев[1*], У. Адилбаева[2], Р. Аманова[3]

[1]Халықаралық көліктік-гуманитарлық университеті, Алматы, Қазақстан;
[2]ALT University, Алматы, Қазақстан;
[3]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: ersaiyn.kurmanbaiuly@mtgu.edu.kz

Майлыбаев Ерсайын — PhD, Халықаралық көліктік-гуманитарлық университетінің «Компьютерлік технологиялар және телекоммуникациялар» кафедрасының қауымдастырылған профессоры, Алматы, Қазақстан
E-mail: ersaiyn.kurmanbaiuly@mtgu.edu.kz. https://orcid.org/0000-0002-1977-3690;
Адилбаева Ұлжалғас — PhD, ALT University «Language Education» кафедрасының қауымдастырылған профессоры, Алматы, Қазақстан

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

https://orcid.org/0000-0003-4976-4178;
Аманова Райхан — PhD докторант, Халықаралық ақпараттық технологиялар университетінің «Ақпараттық жүйелер» кафедрасының оқытушысы, Алматы, Қазақстан
https://orcid.org/0009-0000-8969-582X.

© Е. Майлыбаев, Ұ. Адилбаева, Р. Аманова

Аннотация. Білім беруді цифрландыру жағдайында басқарушылық шешімдер қабылдау үшін білім беру процесіне қатысушылардың пікірлерін жинау мен талдаудың тиімді құралдарын іздестіру өзекті мәселе болып табылады. Дәстүрлі сауалнама әдістері көбінесе айтарлықтай уақыттық және материалдық шығындарды талап етеді, ал алынған деректердің толықтығы мен дұрыстығы жеткіліксіз болуы мүмкін. Зерттеуде шешім қабылдауды қолдау жүйелерімен және бұлтты платформалармен интеграцияланған модификацияланған Дельфи әдісі негізінде онлайн-сауалнамаларды ұйымдастырудың тиімді тәсілі ұсынылған. Жұмыстың мақсаты білім беру процесінің ағымдағы жағдайын талдау үшін сараптамалық бағалауларды жинау мен келісудің моделін әзірлеу және сынақтан өткізу болып табылады. Әдіснамалық база ретінде классикалық нұсқадан интервалдық бағалауды қолдануымен ерекшеленетін модификацияланған Дельфи әдісі алынғандықтан сарапшылардың белгісіздігі мен сенімділік деңгейін ескеруге мүмкіндік туады. Қолжетімділік, интеграцияның қарапайымдылығы және нәтижелерді автоматты түрде экспорттау мүмкіндігі болғандықтан деректерді жинау үшін Google Forms платформасы таңдалды. Деректерді өңдеу және статистикалық көрсеткіштерді есептеу, әзірленген математикалық және ақпараттық қамтамасыз етуді қолдана отырып, автоматты түрде жүргізілді. Сараптамалық топқа оқытушылар, әдіскерлер, IT-мамандары және білім сапасын бағалау саласындағы тәуелсіз сарапшылардан тұратын 12 маман кірді. Зерттеу екі турдан құралды. Бірінші турда сарапшылар бес негізгі көрсеткішті бағалап, пікірлердің бастапқы алшақтығын анықтады. Бірінші турдың жиынтық нәтижелерімен танысқаннан кейін, екінші турда барлық көрсеткіштер бойынша бағалау интервалдарының айтарлықтай тарылғаны және консенсус деңгейінің 0.75-тен жоғары өскені байқалып, кері байланыстың тиімділігі расталды. Білім беру процесінің тиімділігінің қорытынды интегралдық бағасы 10 балдық шкала бойынша 7.73 құрады. Ұсынылған модель деректерді жедел жинап қана қоймай, көп сатылы келісу процедурасы арқылы субъективтілікті азайтуға мүмкіндік береді. Google Forms пен шешім қабылдауды қолдау жүйесінің интеграциясы арқылы деректерді жинау мен өңдеуді автоматтандыру сауалнама жүргізу уақытын едәуір қысқартады және шешім қабылдаушы тұлғалар үшін нәтижелердің көрнекілігін арттырады. Модификацияланған Дельфи әдісі онлайн-құралдармен үйлескен түрде, білім беру мекемелерінде мониторинг, даму бағытын болжау және басқаруды оңтайландырудың тиімді әрі қолжетімді құралы болып табылады, сонымен қатар зерттеу шешімін басқа да пәндік салаларға бейімдеуге болады.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Түйінді сөздер: онлайн сауалнама, модификацияланған Дельфи әдісі, сараптамалық бағалау, шешім қабылдауды қолдау жүйесі, білім беру процесін басқару

Дәйексөздер үшін: Е. Майлыбаев, Ұ. Адилбаева, Р. Аманова (2026). Ұйымдастырылған онлайн сауалнама арқылы білім беру процесіне қатысушылардың пікірлерін жинау және нәтижелерін модификацияланған дельфи әдісі негізінде талдау // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 25. Б. 48–62. https://doi.org/10.54309/IJICT.2026.25.1.003 (Қаз. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

СБОР МНЕНИЙ УЧАСТНИКОВ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПОСРЕДСТВОМ ОРГАНИЗОВАННОГО ОНЛАЙН-АНКЕТИРОВАНИЯ И АНАЛИЗ РЕЗУЛЬТАТОВ НА ОСНОВЕ МОДИФИЦИРОВАННОГО МЕТОДА ДЕЛЬФИ

Е. Майлыбаев[1*], У. Адилбаева[2], Р. Аманова[3]

[1]Международный транспортно-гуманитарный университет, Алматы, Казахстан;

[2]ALT University, Алматы, Казахстан;

[3]Международный университет информационных технологий, Алматы, Казахстан.

E-mail: ersaiyn.kurmanbaiuly@mtgu.edu.kz

Майлыбаев Ерсайын — PhD, ассоциированный профессор, кафедра «Компьютерные технологии и телекоммуникации», Международный транспортно-гуманитарный университет, Алматы, Казахстан
E-mail: ersaiyn.kurmanbaiuly@mtgu.edu.kz, https://orcid.org/0000-0002-1977-3690;
Адилбаева Улжалгас — PhD, ассоциированный профессор, кафедра «Language Education», ALT University, Алматы, Казахстан
https://orcid.org/0000-0003-4976-4178;
Аманова Райхан — докторант PhD, преподаватель, кафедра «Информационные системы», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0009-0000-8969-582X.

© Е. Майлыбаев, У. Адилбаева, Р. Аманова

Аннотация. В условиях цифровой трансформации образования актуальной задачей является поиск эффективных инструментов для сбора и анализа мнений участников образовательного процесса с целью принятия обоснованных управленческих решений. Традиционные методы опроса часто требуют значительных временных и материальных затрат, а полученные данные могут

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

обладать недостаточной полнотой и достоверностью. Данное исследование предлагает эффективный способ организации онлайн-опросов на основе модифицированного метода Дельфи, интегрированного с системами поддержки принятия решений и облачными платформами. Целью работы является разработка и апробация модели сбора и согласования экспертных оценок для анализа текущего состояния образовательного процесса. В качестве методологической базы выступил модифицированный метод Дельфи, отличающийся от классического аналога использованием интервальных оценок, позволяющих учитывать неопределенность и уровень уверенности экспертов. Для сбора данных была выбрана платформа Google Forms благодаря ее доступности, простоте интеграции и возможности автоматизированного экспорта результатов. Обработка данных и расчет статистических показателей производились автоматически с использованием разработанного математического и информационного обеспечения. В экспертную группу вошли 12 специалистов: преподаватели, методисты, IT-специалисты и независимые эксперты в области оценки качества образования. Исследование проводилось в два тура. В первом туре эксперты оценили пять ключевых показателей, что позволило выявить первоначальный разброс мнений. После ознакомления с агрегированными результатами первого тура во втором туре наблюдалось значительное сужение интервалов оценок и рост уровня консенсуса по всем показателям выше 0.75, что подтверждает эффективность обратной связи. Итоговая интегральная оценка эффективности образовательного процесса составила 7.73 по 10-балльной шкале. Предложенная модель позволяет не только оперативно собирать данные, но и минимизировать субъективность за счет многоступенчатой процедуры согласования. Автоматизация сбора и обработки данных через интеграцию Google Forms и системы поддержки принятия решений существенно сокращает время проведения опроса и повышает наглядность результатов для лиц, принимающих решения. Модифицированный метод Дельфи в сочетании с онлайн-инструментами является эффективным и доступным инструментом для мониторинга, прогнозирования развития и оптимизации управления в образовательных учреждениях, а также может быть адаптирован для других предметных областей.

Ключевые слова: онлайн-опрос, модифицированный метод Дельфи, экспертная оценка; система поддержки принятия решений, управление образовательным процессом

Для цитирования: Е. Майлыбаев, У. Адилбаева, Р. Аманова (2026). Сбор мнений участников образовательного процесса посредством организованного онлайн-анкетирования и анализ результатов на основе модифицированного метода дельфи // Международный журнал информационных и коммуникационных технологий. Vol. 7. No. 25. Pp. 48–62. https://doi.org/10.54309/IJICT.2026.25.1.003 (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Кіріспе.

Қазіргі заманда білім беру жүйесін басқару үдерісі ақпараттық технологиялардың қарқынды дамуына байланысты түбегейлі өзгерістерге ұшырап отыр (Xiong et all., 2025). Оқу орнының сапалы дамуын қамтамасыз ету үшін басқарушылық шешімдерді дер кезінде қабылдау және олардың тиімділігін объективті бағалау маңызды міндеттердің біріне айналды. Осы тұрғыда сарапшылардың пікірлерін жүйелі түрде жинау және талдау білім беру процесін жетілдірудің негізгі құралы болып саналады.

Дәстүрлі сауалнама жүргізу әдістері уақыт пен материалдық ресурстарды көп қажет етеді, ал алынған деректердің толықтығы мен нақтылығы көбіне шектеулі болады. Бұл кемшіліктерді жою мақсатында онлайн сауалнама жүргізу технологиялары кеңінен қолданыла бастады. Онлайн платформалар тек деректерді жинауды жеделдетіп қана қоймай, оларды автоматты түрде өңдеу, визуализациялау және нәтижелерді сараптамалық тұрғыда талдау мүмкіндігін береді.

Мақалада білім беру процесіне қатысушылардың пікірлерін жинау және келісімге келтіру үшін модификацияланған Дельфи әдісін қолданудың өзектілігі қарастырылады. Бұл әдіс сараптамалық бағалаулардың дәлдігін арттыруға, пікірлер арасындағы алшақтықты азайтуға және ортақ шешімге қол жеткізуге мүмкіндік береді.

Сараптамалық бағалауларды жинау мен талдау әдістері басқарушылық шешімдер қабылдау жүйесінде ерекше маңызға ие. Солардың ішінде Дельфи әдісі (Delphi method) — көпқадамды, анонимді сауалнама жүргізу арқылы сарапшылардың пікірлерін келісімге келтіруге арналған ең танымал тәсілдердің бірі. Бұл әдіс алғаш рет 1960-жылдары RAND Corporation зерттеу орталығында әзірленіп, бастапқыда әскери-стратегиялық жоспарлау үшін қолданылған (Dalkey et all., 1963). Кейіннен ол білім беру, денсаулық сақтау, экономика және технологиялық даму салаларында кеңінен қолданыла бастады.

Классикалық Дельфи әдісі бірнеше кезеңнен тұрады, әр кезеңде сарапшылар алдыңғы турдың нәтижелерімен танысып, өз бағаларын қайта қарайды. Roy Schmidt атап өткендей, әдістің басты артықшылығы — сарапшылардың пікірлерін жүйелі түрде жақындату және консенсусқа қол жеткізу (Schmidt., 2007). Өзгерістерді талап ететін, дәстүрлі әдістің кемшіліктері де бар: деректерді жинау мен өңдеудің ұзақтығы, ұйымдастырудың күрделілігі және кейбір жағдайларда сарапшылардың белсенділігінің төмендеуі.

Осы олқылықтарды жою үшін соңғы жылдары көптеген ғалымдар әдісті жетілдіруге күш салды. Дельфи әдісінің электрондық нұсқалары зерттеліп, ақпараттық технологиялармен интеграциялаудың артықшылықтары қарастырылды (Hsu et all., 2007). Fred Woudenberg болса, анонимділіктің сарапшылардың пікірлеріне қысым көрсетуді азайтып, шынайылық деңгейін арттыратынын дәлелдеді (Woudenberg., 1991).

Соңғы жылдары модификацияланған Дельфи әдісін қолдану ауқымы

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

кеңейіп, онлайн платформалармен біріктіру үрдісі қалыптасты (Mello et all., 2025). Шешім қабылдауды қолдау жүйелері (ШҚҚЖ) мен Google Forms сияқты онлайн сауалнама құралдарын біріктіру арқылы деректерді жинау мен талдау үдерісін автоматтандырудың тиімділігін көрсетті (Mailybayev et all., 2021). Бұл тәсіл сарапшылар арасындағы географиялық қашықтық мәселесін жойып, жауап беру жылдамдығын арттырады.

Сонымен қатар, Chitu Okoli және Suzanne Pawlowski білім беру жүйесінде Дельфи әдісін қолданудың ерекшеліктерін зерттеп, оның оқу бағдарламаларын жетілдірудегі және білім сапасын бағалаудағы маңызын атап өтті (Okoli et all., 2004). Gene Rowe және George Wright болса, әдістің болжам жасау қабілетін және стратегиялық жоспарлаудағы рөлін нақтылап көрсетті (Rowe et all., 2011).

Жалпы, әдебиеттер көрсеткендей, модификацияланған Дельфи әдісін заманауи онлайн сауалнама құралдарымен біріктіру сараптамалық бағалаулардың сапасын арттырып, басқарушылық шешімдердің тиімділігін айтарлықтай жоғарылатады.

Әдістер мен материалдар.

Бұл зерттеудің негізінде білім беру процесін басқаруда сараптамалық пікірлерді жинау және талдау үшін модификацияланған Дельфи әдісін (МДӘ) қолдану жатыр. Әдістің негізгі мақсаты – сарапшылардың әртүрлі көзқарастарын жүйелі түрде жинақтап, олардың арасындағы айырмашылықтарды біртіндеп азайтып, ортақ келісімге (консенсусқа) қол жеткізу. Модификацияланған нұсқа дәстүрлі әдістен ерекшеленіп, онлайн сауалнама платформаларымен және шешім қабылдауды қолдау жүйесімен (ШҚҚЖ) интеграцияланған (Сурет-1).

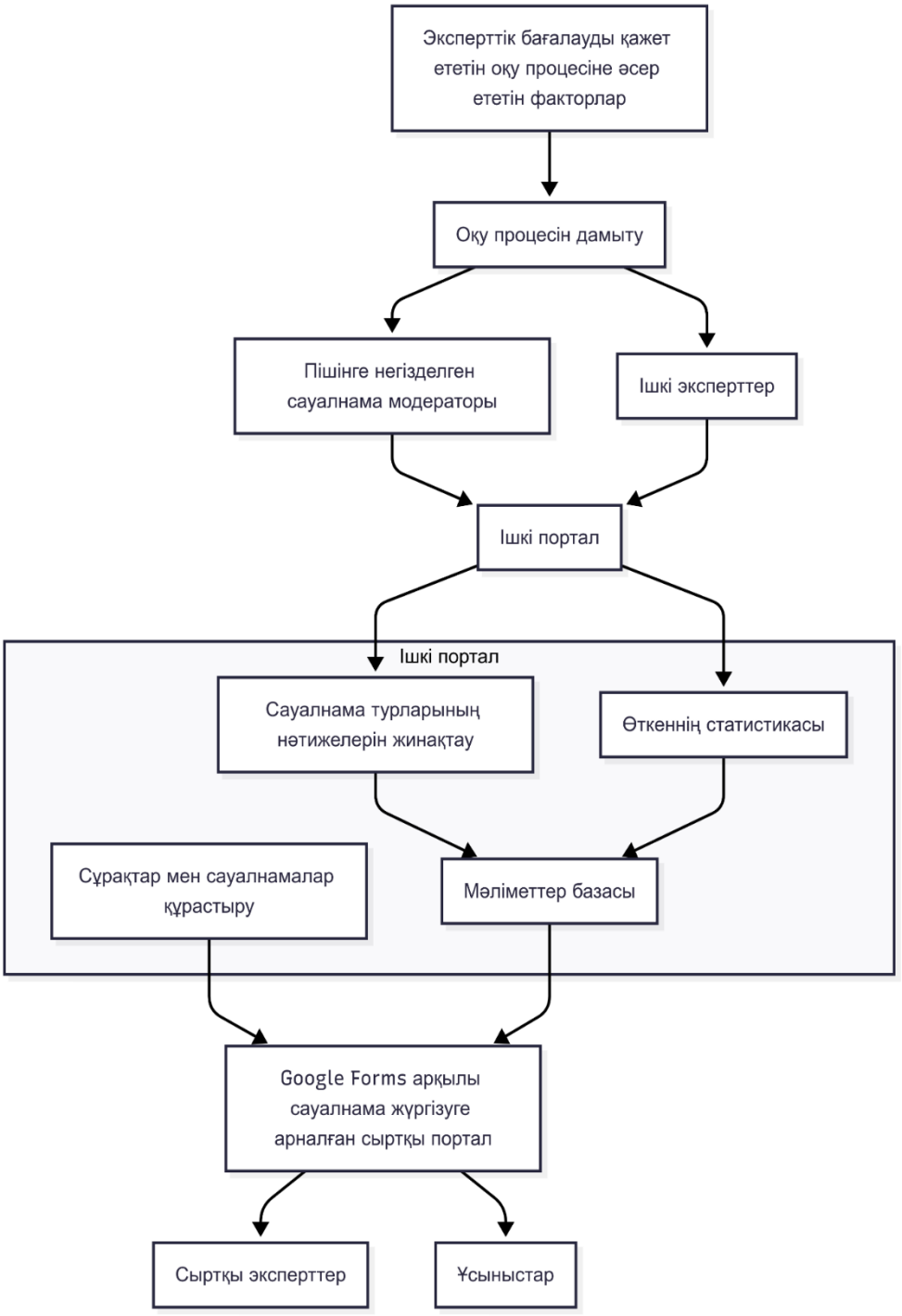
МДӘ қолдану алдында сарапшыларды іріктеу жүргізілді. Сарапшылар құрамына Халықаралық көліктік-гуманитарлық университетінің оқытушылары, оқу процесін басқару бөлімінің қызметкерлері, ақпараттық технологиялар саласының мамандары және білім сапасын бағалау бойынша тәуелсіз сарапшылар кірді. Іріктеу критерийлерінің бірі сарапшының біліктілік көрсеткіші болды, ол келесі формуламен есептелді:

$$E \geq E_{min} \quad (1)$$

мұндағы E – сарапшының жалпы біліктілік көрсеткіші, ал E_min – іріктеуге қатысу үшін қажетті минималды шек. Бұл шарт сарапшылар minқұрамына тек қажетті тәжірибе мен білімге ие тұлғалардың кіруін қамтамасыз етеді. Сарапшылардың нақты біліктілік коэффициенті қосымша түрде келесі формуламен анықталды:

$$K_b = \frac{O_{\text{тәж}}}{O_{max}} \times W_{\text{б}}\text{b} \quad (2)$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Сур.1. Онлайн режимде шешім қабылдауды қолдау жүйелерінің көмегімен білім беру процесінің жұмысын және оны дамыту жолдарын сараптамалық бағалауға арналған платформаның құрылымдық схемасы

мұндағы K_b – сарапшының біліктілік коэффициенті, $0_{\text{тәж}}$ – сарапшының

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

кәсіби тәжірибесінің ұзақтығы (жылмен), O_{max} – сарапшылар тобындағы ең көп тәжірибеге ие қатысушының жұмыс өтілі, ал $W_б$ – сарапшының тәжірибесінің талданатын білім саласына сәйкестік коэффициенті (0-ден 1-ге дейінгі аралықта).

Сауалнама құрастыру кезінде барлық көрсеткіштерге салмақ коэффициенттері (w_j) тағайындалды, бұл әр көрсеткіштің жалпы бағалаудағы үлесін көрсету үшін қажет болды:

$$\sum_{j=1}^{n} w_j = 1 \ (3)$$

мұндағы w_j – j-көрсеткіштің салмағы, ал n – жалпы бағаланатын көрсеткіштердің саны. Бұл тәсіл бағалау кезінде маңызды параметрлердің артық немесе кем есептелуін болдырмауға мүмкіндік береді.

Әр сарапшы бағалауды интервалдық түрде берді (Pankratova et all., 2012: 711–721):

$$I_{ij} = [L_{ij}, U_{ij}] \ (4)$$

мұндағы I_{ij} – i -сарапшының j-көрсеткішке берген бағасы, Lij – бағалаудың төменгі шекарасы, Uij – жоғарғы шекарасы. Мұндай тәсіл сарапшы бағасындағы белгісіздік пен сенімсіздік деңгейін ескеруге мүмкіндік береді.

Әр көрсеткіш бойынша орташа интервал мына формуламен есептелді:

$$\bar{I}_j = \left[\frac{\sum_{i=1}^{m} L_{ij}}{m}, \frac{\sum_{i=1}^{m} U_{ij}}{m}\right] \ (5)$$

мұндағы m – сарапшылар саны. Бұл есептеу барлық сарапшылар пікірін бір интервалға жинақтап, кейінгі талдау үшін бірыңғай шектерді алуға мүмкіндік береді.

Алынған деректерді статистикалық өңдеу кезінде стандартты ауытқу (σ_j) және мәндер диапазоны (R_j) анықталды:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x_j})^2}{m}} \ (6)$$

мұндағы x_{ij} – i-сарапшының нақты мәні, $\bar{x_j}$ – көрсеткіштің орташа мәні. Стандартты ауытқу мәні сарапшылар пікірлерінің бір-бірінен қаншалықты алыс екенін көрсетеді.

Мәндер диапазоны келесі формуламен есептелді:

$$R_j = \max(x_{ij}) - \min(x_{ij}) \ (7)$$

Бұл диапазон сарапшылар пікірлерінің алшақтығын анық көрсетеді. Сарапшылар пікірлерінің үйлесімділік деңгейін сипаттау үшін консенсус деңгейі (КД) есептелді:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

$$КД_j = 1 - \frac{\sigma_j}{R_j} (8)$$

мұндағы КД*j* – *j*-көрсеткіш бойынша келісім деңгейі. Егер КД $_j$ ≥0.7 болса, бұл көрсеткіш бойынша сарапшылар арасында жеткілікті деңгейде келісім бар деп есептеледі (Linstone et all., 1976: 317–318)

Бірінші кейіннен екінші тур жүргізілді. Екі турдың нәтижелері негізінде сарапшылар өз бағаларын қайта қарап, интервалдарын тарылта алды:

$$I_{ij}^{(2)} \subseteq I_{ij}^{(1)} (9)$$

мұндағы $I_{ij}^{(1)}$ – бірінші турдағы интервал, $I_{ij}^{(2)}$ – екінші турдағы жаңартылған интервал. Бұл тәсіл пікірлерді біртіндеп жақындатуға мүмкіндік береді.

Барлық турлар аяқталған соң интегралдық қорытынды баға есептелді:

$$S = \sum_{j=1}^{n} w_j \cdot \bar{x_j} (10)$$

мұндағы S – жалпы интегралдық баға, w_j – көрсеткіштің салмақ коэффициенті, $\bar{x_j}$ – көрсеткіштің орташа мәні. Бұл баға зерттеу нысанының жалпы жағдайын сандық түрде сипаттайды және басқарушылық шешімдер қабылдауға негіз болады (Almaiah et all., 2022).

Сауалнама құрастыру және тарату үшін Google Forms, Microsoft Forms, Yandex Forms, Survey Monkey сынды платформалардың артықшылықтары мен кемшіліктері қарастырылды (Nguyen et all., 2018: 74–79). Аталған платформалардың ішінде, Survey Monkey ақылы, Microsoft Forms шартты ақылы, Yandex Forms платформасында тіркелушілердің аз болдғандығынан бұл платформалар таңдалмады. Google Forms платформасының толық тегін болуы және Android тұтынушыларында Google аккаунттардың бары, сонымен қатар Google Forms кез-келген операциялық жүйеде кедергісіз жұмыс істеуі, аталған платформаны таңдауға негіз болды (Ayandibu., 2025: 411–419). Таңдау барысында Google Forms платформасының артықшылықтары мен кемшіліктері де ескерілді.

Бұл құралдың негізгі артықшылықтары:

Қарапайым әрі интуитивті интерфейс, сауалнаманы тез құрастыру мүмкіндігі;

Құрамында әртүрлі сұрақ түрлері (жабық, көп таңдаулы, Лайкерт шкаласы, ашық сұрақтар) бар;

Автоматты түрде деректерді жинау және кестелік форматта экспорттау мүмкіндігі (Excel, CSV);

Нәтижелерді графикалық түрде көрсету (диаграммалар, гистограммалар);

Респонденттердің анонимдігін сақтау және географиялық орналасуына

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

тәуелсіз қатыстыру.

Google Forms платформасының артықшылықтарымен қатар бірқатар кемшіліктері де бар:

Дизайн мен визуалды бейімдеудің шектеулілігі – сауалнаманың сыртқы көрінісін (түстер, шрифттер, орналасу) кәсіби деңгейде толық өзгерту мүмкіндігі шектеулі;

Күрделі логикалық тармақталудың әлсіздігі – сұрақтар арасындағы шартты қатынастар тек қарапайым деңгейде жүзеге асады, күрделі сценарийлерді құру қиын;

Кеңейтілген статистикалық талдаудың болмауы – Google Forms тек базалық диаграммалар ұсынады, терең статистикалық талдау үшін деректерді басқа бағдарламаларға экспорттау қажет.

Әр сарапшыға Google Forms арқылы бірегей сілтеме жіберілді. Жауаптар автоматты түрде Google Forms бұлттық сақтау жүйесінде жиналып, кейін ШҚҚЖ-ға жүктелді. Бұл интеграция деректерді өңдеу уақытын айтарлықтай қысқартуға мүмкіндік берді.

Нәтижелер және оларды талқылау.

Зерттеу барысында модификацияланған Дельфи әдісі мен онлайн сауалнама платформаларының интеграциясы негізінде білім беру процесінің ағымдағы жағдайы мен даму перспективалары бағаланды. Сарапшылардың пікірлері екі турдан тұратын сауалнама арқылы жиналды.

1. Сарапшылар құрамы және сауалнама статистикасы

Жалпы 12 сарапшы қатысты, оның ішінде 5 — оқытушы, 3 — әдіскер, 2 — IT маманы және 2 — тәуелсіз сарапшы.

Сауалнамаға 12 сарапшының қатысуы зерттеудің мақсаты мен қолданылған сараптамалық бағалау әдісіне сәйкес және келесі факторлармен негізделеді:

Біріншіден, зерттеу жаппай емес, мақсатты, яғни эксперттік сауалнама түрінде жүргізілді. Мұндай жағдайда респонденттердің саны олардың санымен емес, кәсіби құзыреттілігімен және тәжірибесімен айқындалады. Ғылыми-әдістемелік зерттеулерде сарапшылар санының 10–15 адам аралығында болуы алынған нәтижелердің жеткілікті сенімділігі мен репрезентативтілігін қамтамасыз етеді (Mailybayev et all., 2024: 413–420). Екіншіден, сарапшылар құрамы көпсалалы қағида бойынша іріктелген. Оқытушылар педагогикалық аспектілерді, әдіскерлер білім беру процесін ұйымдастыруды, IT мамандары техникалық іске асыруды, ал тәуелсіз сарапшылар объективті сыртқы бағалауды қамтамасыз етеді. Бұл әртүрлі көзқарастарды ескеріп, бағалаудың жан-жақтылығын арттырады. Үшіншіден, сарапшылар санының шектеулі болуы бағалау сапасын тереңдетуге мүмкіндік береді, себебі, әрбір қатысушы сауалнаманы мұқият талдап, дәлелді және саналы жауап береді, бұл формальды жауаптардың ықтималдығын төмендетеді.

Бірінші және екінші турда сауалнамаға қатысқан барлық сарапшылар, қойылған сұрақтардың барлығына толық жауап беру арқылы жауаптардың толықтығын 100% деңгейде қамтамасыз етті.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

2. Бірінші тур нәтижесі бойынша сарапшылар білім беру процесінің 5 негізгі көрсеткішін () бағалады:

LMS жүйесінің тиімділігі ();

Оқу-әдістемелік материалдардың жаңартылу жиілігі ();

Ақпараттық жүйелердің интеграция деңгейі ();

Оқытушылардың цифрлық сауаттылығы ();

Студенттердің қашықтан оқытуға бейімділігі ().

Интервалдық бағалау нәтижелері 1-кестеде берілген.

Кесте 1 – Бірінші турдағы интервалдық бағалау нәтижелері

Көрсеткіш	Орташа төменгі шек	Орташа жоғарғы шек	Орташа мән		Диапазон	КД
LMS тиімділігі	7.2	8.5	7.85	0.48	1.3	0.63
Материалдарды жаңарту	6.8	8.2	7.50	0.42	1.4	0.70
Интеграция деңгейі	6.5	8.0	7.25	0.36	1.5	0.76
Цифрлық сауаттылық	7.0	8.8	7.90	0.50	1.8	0.72
Қашықтан оқытуға бейімділік	6.9	8.6	7.75	0.46	1.7	0.73

Кестеден көріп отырғанымыздай, бірінші турда кейбір көрсеткіштер бойынша (K_1) консенсус деңгейі (КД) 0.7-ден төмен болды, бұл сарапшылар пікірлерінің әлі де толық қалыптаспағанын білдіреді.

Екінші тур нәтижелері бойынша сарапшыларға бірінші турдың қорытындылары ұсынылып, олар өз бағаларын қайта қарастырды. Нәтижесінде бағалау интервалдары тарылды, стандартты ауытқу азайды, ал келісім деңгейі артты.

Кесте 2 – Екінші турдағы интервалдық бағалау нәтижелері

Көрсеткіш	Орташа төменгі шек	Орташа жоғарғы шек	Орташа мән	Стандартты ауытқу	Диапазон	КД
LMS тиімділігі	7.5	8.4	7.95	0.32	0.9	0.82
Материалдарды жаңарту	7.0	8.1	7.55	0.28	1.1	0.75
Интеграция деңгейі	6.9	8.0	7.45	0.25	1.1	0.77
Цифрлық сауаттылық	7.2	8.6	7.90	0.30	1.4	0.79
Қашықтан оқытуға бейімділік	7.1	8.4	7.75	0.27	1.3	0.79

Барлық көрсеткіштер бойынша консенсус деңгейі (КД) 0.75-тен жоғары болды, бұл сарапшылардың екінші турда ортақ көзқарасқа жақындағанын көрсетеді (Chen et all., 2024).

Көрсеткіштердің салмақ коэффициенттері w_j ескеріліп, жалпы интегралдық баға келесі формуламен есептелді:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

$$S = \sum_{j=1}^{n} w_j \cdot \bar{x_j} \ (11)$$

Маңыздылық үлесі: $w_1 = 0.25$, $w_2 = 0.20$, $w_3 = 0.20$, $w_4 = 0.15$, $w_5 = 0.20$.
Есептеу нәтижесі:

$$S = 0.25 \cdot 7.95 + 0.20 \cdot 7.55 + 0.20 \cdot 7.45 + 0.15 \cdot 7.90 + 0.20 \cdot 7.75 = 7.73$$

Бұл мән білім беру процесінің ағымдағы тиімділігін 10 балдық шкала бойынша 7.73 деңгейінде сипаттайды.

Жақсы көрсеткіштерге жеткенмен, таңдалған тәсілдің кемшіліктері де бар екенін ескеру қажет (Bravo-Jaico et all., 2025). Ең алдымен, алынатын нәтижелер сарапшылардың жеке тәжірибесі мен субъективті көзқарасына қатты тәуелді, сондықтан сарапшылар дұрыс іріктелмесе, бағалау объективті болмауы мүмкін. Модификацияланған Дельфи әдісі бірнеше кезеңнен тұратындықтан, келісімге келу процесі кезінде кейбір қатысушылардың белсенділігі төмендеуі ықтимал (Naeem et all., 2025: 4–12). Сонымен қатар, онлайн сауалнама ШҚҚЖ мен Google Forms платформаларына сүйенетіндіктен, интернет сапасына және техникалық жүйелердің тұрақтылығына тәуелді болады, ал бұл кей жағдайда деректердің толық жиналмауына әсер етуі мүмкін. Интервалдық бағалау форматы барлық сарапшыларға бірдей түсінікті бола бермейді, сондықтан кейбір жауаптар дәл болмай қалу қаупі бар. Сондай-ақ автоматтандырылған өңдеу сарапшылардың күрделі әрі контекстке бай пікірлерін толық қамти алмай, сапалық ақпараттың бір бөлігін жоғалтуы мүмкін. ШҚҚЖ мен Google Forms қолданбаған кезде, сарапшылар білім беру процесінің жай-күйін оптимистік түрде бағалап жібереді, кейінен тағайындалған аудит сарапшылардың бағасын көп жағдайда растай бермейді, керсінше аудит нәтижелері ШҚҚЖ мен Google Forms нұсқаларына сәйкес келеді.

Нәтижелер көрсеткендей, модификацияланған Дельфи әдісін онлайн сауалнама құралдарымен біріктіру сараптамалық бағалаулардың нақтылығын арттырды. Бірінші тур мен екінші тур арасындағы салыстыру консенсус деңгейінің айтарлықтай өскенін көрсетеді.

Сонымен қатар, LMS тиімділігі () мен оқу материалдарын жаңарту жиілігі () бойынша пікірлер- дегі ауытқу көбірек қысқарды, бұл деректерді алдын ала талдаудың және сарапшыларға объективті ақпарат ұсынудың тиімділігін көрсетеді.

Google Forms платформасы сауалнаманы онлайн режимде таратуды қамтамасыз етті және қатысушылардың жауап беру белсенділігін арттырды. Бірінші турда сілтеме e-mail және мессенджерлер арқылы жіберілді, нәтижесінде жауап қайтару уақыты орта есеппен 2,3 күнді құрады.

Платформа жинақтаған деректер бірден Excel кестелеріне экспортталып, ШҚҚЖ-ға енгізілді. Бұл тәсіл статистикалық көрсеткіштерді (орташа мән, медиана, стандартты ауытқу, диапазон) автоматты түрде есептеуге және оларды визуалды түрде көрсетуге мүмкіндік берді (Ani et all., 2025: 7484–7497).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Екінші турда сарапшылар Google Forms интерфейсіндегі нәтижелермен танысып, пікірлерін қайта қарады. Платформадағы Branching функциясының арқасында әр сарапшы алдыңғы турдағы өз бағасына және орташа мәндерге байланысты жеке сұрақтар тізімін алды (Anih et all., 2025: 189–198). Бұл жекелеген түзетулер енгізуді жеңілдетіп, консенсус деңгейінің артуына әсер етті.

Қорытынды.

Зерттеу нәтижелері білім беру процесін бағалау мен жетілдіруде модификацияланған Дельфи әдісін онлайн сауалнама құралдарымен, атап айтқанда Google Forms платформасымен біріктірудің жоғары тиімділігін көрсетті. Бұл тәсіл сарапшылардың пікірлерін жедел жинауға, оларды статистикалық тұрғыдан өңдеуге және нәтижелерді визуализациялауға мүмкіндік берді.

Модификацияланған Дельфи әдісі дәстүрлі әдіспен салыстырғанда бірнеше артықшылыққа ие екені анықталды:

Интервалдық бағалау сарапшылардың пікірлерін дәлірек көрсетуге және белгісіздікті азайтуға мүмкіндік берді;

Көптурлы кері байланыс сарапшылар арасындағы пікір алшақтығын қысқартып, келісім деңгейін арттырды;

Онлайн интеграция (Google Forms + ШҚҚЖ) деректерді жинау уақытын қысқартты;

Автоматты есептеу және визуализация нәтижелердің көрнекілігін арттырды және шешім қабылдау процесін жеделдетті.

Алынған нәтижелер көрсеткендей, екінші турда барлық көрсеткіштер бойынша консенсус деңгейі 0.75-тен жоғары болды, бұл сарапшылардың ортақ көзқарасқа жақындағанын дәлелдейді. Жалпы интегралдық баға 10 балдық шкала бойынша 7.73 деңгейінде қалыптасты, бұл білім беру процесінің тиімділігінің жоғары екенін, бірақ жетілдіруді қажет ететін аспектілердің бар екенін көрсетеді.

Практикалық тұрғыдан алғанда, ұсынылған әдіс университеттерде, колледждерде және басқа да білім беру ұйымдарында оқу процесін автоматтандыру, талдау, стратегиялық жоспарлау және басқарушылық шешімдер қабылдау кезінде пайдалануға болады. Болашақ зерттеулерде бұл тәсілді нақты пәндік салаларға бейімдеу және жасанды интеллект құралдарын интеграциялау арқылы сараптамалық бағалауды одан әрі автоматтандыру мүмкіндігі қарастырылады.

REFERENCES

Almaiah M., Hajjej F., Lutfi A., Al-Khasawneh A., Alkhdour T., Almomani O., & Shehab R. (2022). A Conceptual Framework for Determining Quality Requirements for Mobile Learning Applications Using Delphi Method. — Electronics, Basel, Switzerland. — Vol. 11. — Issue 5. Article 788. 10.3390/electronics11050788. [In Eng.].

Ani A., Dalimunthe M., & Daulay E. (2025). The Effectiveness of Creating Assessment Questions Through Google Form as a Digital Learning Assessment Tool // IDEAS: *Journal on English Language Teaching and Learning, Linguistics and Literature*, Palopo, Indonesia. Vol. 13. — No. 2. Pp. 7484–7497. 10.24256/ideas.v13i2.8578. [In Eng.].

Anih A., & Alibo T. (2025). Impact of formative assessment via Google Forms on learning outcomes of undergraduate students in educational technology and media literacy // FUO-Journal of Science Education and Multidisciplinary Research, Bayelsa State. —Nigeria. — Vol. 1. — Issue. 1. Pp. 189–198. 10.5281/zenodo.17633451. [In Eng.].

Ayandibu A. (2025). A comparative analysis of interactive tools in higher education's teaching and learn-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ing: The strengths and weaknesses of Mentimeter, Google forms, Socrative, and Kahoot for playful learning // *International Journal of Research in Business and Social Science.* — Istanbul, Turkey. — Vol. 14. — No. 5. Pp. 411–419. 10.20525/ijrbs.v14i5.3821. [In Eng.].

Bravo-Jaico J., Maquen-Niño G., Germán N., Valdivia C., Alarcón, R., Aquino J., & Serquén, O. (2025). Assessing digital transformation maturity in higher education institutions: a correlational analysis by actors and dimensions // Frontiers in Computer Science, Lausanne, Switzerland. — Vol. 7. Article 1549262. 10.3389/fcomp.2025.1549262. [In Eng.].

Chen A., Sobieraj D., Beckett R., Augustin J., Shah B., & Bechtol R. (2024). Determining Ideal Practices for Student Course Evaluations Using a Modified Delphi Approach // *American Journal of Pharmaceutical Education.* — Alexandria, USA. — Vol. 88. — Issue 12. Article 101330. 10.1016/j.ajpe.2024.101330. [In Eng.].

Dalkey N., Helmer O. (1963). An Experimental Application of the Delphi Method to the Use of Experts // Informs. — Catonsville, USA. — Vol. 9. — No. 3. Pp. 458–467. 10.1287/mnsc.9.3.458. [In Eng.].

Hsu C., Sandford B. (2007). The Delphi Technique: Making Sense Of Consensus // Practical Assessment, Research and Evaluation. — Amherst, USA. — Vol. 2. — No. 10. Pp. 1–8. [In Eng.].

Linstone H., Turoff M. (1976). The Delphi Method: Techniques and Applications // *Journal of Marketing Research.* — Chicago, USA. — Vol. 13. — No. 3. Pp. 317–318. 10.2307/3150755. [In Eng.].

Mello C., Akojie P., & Blake M. (2025). Empowering educators to enhance engagement in a virtual learning environment // *Research in Educational Management.* — Pasuruan, Indonesia. — Vol. 12. — No. 1. Pp. 42–48. 10.2478/rem-2025-0005. [In Eng.].

Mailybayev Y., Umbetov U., Lakhno V., Omarov A., Abuova A., Amanova M., & Sauanova K. (2021). Development of mathematical and information support for solving prediction tasks of a railway station development // *Journal of Theoretical and Applied Information Technology.* — Islamabad, Pakistan. — Vol. 99. — No. 3. Pp. 583–593. [In Eng.].

Mailybayev Y., Shinykulova A., & Syrlybayev Y. (2024). Utilizing information technologies to organize a railway junction survey // *The Bulletin of KazATC.* — Almaty, Kazakhstan. — Vol. 131. — No. 2. Pp. 413–420. 10.52167/1609-1817-2024-131-2-413-420. [In Eng.].

Naeem N., Hadie S., Ismail I., Naeem Z., Khan A., & Yusoff M. (2025). Experts' consensus over key components of online learning environments in medical education: A modified e-Delphi study // *Khyber Medical University Journal.* — Kohat, Pakistan. — Vol. 17. — Issue. 1. Pp. 4–12. 10.35845/kmuj.2025.23743. [In Eng.].

Nguyen H., Stehr E., Eisenreich H., & An T. (2018). Using Google Forms to Inform Teaching Practices // Proceedings of the Interdisciplinary STEM Teaching & Learning Conference. — Statesboro, USA. — Vol. 5. Article 10. Pp. 74–79. 10.20429/stem.2018.020110. [In Eng.].

Okoli, C., & Pawlowski, S. (2004). The Delphi method as a research tool: An example, design considerations and applications // *Information & Management.* — Amsterdam, Netherlands. — Vol. 42. — Issue 1. Pp. 15–29. 10.1016/j.im.2003.11.002. [In Eng.].

Pankratova, N. D., & Malafeeva, L. Y. (2012). Formalizing the consistency of experts' judgments in the Delphi method // *Cybernetics and Systems Analysis.* — New York, USA. — Vol. 48. — No. 5. Pp. 711–721. 10.1007/s10559-012-9451-6. [In Eng.].

Rowe G., Wright G. (2011). The Delphi technique: Past, present, and future prospects // Introduction to the special issue. // Technological Forecasting and Social Change. — Amsterdam, Netherlands. — Vol. 78. — Issue 9. Pp. 1487–1490. 10.1016/j.techfore.2011.09.002. [In Eng.].

Schmidt R. (2007). Managing Delphi Survey Using Nonparametric Statistical Techniques // Decision Sciences, Hoboken, USA. — Vol. 28. — Issue 3. Pp. 763–764. 10.1111/j.1540-5915.1997.tb01330.x. [In Eng.].

Woudenberg F. (1991). An evaluation of Delphi // Technological Forecasting and Social Change, Amsterdam, Netherlands. — Vol. 40. — Issue. 2. Pp. 131–150. 10.1016/0040-1625(91)90002-W. [In Eng.].

Xiong X., & Tsai C. (2025). The Impact of Digital Transformation on Educational Management Models // Interdisciplinary *Academic and Research Journal, Mahasarakham Province.* — Thailand. — Vol. 5. — No. 1. Pp. 825–842. 10.60027/iarj.2025.286903. [In Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 63–77
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.004

SIMULATION-BASED ROBUSTNESS ASSESSMENT OF ASTANA'S BUS NETWORK UNDER RANDOM AND TARGETED FAILURES

V.A. Takizhanov*, A.Z. Ibragimov, A. Shalakhmetov
Astana IT University, Astana, Kazakhstan.
E-mail: ezpigeon.hill@gmail.com

Vyacheslav Takizhanov — Master's student in Computer Science and Engineering, Astana IT University
E-mail: ezpigeon.hill@gmail.com, https://orcid.org/0009-0000-3880-7334;
Aldiyar Ibragimov — PhD student in Computer Science, Astana IT University
https://orcid.org/0000-0002-3697-8647;
Aidarbek Shalakhmetov — PhD student in Computer Science, Astana IT University
https://orcid.org/0009-0001-6779-770X.

© V.A. Takizhanov, A.Z. Ibragimov, A. Shalakhmetov

Abstract. This study investigates the structural robustness of Astana's bus transport network using complex network theory. The network was modeled in L-space based on open 2GIS data, where bus stops represent nodes and route segments represent edges. Key metrics, including the size of the largest connected component, mean shortest path length and mean inverse path length, were analyzed under both random and targeted failure scenarios. The results show that the network maintains high resilience to random disruptions but is highly vulnerable to targeted removals of high-betweenness nodes and bridge edges, which act as critical connectors between major urban areas. Their removal leads to rapid fragmentation and a collapse of global connectivity. The findings highlight the imbalance between local redundancy and global dependency within the city's public transport system and provide a foundation for resilience-oriented planning, emphasizing redundancy reinforcement and diversification of inter-hub connections to ensure uninterrupted service during disruptions.

Keywords: public transport network, complex network theory, structural robustness, network resilience, betweenness centrality, targeted attacks, network vulnerability analysis

For citation: V.A. Takizhanov, A.Z. Ibragimov, A. Shalakhmetov (2026). Simulation-based robustness assessment of Astana's bus network under random and targeted failures // International journal of information and communication technologies. 2026. Vol. 7. No. 25. Pp. 63–77. https://doi.org/10.54309/IJICT.2026.25.1.004. (In Eng.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

МОДЕЛЬДЕУ НЕГІЗІНДЕ АСТАНАНЫҢ АВТОБУС ЖЕЛІСІНІҢ ТҰРАҚТЫЛЫҒЫН БАҒАЛАУ: КЕЗДЕЙСОҚ ЖӘНЕ МАҚСАТТЫ ІСТЕН ШЫҒУЛАР ЖАҒДАЙЫНДА

В.А. Такижанов*, *, А.Ж. Ибрагимов, А. Шалахметов
Astana IT University, Астана, Қазақстан.
E-mail: ezpigeon.hill@gmail.com

Такижанов Вячеслав Александрович — Компьютерлік ғылымдар және инженерия мамандығы бойынша магистрант, Astana IT University
E-mail: ezpigeon.hill@gmail.com, https://orcid.org/0009-0000-3880-7334;
Ибрагимов Алдияр Жаксылыкович — Компьютерлік ғылымдар мамандығы бойынша PhD докторанты, Astana IT University
https://orcid.org/0000-0002-3697-8647;
Шалахметов Айдарбек — Компьютерлік ғылымдар мамандығы бойынша PhD докторанты, Astana IT University
https://orcid.org/0009-0001-6779-770X.

© В.А. Такижанов, А.Ж. Ибрагимов, А. Шалахметов

Аннотация. Бұл зерттеуде Астананың автобус көлік желісінің құрылымдық орнықтылығы күрделі желілер теориясы тұрғысынан талданды. Желі 2GIS платформасының ашық деректері негізінде L-кеңістікте модельденді, мұнда автобус аялдамалары түйіндер ретінде, ал маршрут бөліктері қырлар ретінде қарастырылды. Негізгі көрсеткіштер, ең үлкен байланысқан компоненттің өлшемі, орташа ең қысқа жол және орташа кері жол ұзындығы, кездейсоқ және мақсатты істен шығу сценарийлері жағдайында талданды. Нәтижелер көрсеткендей, желі кездейсоқ бұзылуларға жоғары төзімділік танытқанымен, жоғары делдалдық орталықтылыққа ие түйіндер мен көпірлік қырларды жою кезінде едәуір осал болып келеді. Мұндай элементтердің жойылуы желінің тез фрагментациялануына және жаһандық байланыстылықтың үзілуіне әкеледі. Зерттеу нәтижелері қалалық қоғамдық көлік жүйесіндегі жергілікті артықтық пен ғаламдық тәуелділіктің теңгерімсіздігін айқындай отырып, орнықтылыққа бағытталған жоспарлауға негіз қалайды. Бұл бағытта көпір өткелдеріндегі қайталама байланыстарды күшейту мен торапаралық қосылыстарды әртараптандырудың маңыздылығы атап өтіледі.

Түйін сөздер: қоғамдық көлік желісі, күрделі желілер теориясы, құрылымдық орнықтылық, желінің орнықтылығы, делдалдық орталықтылық, мақсатты шабуылдар, желінің осалдықтарын талдау

Дәйексөздер үшін: В.А. Такижанов, А.Ж. Ибрагимов, А. Шалахметов (2026). Модельдеу негізінде Астананың автобус желісінің тұрақтылығын бағалау: кездейсоқ және мақсатты істен шығулар жағдайында // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 25. 63–77 бет. https://doi.org/10.54309/IJICT.2026.25.1.004. (Ағыл. Тіл.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ОЦЕНКА УСТОЙЧИВОСТИ АВТОБУСНОЙ СЕТИ АСТАНЫ НА ОСНОВЕ МОДЕЛИРОВАНИЯ ПРИ СЛУЧАЙНЫХ И ЦЕЛЕНАПРАВЛЕННЫХ ОТКАЗАХ

В.А. Такижанов*, *, А.Ж. Ибрагимов, А. Шалахметов
Astana IT University, Астана, Казахстан.
E-mail: ezpigeon.hill@gmail.com

Такижанов Вячеслав Александрович — магистрант по направлению "Компьютерные науки и инженерия", Astana IT University
E-mail: ezpigeon.hill@gmail.com, https://orcid.org/0009-0000-3880-7334;
Ибрагимов Алдияр Жаксылыкович — PhD-докторант по направлению "Компьютерные науки", Astana IT University
https://orcid.org/0000-0002-3697-8647;
Шалахметов Айдарбек — PhD-докторант по направлению "Компьютерные науки", Astana IT University
https://orcid.org/0009-0001-6779-770X.

© В.А. Такижанов, А.Ж. Ибрагимов, А. Шалахметов

Аннотация. В настоящем исследовании анализируется структурная устойчивость автобусной транспортной сети Астаны с использованием методов теории сложных сетей. Сеть моделировалась в L-пространстве на основе открытых данных платформы 2GIS, где автобусные остановки представлены в виде узлов, а участки маршрутов в виде рёбер. Основные показатели, включая размер наибольшего связанного компонента, среднюю кратчайшую длину пути и среднюю обратную длину пути, были проанализированы при случайных и целенаправленных сценариях отказов. Результаты показали, что сеть обладает высокой устойчивостью к случайным сбоям, но проявляет значительную уязвимость при целенаправленном удалении узлов с высокой посреднической центральностью и мостовых рёбер, выполняющих функцию ключевых соединителей между основными городскими зонами. Их удаление приводит к быстрой фрагментации сети и потере глобальной связности. Полученные результаты выявляют дисбаланс между локальной избыточностью и глобальной зависимостью в системе общественного транспорта города и создают основу для планирования, ориентированного на повышение устойчивости, с акцентом на усиление дублирующих связей и диверсификацию межузловых соединений для обеспечения непрерывности транспортного обслуживания при сбоях.

Ключевые слова: сеть общественного транспорта, теория сложных сетей, структурная устойчивость, устойчивость сети, посредническая центральность, целенаправленные атаки, анализ уязвимости сети

Для цитирования: В.А. Такижанов, А.Ж. Ибрагимов, А. Шалахметов (2026). Оценка устойчивости автобусной сети Астаны на основе моделирования

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

при случайных и целенаправленных отказах // Международный журнал информационных и коммуникационных технологий. 2026. Т. 7. No. 25. Стр. 63–77. (На анг.). https://doi.org/10.54309/IJICT.2026.25.1.004.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

In the context of rapid urban growth and increased uncertainty in transport demand, the resilience and reliability of public transport become key factors of urban safety and quality of life. The approach of complex network theory has proven effective for quantitatively assessing the connectivity and vulnerability of transport systems: such networks often exhibit robustness to random failures while showing heightened sensitivity to targeted impacts on the most significant elements (Ge et al., 2022). However, for Astana there is no systematic robustness assessment that directly contrasts random failures with targeted disruptions to nodes and edges. To address this gap, we quantify the robustness of Astana's bus network under both random and targeted scenarios and outline planning-relevant implications. We analyze Astana's bus network using stop and route data sourced from 2GIS, within the city's administrative boundaries.

Astana is a young capital with a rapidly changing spatial structure and pronounced monocentricity, divided by the Yesil River into large functional zones and connected by a limited set of bridge crossings that form potential "bottlenecks" for passenger flows. Climatic extremes (a long and cold winter with strong winds) reinforce the requirements for network reliability and route redundancy: even short-term outages of nodes and links during peak periods can lead to cascading overloads of adjacent corridors (Huang et al., 2023). Given the dominance of bus services (in the absence of a fully functioning inner-city rail system), redundancy of transfer alternatives is critical, especially on the arteries crossing the river and in areas of mass residential development on the right and left banks (Jia et al., 2019).

The aim of this study is to quantitatively assess the structural robustness of Astana's bus network and to identify the key elements that determine its reliability under various disruption scenarios. We model the network in the L-space representation: nodes are stops, edges are sequential links along routes; we analyze the network's response to sequential removal of nodes/edges both in random order and in order of decreasing degree and betweenness centrality; we track the dynamics of the share of the largest connected component S(c), the mean and maximum shortest path length, as well as the mean inverse path length, correctly defined for disconnected graphs (Tran et al., 2019). Additionally, we consider scenarios of "cascading effects", wherein the failure of a critical stop leads to the removal of the routes serving it, thereby simulating realistic failure propagation (Li et al., 2025). Details of the materials, metrics and experimental procedures are provided in the "Materials and Methods" section.

The novelty of this study lies in integrating a comprehensive complex-network assessment of Astana's bus network robustness, explicitly contrasting random and tar-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

geted failure scenarios, with a direct translation of degradation curves and identified critical nodes and links into practice-oriented recommendations. This perspective links network indicators to Astana's public transport development strategy, prioritizing uninterrupted service under challenging climatic and spatial conditions. The object of the study is the bus transport network of Astana; the subject is its structural robustness under random and targeted node and edge removals. We hypothesize that the network maintains substantial tolerance to random failures but exhibits pronounced vulnerability to targeted disruptions of high-betweenness transfer nodes and bridge edges.

Materials and methods.

Observations of urban public transport routes show that their paths form a network with a complex structure. This approach makes it possible to view transport systems through the lens of complex network theory, where public transport infrastructure is modeled and analyzed using graph-theoretic tools (Derrible & Kennedy, 2011). Over the past decades, the concept of complex networks has become a central field of research, combining methods from graph theory and statistical physics. Within this framework, a network is defined as a set of nodes connected by edges, allowing interactions to be described in both natural and man-made systems.

Particular attention in complex network studies is paid to nontrivial properties that differ sharply from those of classical random graphs. In particular, it has been found that such networks exhibit the small-world effect, short distances between nodes, a high level of local clustering and a pronounced ability for self-organization (Xiao et al., 2024). These characteristics ensure robustness against random failures while making the system vulnerable to targeted attacks on key elements, a pattern repeatedly observed in real-world transit networks (Cicchini et al., 2024).

Applying this approach to transport systems makes it possible to reveal hidden patterns in their structure, assess connectivity and resilience levels and detect correlations that go beyond random link distributions (Chen et al., 2024; Hassan et al., 2022). Thus, analyzing a bus network as a complex system allows not only for a graph-based structural description but also for conclusions about its reliability and operational robustness.

For this study, a custom Python parser was developed to automatically collect open data from the 2GIS platform. The data were collected in September 2025, meaning that all network topology and route information used in the analysis reflects the actual state of the public transport system at that moment. The parser extracts the complete set of bus stops and routes within the administrative boundaries of Astana, recording the ordered sequence of stops for each route. The resulting data are stored in CSV format to ensure reproducibility and subsequent processing. Based on these data, the transport network is represented as a graph, which enables the application of methods of complex network theory to analyze its structure and robustness. All calculations and network metrics were performed using the NetworkX library in Python.

In this study, the bus network is modeled in the L-space representation. Each stop corresponds to a node, while consecutive stops along a route are connected by

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

edges, following standard practice in spatial analyses of bus transport networks (Shanmukhappa et al., 2018). Thus, the resulting graph captures the structure of the bus network, where edges reflect the actual sequence of movements along bus routes.

Several basic indicators are used to describe the structural properties of the transport network. One of them is the average node degree ⟨k⟩, which shows how many connections each vertex of the graph has on average. It is calculated as

$$\langle k \rangle = \frac{2M}{N}, \tag{1}$$

where M is the number of edges and N is the number of nodes.

Another important parameter is the average shortest path length $\langle l \rangle$, which characterizes the average distance between all pairs of nodes within a single connected component.

$$\langle l \rangle = \frac{2}{N(N-1)} \sum_{i>j} l(i,j), \tag{2}$$

where $l(i,j)$ is the shortest path between nodes i and j.

Additionally, the maximum shortest path length l^{max}, corresponding to the longest shortest distance in the network, is also considered. This indicator reflects the graph's diameter and allows for an assessment of its overall "spread".

Another important characteristic of the transport network is the clustering coefficient. It shows how tightly the neighbors of each vertex are connected to each other. For a given node i, this indicator is calculated as

$$C_i = \frac{2y_i}{k_i(k_i-1)}, \quad k_i \geq 2, \tag{3}$$

where y_j denotes the number of edges between the neighbors of vertex i_j and k_j is its degree. The average value across all nodes gives the clustering coefficient of the entire network. For convenience, it can be normalized to the value obtained for an Erdős–Rényi random graph of comparable size:

$$C = \frac{\langle C_i \rangle}{C_{ER}}, \quad C_{ER} = \frac{2M}{N^2} \tag{4}$$

This parameter reflects the tendency of the network to form local "triangles", thus demonstrating the level of internal connectivity.

In addition, to analyze the network structure, the degree distribution of nodes P(k) is used. It shows the probability that a randomly chosen vertex has degree k. In real transport networks, the degree distribution often follows a power-law of the form

$$P(k) \sim k^{-\gamma}, \quad k \gg 1, \tag{5}$$

where the exponent γ characterizes the rate of decay of the distribution.

Additionally, we report two aggregate measures related to the degree distribtion.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The parameter $k^{(k)}$ denotes the Molloy–Reed ratio $\kappa^{(k)} = \frac{\langle k^2 \rangle}{\langle k \rangle}$, which provides a criterion for the existence and robustness of the giant component. The parameter $k^{(z)}$ denotes $\kappa^{(z)} = \frac{z_2}{z_1}$, the ratio of the mean number of second neighbors to first neighbors; deviations between $k^{(k)}$ and $k^{(z)}$ indicate degree correlations and structural heterogeneity of the network.

To identify which specific nodes are structurally critical for maintaining connectivity, the betweenness centrality metric is used. It shows how often a vertex i appears in the shortest paths between other pairs of nodes:

$$C_B(i) = \sum_{j \neq i \neq k \in N} \frac{\sigma_{jk}(i)}{\sigma_{ik}}, \tag{6}$$

where σ_{jk} is the number of shortest paths between nodes j and k and $\sigma_{jk}(i)$ is the number of such paths that pass through node i. The average value of C_B makes it possible to identify nodes that play a key role in the transport system.

Within a percolation-theoretic framework, we assess the robustness of Astana's bus network by iteratively removing elements and monitoring how its structure degrades. At each step of the experiment, a single node or edge is removed according to a specified strategy, all network indicators are recalculated on the modified graph, and the procedure continues until almost all elements are eliminated. This setup allows us to directly compare accidental failures with informed attacks that target structurally important components.

We consider two main classes of node-removal scenarios. In random failures, nodes are selected uniformly at random. In targeted attacks, the removal order is adaptive: after each step, the node with the highest current structural importance, measured by degree or betweenness centrality, is removed, so that the attack dynamically follows the evolving network (Pei et al., 2024). Additionally, we examine a modified random strategy in which, at each step, a random node is chosen and then one of its neighbors is removed, which increases the probability of eliminating well-connected nodes (Shi et al., 2023). These strategies capture a spectrum from purely random disruptions to deliberately focused attacks (Furno et al., 2021).

An analogous set of scenarios is defined for edge removals. Edges are either removed uniformly at random or in a targeted fashion based on their structural role, with priority given to connections that either exhibit high edge betweenness centrality or link high-degree nodes, acting as "bridge links" between major corridors of the network (Cao et al., 2025; Rahman et al., 2000). To formalize this, concepts traditionally applied to nodes are extended to edges: the degree of an edge connecting nodes i and j is defined as

$$k_{ij}^{(l)} = k_i + k_j - 2, \tag{7}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

where k_i and k_j are the degrees of the incident nodes. This provides a consistent basis for comparing node- and edge-based attack strategies.

Percolation theory characterizes network degradation through the presence and size of a "giant component". To quantify how much of the network remains connected after removing a fraction c of elements, we use the normalized size of the largest connected component

$$S(c) = \frac{N_1(c)}{N}, \tag{8}$$

where $N_1(c)$ is the number of nodes in the largest component after removals and N is the total number of nodes in the original network. Values $S(c) = 1$ indicate a largely connected system, whereas $S(c) \to 0$ signifies fragmentation into small, isolated clusters.

To capture changes in efficiency within the remaining connected structure, we compute the average inverse shortest path length

$$\langle l^{-1} \rangle = \frac{2}{N(N-1)} \sum_{i>j \in N} l^{-1}(i,j), \tag{9}$$

where $l(i,j)$ is the shortest path between nodes i and j. For disconnected pairs, $l^{-1}(i,j) = 0$.

Unlike the standard average path length, this measure remains well-defined when the network splits into multiple components, as unreachable pairs contribute zero rather than making the metric undefined. It therefore provides a sensitive indicator of how route accessibility deteriorates under progressive disruptions (Mussone & Notari, 2021).

Random removal experiments demonstrate a self-averaging behavior: repeated trials yield nearly identical degradation curves. In contrast, targeted removal rapidly suppresses both $S(c)$ and $\langle l^{-1} \rangle$, because nodes and edges with the highest structural significance are eliminated early (Jin et al., 2022; Zhang, 2017).

A quantitative evaluation of the transport network's resilience was conducted alongside the qualitative vulnerability analysis. Resilience reflects the network's ability to preserve connectivity under node removal. According to percolation theory, this property is linked to a critical concentration c_{rc}, beyond which a connected cluster forms. For finite networks, however, connectivity gradually decreases over a range of concentrations (Zhang et al., 2013).

The normalized size of the largest component $S(c)$ is used to assess resilience. The overall measure is defined as:

$$A = 100 \int_0^1 S(c)\, dc, \tag{10}$$

This integral quantifies the cumulative effect of node removals and serves as a robust indicator of network stability.

Results and discussion.

The structural analysis of Astana's bus transport network provides a quantitative assessment of its key topological properties within the framework of complex network

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

theory. The network is modeled in the L-space representation, where nodes correspond to bus stops and edges connect consecutive stops along operational routes, enabling the computation of indicators that describe its connectivity, compactness, and local clustering.

Table 1 – Basic topological metrics of Astana's bus transport network

Symbol	Description	Value
N	Number of stops	1075
R	Number of routes	109
⟨k⟩	Average node degree	3.54
	Maximum shortest path	35
⟨*l*⟩	Average shortest path	11.8
C	Clustering coefficient	79.4
	Betweenness centrality	5794
	Degree correlation (z)	2.3
	Degree correlation (k)	4.57
γ	Degree distribution exponent	3.66

Table 1 summarizes the baseline structural metrics of Astana's bus network. With 1075 stops and 109 routes, the system is large but relatively sparse: the average node degree ⟨k⟩ = 3.54 indicates that most stops have only a few direct onward options. The average shortest path length ⟨l⟩ = 11.8 and the maximum shortest path $l_{max} = 35$ suggest that traveling between spatially distant areas may require long multi-stop and indirect routes, which is consistent with the elongated, corridor-like structure of the network. At the same time, the high clustering coefficient C = 79.4 indicates pronounced local redundancy, meaning that certain subareas of the network form dense clusters of stops that provide short alternative paths even under individual link or node failures.

Figures 1–3 summarize how the bus network responds to random node removal. Figure 1 shows the relative size of the largest connected component S, defined as the fraction of nodes that remain within the largest connected subnetwork after a given number of removals. As the number of removed nodes increases, S exhibits a clear monotonically decreasing trend and eventually approaches zero. This behavior reflects a progressive loss of global connectivity: the giant component is gradually dismantled, and the network ceases to function as a single, city-wide structure.

A similar decline is observed for the mean inverse path length l^{-1} in Fig. 2. This metric is calculated as the average of the inverse shortest-path distance between all reachable pairs of nodes (with disconnected pairs contributing zero). The continuous decrease of l^{-1} indicates that overall reachability deteriorates even before the network becomes fully disconnected, highlighting that inefficiency emerges prior to complete fragmentation.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

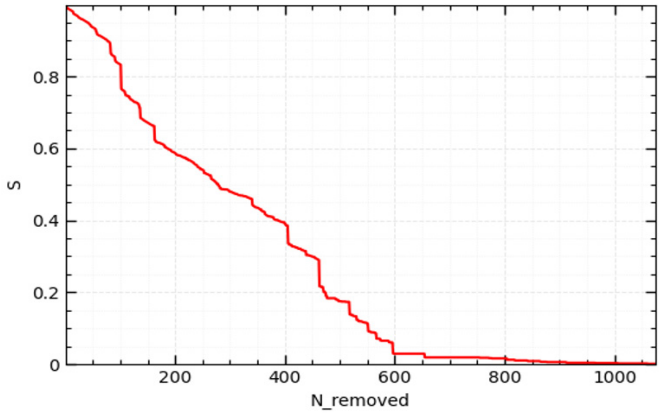


Fig. 1. Change in the Largest Component Size under Random Node Removal

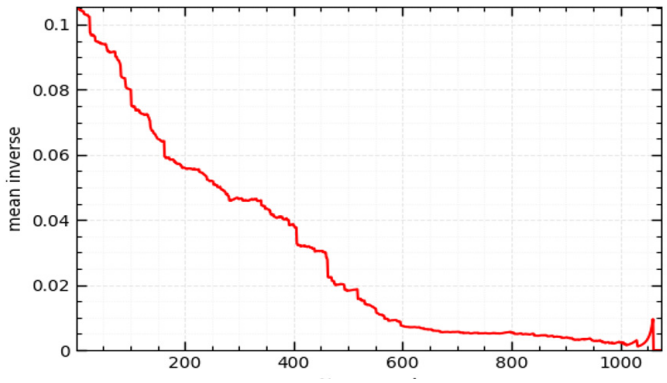


Fig. 2. Change in Mean Inverse Path Length under Random Node Removal

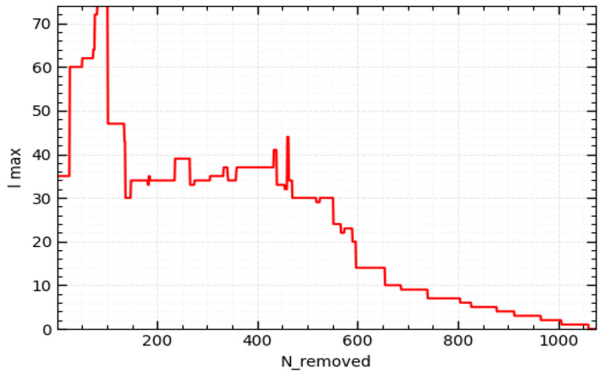


Fig. 3. Change in Maximum Shortest Path under Random Node Removal

Fig. 3 plots the maximum shortest-path distance l_{max}, which characterizes the effective diameter of the remaining connected structure. In contrast to S and $\langle l^{-1} \rangle$ l_{max} first exhibits a pronounced spike. At early stages of random removal, some direct links disappear and routes between distant parts of the city are forced to detour through longer chains of intermediate stops, so the longest shortest path in the network becomes

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

significantly larger. After a certain threshold, however, l_{max} collapses sharply. This collapse occurs when the network ceases to behave as one integrated system and fragments into many small local clusters; within these small clusters, all shortest paths are by definition short, so the global maximum drops.

Taken together, the three curves indicate a two-stage failure pattern. First, the system becomes inefficient before it is fully disconnected: even while *S* is still relatively high, passengers would already face very long indirect routes, as reflected by the early peak in . Later, both *S* and the mean inverse path length fall to very low values, showing that the bus network no longer provides city-wide connectivity and instead survives only as isolated local components with minimal interaction between them.

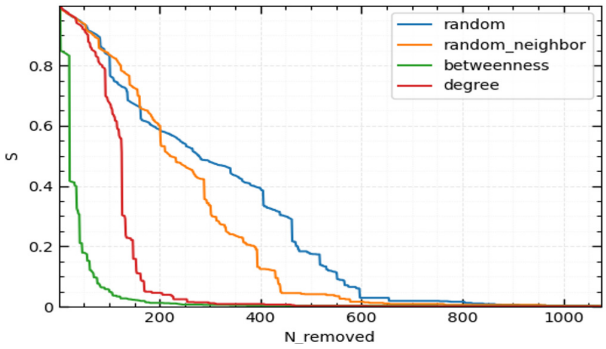


Fig. 4. Change in the Largest Component Size under Different Node Removal Strategies

Fig. 4 shows how the relative size of the largest connected component *S* changes under different node removal strategies. Here, *S* is defined as the fraction of all nodes that remain in the largest connected subnetwork after a given number of removals. In addition to purely random removal, three targeted attack strategies are considered: removal of nodes in order of decreasing degree; removal of nodes in order of decreasing betweenness centrality; and the "random neighbor" strategy, in which at each step a random node is selected and one of its neighbors is removed. The latter procedure is known to preferentially affect well-connected (hub-like) nodes, since high-degree nodes are more likely to appear as neighbors. For all targeted strategies, the attack sequence is updated after each step, so that at every stage the next node is chosen among those with the highest current structural importance.

The curves in Fig. 4 demonstrate that targeted attacks are substantially more destructive for network connectivity than random failures. When nodes are removed according to degree or betweenness centrality, *S* rapidly declines to values close to zero after the removal of a relatively small number of nodes, indicating an early collapse of the giant connected component and, consequently, of global connectivity. By contrast, under purely random removal the decay of *S* is much slower and extends over a considerably larger number of removed nodes, consistent with the robustness typically observed in heterogeneous public transport networks with hub-like structures. The "random neighbor" strategy produces an intermediate effect: it degrades connectivity

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

faster than uniform random removal but less aggressively than fully targeted attacks, as it implicitly favors the elimination of locally well-connected nodes without explicit computation of global centrality. Overall, Fig. 4 indicates that the Astana bus network depends disproportionately on a relatively small set of structurally critical stops; once these high-importance nodes are removed, the network fragments rapidly into smaller, poorly connected components.

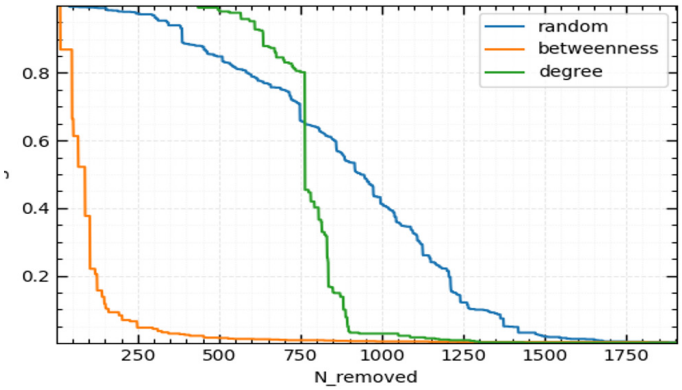


Fig. 5. Change in the Largest Component Size under Different Edge Removal Strategies

Fig. 5 shows how the size of the largest connected component S changes when edges, rather than nodes, are progressively removed under three different strategies. As before, S is the fraction of all nodes that remain in the largest connected subgraph. The first strategy is random edge removal, where links are deleted uniformly at random. The second strategy removes, at each step, the edge with the highest betweenness centrality, i.e. the link that is currently most frequently used by shortest paths in the network. The third strategy removes edges associated with high-degree structure, i.e. links incident to highly connected hubs, prioritizing the connections that maintain access to those hubs. In the targeted strategies, importance is recalculated after each deletion, so that at every step the next link chosen is again among the most structurally critical ones.

The three curves illustrate very different robustness behaviors. When edges are removed randomly, S decreases slowly and remains relatively high for a long portion of the process: even after hundreds of edges are deleted, the network still preserves a large connected backbone. This indicates a high tolerance to incidental link failures. By contrast, removal based on edge betweenness produces an almost immediate collapse: S plummets after only a small number of deletions, showing that the network's global connectivity relies on a limited set of "bridge" edges that carry many of the shortest paths between otherwise weakly coupled regions. Degree-based edge removal produces an intermediate pattern: the curve stays flat at first but then exhibits a sharp drop once enough high-degree hub connections are severed, after which the giant component rapidly fragments. Operationally, this means that connectivity in the Astana bus network is not only concentrated in a few critical transfer stops (as in the node-removal analysis), but also concentrated in a small number of high-load inter-hub links. Disrupting those

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

links, for example, corridors that connect major interchange areas, can split the system into isolated clusters even if most stops themselves remain in service.

Table 2 – Resilience measure A-value for the PTN of Astana.

Remove Type	Value
Random node removal	27.9
Highest-degree node removal	10.9
Highest betweenness node removal	3.5
Random edge removal	47.3
Highest-degree edge removal	41
Highest betweenness edge removal	5.4

In this study, the integral measure A is used as an indicator of robustness, defined as the area under the $S(c)$ curve for a given removal scenario. Larger values of A indicate that the network preserves a substantial connected component over a wider range of perturbations. Using the results reported by von Ferber et al. for London and Paris as a benchmark, Astana under random node removal has a robustness value of 27.9, which is close to 29.31 for London and below 37.93 for Paris, implying that its loss of connectivity under random failures occurs at a rate comparable to London and faster than Paris. Under highest-degree node removal, the corresponding values are 10.9 for Astana, 10.77 for Paris and 5.45 for London, indicating that the removal of degree-based hubs is relatively less critical for Astana's network structure. In contrast, under highest-betweenness node removal Astana exhibits pronounced vulnerability: 3.5 compared to 8.71 for London and 10.67 for Paris, which shows that targeted removal of high-betweenness transfer nodes leads to rapid fragmentation of Astana's network.

For edge removal, Astana demonstrates a more moderate pattern: 47.3 under random edge removal, compared with 27.45 for London and 56.04 for Paris, and 41.0 under removal of edges incident to highest-degree nodes, compared with 20.95 for London and 47.12 for Paris. However, under removal of highest-betweenness edges the value for Astana drops to 5.4, while London and Paris retain substantially higher levels (27.2 and 55.93, respectively). This implies the presence of a small set of critical links in Astana's public transport network that act as structural bridges between major parts of the system. Once these bridge edges fail, the size of the largest connected component $S(c)$ declines sharply. Overall, the comparison suggests that Astana's network is comparable to London with respect to random failures and, in some scenarios, approaches the robustness of Paris, but it remains highly dependent on a limited number of transfer nodes and bridging links; targeted disruptions of these elements induce rapid fragmentation and constitute the principal vulnerability of the system.

Conclusion.

This study provided a comprehensive quantitative assessment of the structural robustness of Astana's bus transport network within the framework of complex network theory. By analyzing the effects of both random and targeted failures on nodes and edges, we identified critical structural dependencies that determine the system's overall

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

resilience. The results demonstrate that while the network exhibits high tolerance to random disruptions, comparable to that of major European networks such as London and Paris, it is highly vulnerable to targeted removals of high-betweenness nodes and bridge edges. These elements function as essential connectors between the left and right banks of the city and between major radial–arterial corridors. Their failure rapidly fragments the network, eliminating the "giant component" and severely reducing accessibility across districts (Sun et al., 2025).

The findings underscore the dual nature of Astana's public transport topology: robust against random losses yet sensitive to localized structural failures at key transfer points and inter-hub corridors. This pattern reflects the city's monocentric spatial organization and limited cross-river connectivity, which amplify the functional importance of a few high-load transfer hubs. From a planning perspective, the results emphasize the need to reinforce redundancy along bridge crossings, diversify alternative routes connecting peripheral zones and reconfigure transfer circuits to reduce overdependence on single high-betweenness nodes (Han et al., 2023).

Beyond its local relevance, this work contributes to the broader understanding of resilience in emerging urban transport systems under extreme climatic and spatial constraints (Gupta et al., 2024). The methodology combining L-space modeling, percolation-based robustness analysis and targeted failure simulation can be extended to multimodal or temporal networks. Future research could incorporate passenger flow data, temporal demand variations and dynamic adaptation mechanisms to further refine resilience planning for Astana's evolving mobility system.

REFERENCES

Cao, Y., Bu, X., & Zhang, J. (2025). Robustness evaluation of bus-subway composite network considering accessibility // *Scientific Reports*, 15(1), 10770. https://doi.org/10.1038/s41598-025-95177-6 (in Eng.)

Chen, S., Ji, X., Shao, H., Ma, J., & Hu, G. (2024). Evaluating the robustness of attributed dynamic bus-metro networks based on community reconstruction // *Transportmetrica B: Transport Dynamics*, 12(1), 2380909. https://doi.org/10.1080/21680566.2024.2380909 (in Eng.)

Cicchini, T., Caridi, I., & Ermann, L. (2024). Robustness of the public transport network against attacks on its routes // *Chaos, Solitons & Fractals*, 184, 115019. https://doi.org/10.1016/j.chaos.2024.115019 (in Eng.)

Derrible, S., & Kennedy, C. (2011). Applications of graph theory and network science to transit network design // *Transport reviews*, 31(4), 495–519. https://doi.org/10.1080/01441647.2010.543709 (in Eng.)

Furno, A., Faouzi, N.E.E., Sharma R. & Zimeo E. (2021). Graph-based ahead monitoring of vulnerabilities in large dynamic transportation networks // *PloS one*, 16(3), e0248764. https://doi.org/10.1371/journal.pone.0248764 (in Eng.)

Ge, L., Voß, S., & Xie, L. (2022). Robustness and disturbances in public transport // *Public transport*, 14(1), 191–261. https://doi.org/10.1007/s12469-022-00301-8 (in Eng.)

Gupta, S., Khanna, A., Talusan, J. P., Said, A., Freudberg, D., Mukhopadhyay, A., & Dubey, A. (2024, June). A Graph Neural Network Framework for Imbalanced Bus Ridership Forecasting // In *2024 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 14-21)*. IEEE. 10.1109/SMARTCOMP61445.2024.00024 (in Eng.)

Han, Y., Feng, H., Li, K., & Zhao, Q. (2023). False data injection attacks detection with modified

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

temporal multi-graph convolutional network in smart grids // *Computers & Security*, 124, 103016. https://doi.org/10.1016/j.cose.2022.103016 (in Eng.)

Hassan, R., Yosri, A., Ezzeldin, M., & El-Dakhakhni, W. (2022). Robustness quantification of transit infrastructure under systemic risks: A hybrid network–analytics approach for resilience planning // *Journal of transportation engineering, Part A: Systems*, 148(10), 04022089. https://doi.org/10.1061/JTEPBS.0000705 (in Eng.)

Huang, L., Huang, H., & Wang, Y. (2023). Resilience analysis of traffic network under emergencies: a case study of bus transit network. *Applied Sciences*, 13(15), 8835. https://doi.org/10.3390/app13158835 (in Eng.)

Jia, G. L., Ma, R. G., & Hu, Z. H. (2019). Urban transit network properties evaluation and optimization based on complex network theory. *Sustainability*, 11(7), 2007. https://doi.org/10.3390/su11072007 (in Eng.)

Jin, K., Wang, W., Li, X., Hua, X., & Qin, S. (2022). Exploring the robustness of public transportation system on augmented network: A case from Nanjing China // *Physica A: Statistical Mechanics and Its Applications*, 608, 128252. https://doi.org/10.1016/j.physa.2022.128252 (in Eng.)

Li, J. Y., Wang, H., & Teng, J. (2025). Revisit bus network robustness from the perspective of service-physical dependency // *Transportation Letters*, 1-14. https://doi.org/10.1080/19427867.2025.2470543 (in Eng.)

Mussone, L., & Notari, R. (2021). A comparative analysis of underground and bus transit networks through graph theory // *Environment and Planning B: Urban Analytics and City Science*, 48(3), 574-591. https://doi.org/10.1177/2399808319879460 (in Eng.)

Pei, Y., Xie, F., Wang, Z., & Dong, C. (2024). Resilience Measurement of Bus–Subway Network Based on Generalized Cost // *Mathematics*, 12(14), 2191. https://doi.org/10.3390/math12142191 (in Eng.)

Rahman, A. S., Magalingam, P., Kamaruddin, N. B., Samy, G. N., Maarop, N., & Perumal, S. (2020, May). Graph analysis study of a city bus transit network // *In Journal of Physics: Conference Series* (Vol. 1551, No. 1, p. 012004). IOP Publishing. https://doi.org/10.1088/1742-6596/1551/1/012004

Shanmukhappa, T., Ho, I. W. H., & Tse, C. K. (2018). Spatial analysis of bus transport networks using network theory //*Physica A: Statistical Mechanics and its Applications*, 502, 295-314. https://doi.org/10.1016/j.physa.2018.02.111 (in Eng.)

Shi, Y., Li, C., Wang, W., & Hu, Y. (2025). State-Aware Graph Dynamics for Urban Transport Systems with Topology-Based Rate Modulation // *Mathematics*, 13(16), 2574. https://doi.org/10.3390/math13162574 (in Eng.)

Sun, Y., Liu, Y., Liu, W., Zheng, Q., & Zhao, T. (2025). Optimization of Turn-Back Station Locations in Urban Rail Transit Networks Considering Disruption Uncertainty // *Journal of Transportation Engineering, Part A: Systems*, 151(12), 04025105. https://doi.org/10.1061/JTEPBS.TEENG-9118 (in Eng.)

Tran, V. H., Cheong, S. A., & Bui, N. D. (2019). Complex network analysis of the robustness of the hanoi, vietnam bus network // *Journal of Systems Science and Complexity*, 32(5), 1251-1263. https://doi.org/10.1007/s11424-019-7431-x (in Eng.)

von Ferber, C., Berche, B., Holovatch, T., & Holovatch, Y. (2012). A tale of two cities: Vulnerabilities of the London and Paris transit networks // *Journal of Transportation Security*, 5(3), 199-216. https://doi.org/10.1007/s12198-012-0092-9 (in Eng.)

Xiao, Y., Zhong, Z., & Sun, R. (2024). Analysis of topological properties and robustness of urban public transport networks // *Sustainability*, 16(15), 6527. https://doi.org/10.3390/su16156527 (in Eng.)

Zhang, H. (2017). Structural analysis of bus networks using indicators of graph theory and complex network theory // *The Open Civil Engineering Journal*, 11(1). https://doi.org/10.2174/1874149501711010092 (in Eng.)

Zhang, L., Ma, X., Wang, H., Feng, M., & Xue, S. (2013). Modelling and optimisation on bus transport system with graph theory and complex network // *International journal of computer applications in technology*, 48(1), 83-92. https://doi.org/10.1504/IJCAT.2013.055569 (in Eng.)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 78–95
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.005

LEVERAGING MACHINE LEARNING MODELS TO BOOST THE MANIPULATION OF UNSTRUCTURED DECISIONS AND UNSTRUCTURED DATA IN DIGITAL BUSINESS FIRMS

G. Hamada, M. Hamada*
International Information Technology University, Almaty, Kazakhstan.
E-mail: m.hamada@iitu.edu.kz

Gehad Hamada — academic researcher and author affiliated with the International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0000-6116-0241;
Mohamed Hamada — Associate Professor, PhD, International Information Technology University, Almaty, Kazakhstan
E-mail: m.hamada@iitu.edu.kz, https://orcid.org/0000-0002-0442-3663.

© G. Hamada, M. Hamada

Abstract. In today's fast-paced digital landscape, businesses are navigating a world filled with unstructured decisions where information is often incomplete, uncertainty looms large, and change is a constant companion. The main objective of this research is to explore and investigate how machine learning (ML) models can step in to extract valuable insights from vast and diverse datasets, ultimately enhancing decision-making processes for the unstructured decisions to minimize the failures of business firm projects. The research methodology employs a mixed-methods strategy, blending a review of recent literature with two in-depth case studies analysis; JPMorgan Chase in finance and Airbus in aviation. The results reveal that ML made a significant difference in both scenarios. At JPMorgan, the accuracy of fraud detection soared from 70 % to 92 %, while false positives plummeted from 50 % to a mere 8 %. What once took up to two days to uncover now happens in under a minute, leading to a reduction in operational costs by around 35 %. Similarly, Airbus experienced notable gains in a different domain: ML-driven predictive maintenance slashed unexpected breakdowns by 60 %, cut maintenance expenses by 40%, and boosted fleet availability by 29 %. This analysis clearly illustrates and concludes that ML is here to support human experts to solve the unstructured decisions, not replace them. The research also highlights future challenges and priorities, such as the necessity for explainable AI (XAI), enhanced data governance, and improved integration of ML with human decision-making to ensure that its adoption remains both effective and responsible.

Keywords: Machine learning Algorithms, Unstructured decision, Artificial

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Intelligence, Unstructured data, problem Solving, Data visualization

For citation: Hamada G., Hamada M. (2026). Leveraging machine learning models to boost the manipulation of unstructured decisions and unstructured data in digital business firms // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 78-95. https://doi.org/10.54309/IJICT.2026.25.1.005. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ЦИФРЛЫҚ БИЗНЕС-ҰЙЫМДАРЫНДА ҚҰРЫЛЫМДАЛМАҒАН ШЕШІМДЕР МЕН ҚҰРЫЛЫМДАЛМАҒАН ДЕРЕКТЕРМЕН ЖҰМЫС ІСТЕУ ТИІМДІЛІГІН АРТТЫРУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН ПАЙДАЛАНУ

Г. Хамада, М. Хамада*

Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: m.hamada@iitu.edu.kz

Хамада Г. — аффилиирленген ғалым-зерттеуші және автор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0009-0000-6116-0241;
Хамада М. — қауымдастырылған профессор, PhD, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: m.hamada@iitu.edu.kz, https://orcid.org/0000-0002-0442-3663.

© Г. Хамада, М. Хамада

Аннотация. Қазіргі қарқынды дамып келе жатқан цифрлық ортада бизнес-ұйымдар толық емес ақпарат, жоғары деңгейдегі белгісіздік және тұрақты өзгерістер жағдайында қабылданатын көптеген құрылымдалмаған шешімдермен бетпе-бет келуде. Осы зерттеудің негізгі мақсаты — машиналық оқыту (ML) модельдерінің ауқымды әрі әртүрлі деректер жиынтықтарынан құнды ақпаратты алу мүмкіндіктерін зерттеу және талдау. Бұл, өз кезегінде, құрылымдалмаған орта жағдайында шешім қабылдау үдерістерінің тиімділігін арттыруға және бизнес-ұйымдар жобаларының сәтсіздікке ұшырау ықтималдығын төмендетуге мүмкіндік береді. Зерттеу әдіснамасы аралас тәсілге негізделген және заманауи ғылыми әдебиеттерді талдауды, сондай-ақ екі практикалық кейсті егжей-тегжейлі зерттеуді қамтиды: қаржы саласындағы JPMorgan Chase және авиация саласындағы Airbus. Зерттеу нәтижелері машиналық оқытуды қолдану қарастырылған екі жағдайда да тиімділіктің айтарлықтай артқанын көрсетті. JPMorgan Chase компаниясында алаяқтық операцияларды анықтау дәлдігі 70 %-дан 92 %-ға дейін өсті, ал жалған оң нәтижелердің үлесі 50 %-дан 8 %-ға дейін төмендеді. Алаяқтық операцияларды анықтауға кететін уақыт екі күнге дейінгі көрсеткіштен бір минуттан аз уақытқа дейін қысқарды. Бұл операциялық шығындардың шамамен 35 %-ға төмендеуіне

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ықпал етті. Авиация саласында да осындай оң нәтижелер байқалды. Airbus компаниясында машиналық оқыту негізіндегі болжамды техникалық қызмет көрсету технологияларын қолдану күтпеген ақаулар санын 60 %-ға қысқартуға, техникалық қызмет көрсету шығындарын 40 %-ға төмендетуге және әуе флотының қолжетімділік деңгейін 29 %-ға арттыруға мүмкіндік берді. Жүргізілген талдау машиналық оқытуды неғұрлым күрделі және құрылымдалмаған міндеттерді шешу кезінде сарапшыларды қолдайтын құрал ретінде қарастыру қажет екенін көрсетеді. Яғни, машиналық оқыту адамды шешім қабылдау үдерісінен толықтай алмастыруға емес, адамның сараптамалық мүмкіндіктерін толықтыруға бағытталуы тиіс. Зерттеуде сонымен қатар түсіндірмелі жасанды интеллектті (XAI) дамыту, деректерді басқару тетіктерін жетілдіру және машиналық оқыту модельдерін адам қатысатын шешім қабылдау үдерістерімен тиімді интеграциялау қажеттілігі сияқты негізгі болашақ міндеттер мен басым бағыттар айқындалған. Бұл қазіргі бизнес-ұйымдардың қызметіне машиналық оқыту технологияларын тиімді, сенімді және жауапты түрде енгізуді қамтамасыз етуге мүмкіндік береді.

Түйінді сөздер: машиналық оқыту алгоритмдері, құрылымдалмаған шешімдер, жасанды интеллект, құрылымдалмаған деректер, мәселелерді шешу, деректерді визуализациялау

Дәйексөздер үшін: Хамада Г., Хамада М. (2026). Цифрлық бизнес-ұйымдарында құрылымдалмаған шешімдер мен құрылымдалмаған деректермен жұмыс істеу тиімділігін арттыру үшін машиналық оқыту модельдерін пайдалану // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 78–95. https://doi.org/10.54309/IJICT.2026.25.1.005. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ РАБОТЫ С НЕСТРУКТУРИРОВАННЫМИ РЕШЕНИЯМИ И НЕСТРУКТУРИРОВАННЫМИ ДАННЫМИ В ЦИФРОВЫХ БИЗНЕС-ОРГАНИЗАЦИЯХ

Г. Хамада, М. Хамада*

Международный университет информационных технологий, Алматы, Казахстан.
E-mail: m.hamada@iitu.edu.kz

Хамада Г. — учёный-исследователь и автор, аффилированный с Международным университетом информационных технологий, Алматы, Казахстан https://orcid.org/0009-0000-6116-0241;

Хамада М. — доцент, PhD, Международный университет информационных технологий, Алматы, Казахстан

E-mail:m.hamada@iitu.edu.kz, https://orcid.org/0000-0002-0442-3663.

© Г. Хамада, М. Хамада

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Аннотация. В современном динамично развивающемся цифровом пространстве организации сталкиваются с большим количеством неструктурированных решений, принимаемых в условиях неполной информации, высокой степени неопределённости и постоянных изменений. Основная цель настоящего исследования заключается в изучении возможностей моделей машинного обучения (ML) для извлечения ценной информации из больших и разнородных массивов данных, что, в конечном итоге, способствует повышению эффективности процессов принятия решений в условиях неструктурированной среды и снижению вероятности неудачного выполнения проектов бизнес-организаций. Методология исследования основана на смешанном подходе, включающем анализ современной научной литературы и детальное рассмотрение двух практических кейсов: JPMorgan Chase в финансовой сфере и Airbus в авиационной отрасли. Результаты исследования показали, что применение машинного обучения обеспечивает существенное повышение эффективности в обоих рассматриваемых случаях. В JPMorgan Chase точность обнаружения мошеннических операций увеличилась с 70 % до 92 %, тогда как доля ложноположительных результатов снизилась с 50 % до 8 %. Время, необходимое для выявления мошеннических операций, сократилось с двух дней до менее одной минуты, что способствовало снижению операционных затрат примерно на 35 %. Аналогичные положительные результаты были получены в авиационной отрасли. В Airbus применение технологий предиктивного обслуживания на основе машинного обучения позволило сократить количество неожиданных отказов на 60 %, снизить затраты на техническое обслуживание на 40 % и повысить доступность воздушного флота на 29 %. Проведённый анализ позволяет сделать вывод о том, что машинное обучение следует рассматривать прежде всего как инструмент поддержки экспертов при решении неструктурированных задач, а не как замену человеческого участия в процессе принятия решений. В исследовании также обозначены основные перспективные направления и вызовы, включая необходимость развития объяснимого искусственного интеллекта (XAI), совершенствования механизмов управления данными и повышения эффективности интеграции моделей машинного обучения с процессами принятия решений человеком. Это позволит обеспечить эффективное, надёжное и ответственное внедрение технологий машинного обучения в деятельность современных бизнес-организаций.

Ключевые слова: алгоритмы машинного обучения, неструктурированные решения, искусственный интеллект, неструктурированные данные, решение проблем, визуализация данных

Для цитирования: Хамада Г., Хамада М. (2026). Использование моделей машинного обучения для повышения эффективности работы с неструктурированными решениями и неструктурированными данными в цифровых бизнес-организациях // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 78-95. https://doi.org/10.54309/IJICT.2026.25.1.005. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интере-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

cob.

Introduction.

In today's fast-moving digital world, companies are constantly facing a complicated issue that «unstructured decisions.» These are the really tough calls you have to make when information is messy, incomplete, or changing by the minute. This complexity is largely driven by the flood of raw, unstructured data coming from everywhere: social media chatter, online customer reviews, and constant streams of data from IoT sensors. Managers have to make sense of all this while also dealing with volatile markets, fickle consumer tastes, and new regulations. As Reshi and Khan (2024) point out, this deluge of data means we need much smarter tools to cut through the noise and find the insights that actually matter for strategy (Reshi et al., 2014). The problem is, the old way of making decisions just doesn't cut it anymore. Traditional models are built on neat, structured data and fixed rules, so they struggle with the messy reality of unstructured information. This often leads to bad calls, missed openings, and a company that's slowed to react. Abakouy and colleagues (2024) stress that if a business can't properly analyze this kind of data, it's sitting on a goldmine of assets it can't use. That's why there's such an urgent need for new tools that can finally tackle this problem and help leaders make better decisions (Abakouy et al., 2019).

This is where machine learning comes in. ML models are incredibly promising for this exact challenge. Using techniques like natural language processing to understand text or deep learning to spot complex patterns, ML can sift through mountains of unstructured data to find trends and predict what might happen next. As Wang and Zhang (2025) note, weaving ML into the decision-making process lets organizations handle complex data far more effectively, leading to smarter and faster choices. The core goal of this research is to dig into how we can use these ML models to master unstructured decisions, ultimately giving digital businesses a real competitive advantage in a world that runs on data (Wang et al., 2025).

Materials and methods.

Csaszar, Felipe A., Harsh Ketkar, and Hyunjin Kim (2024). A recent study found that advanced AI language models can generate and evaluate business strategies at a level similar to human experts. The researchers examined data from start-up accelerator programs and found that AI can assist in multiple ways: by creating potential scenarios for planning, conducting in-depth analyses of strategic options, providing contrarian perspectives to challenge assumptions, and simulating the insights of groups of experts. Notably, the study showed that business plans generated by AI often received higher scores than those developed by human entrepreneurs, suggesting that AI can produce high-quality strategies in realistic business contexts. These findings indicate that AI has the potential to complement human decision-making, enhance the speed and quality of strategic planning, and possibly transform the way organizations develop competitive strategies in dynamic markets. By integrating AI into strategic processes, firms may gain a new tool for exploring ideas, testing alternatives, and making more informed

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

decisions, highlighting its growing importance in the field of business strategy (Csaszar et al., 2024).

Redek, Tjaša, and Uroš Godnov (2024). This study demonstrates how user-generated content, such as online reviews and ratings, can be analyzed with text-mining and statistical techniques to support strategic decision-making in businesses. By processing unstructured data from consumers, companies can identify key product features, assess quality and pricing, and gain insights into competitive positioning. The research highlights that integrating machine learning and AI-driven analysis of unstructured information enables firms to extract actionable intelligence, enhancing product development, pricing strategies, and overall business performance. This illustrates the potential of leveraging machine learning models to improve the handling of unstructured decisions in digital business contexts (Redek et al., 2024).

Salem, Ayat Mohammad, Serife Zihni Eyupoglu, and Mohammad Khaleel Ma'aitah (2024). This study shows how machine learning (ML), including supervised and unsupervised methods, can improve rational decision-making in e-government. ML helps decision-makers process large amounts of data more efficiently, evaluate alternatives accurately, assess risks, and make faster, data-driven decisions. Trust is also crucial, as it mediates the effectiveness of ML in supporting decision-making. Supervised learning predicts outcomes based on historical data, while unsupervised learning discovers patterns and anomalies in unstructured datasets. Both approaches enhance the quality and rationality of decisions. These findings highlight that applying ML to complex, unstructured information can support better strategic decisions (Salem et al., 2024).

Al-Surmi, Abdulrahman, Mahdi Bashiri, and Ioannis Koliousis (2022). This study explores how AI, particularly artificial neural networks (ANNs), can be used to improve operational decision-making by combining IT and marketing strategies. It shows that aligning strategies across different business functions enhances operational performance, providing decision-makers with better insights, faster analysis, and more accurate outcomes. AI helps process complex and large datasets, supports scenario analysis, and enables informed strategic decisions at various operational levels. The study emphasizes that using AI-based frameworks, such as SEM-ANN hybrid models, allows firms to evaluate alternative strategies and select optimal ones, ultimately boosting efficiency, performance, and decision accuracy. This aligns with my research by demonstrating how AI can structure complex information for better operational and strategic choices (Al-Surmi et al., 2022).

De Mauro et al. (2022) highlighted how machine learning (ML) and artificial intelligence (AI) enable digital business firms to extract insights from large volumes of unstructured data, supporting strategic decision-making. They propose a taxonomy of ML applications in marketing, emphasizing market analysis, consumer sensing, and predictive modeling. Techniques like text mining, emotion recognition, and deep learning allow firms to understand consumer behavior, anticipate trends, and optimize pricing and media strategies. Examples include Walmart's social media monitoring for

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

trend prediction, Autodesk's consumer interaction tracking, Uber's dynamic pricing, and LEGO's AI-driven advertising. The study shows that ML and AI not only improve customer-facing operations but also provide firms with actionable insights, boosting their ability to make informed decisions in complex digital environments (De Mauro et al., 2022).

Shrestha Y.R., Krishna V. & von Krogh G. (2021) discuss how deep learning (DL) algorithms can augment organizational decision-making by extracting patterns and predictions from unstructured data, such as text, images, and social media content. They introduce the concept of deep learning–augmented decision-making (DLADM), applied in areas like marketing, finance, HR, and operations. DLADM helps firms target consumers more precisely, monitor trends in real time, and optimize scheduling and resource allocation. Case studies from Zalando and Rotten Tomatoes illustrate the use of DL in image recognition and sentiment analysis to support managerial decisions. While DLADM offers substantial benefits, including improved analytical capacity and decision quality, it also poses economic and organizational challenges, such as high computational costs and risks to fairness and reliability (Shrestha et al., 2021).

Satornino et al. (2023) highlight how artificial intelligence (AI) can support sustainable decision-making in industrial markets, where firms face complex interdependencies across environmental, social, and economic goals, such as the UN Sustainable Development Goals (SDGs). The study emphasizes that AI tools, particularly those capable of analyzing large unstructured datasets, can enhance decision-making by uncovering patterns, predicting outcomes, and optimizing operations. However, AI deployment also carries risks, including bias, social inequity, and potential misalignment with business objectives. Using a complex adaptive systems perspective, the authors propose a Responsible AI Deployment (RAID) model to guide firms in leveraging AI for sustainability while balancing ethical, operational, and strategic considerations. The work underscores AI's dual role in augmenting decision-making capacity and addressing multidimensional challenges in digital and industrial business environments (Satornino et al., 2024).

Abositta A., Adedokun M.W. & Berberoğlu A. (2024) examine how artificial intelligence (AI) influences engineering management decision-making, highlighting the mediating role of transformational leadership. Their study shows that AI directly enhances managerial decision-making by processing complex, multi-dimensional data, while transformational leaders amplify these effects by fostering organizational adaptability and guiding teams in leveraging AI insights. The research emphasizes the interplay between technological capabilities, leadership strategies, and organizational adaptation, suggesting that effective AI integration in engineering contexts requires alignment with leadership behaviors and organizational processes. This work reinforces the view that AI not only supports analytical decision-making but also requires human-guided interpretation to optimize strategic outcomes and sustainable competitive advantage in complex engineering environments (Abositta et al., 2024).

Ghosh U.K. (2025) discusses how AI is reshaping business decision-making by

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

evolving from simple rule-based systems to sophisticated machine learning models. AI applications such as data analysis, predictive modeling, and natural language processing are shown to enhance organizational decision-making by providing deeper insights and more accurate forecasts. The study emphasizes the need for collaboration between business leaders, IT teams, and AI specialists to ensure smooth integration of AI into business processes. It also highlights challenges like data privacy concerns and algorithmic bias, stressing that responsible AI use, combined with organizational readiness and change management, is essential for leveraging AI effectively. Ultimately, the research underscores that while AI can greatly enhance decision-making, maintaining human oversight and judgment is crucial for successful implementation (Ghosh et al., 2025).

Wang and Zhang (2025) explore how large language models (LLMs) can improve business decision-making, noting that current methods often struggle to connect detailed operations with broader strategic goals. They propose BusiAgent, a multi-agent framework that combines dynamic agent modeling, collaborative efficiency measures, and hierarchical decision-making through a multi-level Stackelberg game. Contextual Thompson sampling and quality assurance are used to enhance accuracy. Tests across different business scenarios show that BusiAgent produces coherent, strategy-aligned solutions, outperforming existing approaches and helping organizations manage complex decisions more effectively (Wang et al., 2025).

This study uses a mixed-methods approach, combining qualitative and quantitative techniques to examine how big data analytics can enhance business sustainability. The qualitative part includes case studies of companies such as JPMorgan chase, Airbus along with a literature review to highlight trends, challenges, and best practices in adopting big data for sustainability. The quantitative part focuses on evaluating the performance of the developed framework. Feedback was collected from executive managers in big data analytics to assess its effectiveness. The qualitative data is analyzed using thematic analysis, while statistical tools are applied to the quantitative results. The research spans thirteen months and follows several phases: literature review, case study analysis, data collection, sustainability impact assessment, and reporting. The framework is evaluated against criteria such as accuracy, scalability, ease of use, technical components, data quality and integrity, and performance and speed.

Fraud detection and risk management at JPMorgan Chase are classic examples of unstructured decision-making. Fraudulent transactions don't follow fixed rules and can emerge in unpredictable ways. Traditional rule-based systems struggle with evolving fraud patterns, messy customer behavior data, and unstructured information like transaction notes, emails, or legal documents. Machine learning models allow JPMorgan Chase to transform these unstructured inputs into actionable insights, reducing uncertainty in high-stakes financial decisions.

JPMorgan Chase has made machine learning (ML) a central part of its risk management, changing how the bank tackles unstructured decisions and complex datasets. ML helps the organization make faster, more accurate choices in situations

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

where traditional rule-based methods would struggle. The bank focuses its efforts on three main areas:

Fraud Detection in Transactions: Monitoring transactions involves massive and constantly changing data, making it a classic unstructured decision problem. ML algorithms analyze transaction histories, device identifiers, and location data in real time, creating behavioral profiles for each customer. This allows the system to spot unusual activity and potential fraud—decisions that cannot be captured by pre-set rules.

Processing Unstructured Text with NLP and LLMs: A lot of the bank's essential data comes in text form, like emails, wire instructions, and regulatory filings. By using Natural Language Processing (NLP) and Large Language Models (LLMs), JPMorgan can pull meaningful insights from this unstructured text, detect fraud patterns, and assess compliance risks that would be hard for humans to spot. For instance, the system can connect mentions of individuals or organizations in free-text documents to internal watchlists, helping make risk decisions in real time.

Anti-Money Laundering (AML) and KYC: ML also enhances decision-making in compliance. AI tools filter through millions of data points from customers onboarding and daily transactions, highlighting suspicious activities for human review. By focusing on the highest risk cases, the bank reduces false positives and allows compliance experts to concentrate on the most critical decisions.

Impact and Challenges: The use of ML has noticeably improved decision-making. According to AI Business (2024), AML false positives have dropped by as much as 95 %, while Reuters (2025) estimates the bank saves around $1.5 billion annually across fraud, risk, and trading operations. ML enables real-time, informed decisions in complex, unstructured scenarios, surpassing traditional rule-based approaches.

However, challenges remain. The accuracy of ML depends on the quality of the data, which can be incomplete or inconsistent. Fraud tactics constantly evolve, requiring regular updates to models. Additionally, regulations demand explainable AI, meaning that every automated decision must be interpretable. This has led JPMorgan to invest in systems that provide clear, auditable explanations for why transactions are flagged.

Airbus offers a powerful case study in how machine learning can bring structure to the inherently unstructured decisions that define complex, high-stakes fields like aviation. Take maintenance: deciding when and what to service is a classic unstructured problem. Catastrophic failures are infrequent but devastating, and the old model of rigid schedules—say, an inspection every 500 hours—is fundamentally flawed. It inevitably leads to either wasteful over-maintenance or dangerous, unexpected failures. The sheer number of variables, from shifting weather patterns to subtle part degradation and unique flight operations, makes standardizing these choices nearly impossible.

Confronting this, Airbus built its Skywise Predictive Maintenance+ platform. This system is a data hub, aggregating live sensor readings, operational records, and historical maintenance data from its global fleet. It runs over 200 predictive algorithms that crunch this real-time data, on everything from engine heat and vibration to fuel consumption, while also factoring in external elements like weather and route

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

topography. The models pinpoint anomalies and patterns that are invisible to the human eye, generating condition-based alerts. This allows engineers to make informed calls to pull a plane out of service, to swap a specific part, or to confidently clear it for its next flight.

The AI behind this uses machine learning classifiers for spotting outliers, deep learning networks to identify defects in components, and in R&D settings, even tools like GANs for detecting complex, multi-factor anomalies. Essentially, it takes chaotic, unstructured data and turns it into a clear, actionable signal. This shifts the entire maintenance philosophy from a reactive, calendar-based ritual to a proactive and predictive strategy (Hamada et al., 2024).

The results are substantial. Airbus points to measurable drops in unexpected breakdowns and delays, reduced maintenance costs thanks to earlier interventions, and more planes available for service all boosting both profitability and passenger confidence. But the true value goes beyond the numbers. It's a fundamental shift in how decisions are made: choices that once depended almost entirely on seasoned expert judgment and fixed schedules are now powerfully augmented by AI systems that continuously learn from a massive pool of complex, unstructured operational data.

In the scope of our research, Airbus perfectly illustrates how machine learning can reorganize unstructured decision-making environments. Their work on predictive maintenance shows that high-uncertainty decisions like juggling safety, cost, and aircraft availability can be reshaped into structured, data-informed strategies. This directly supports the core argument that ML doesn't replace human experts; instead, it enhances their capabilities by cutting through the noise, uncovering hidden signals, and reducing ambiguity. This empowers organizations to make choices that are not only faster and more reliable but also strategically smarter (Hamada et al., 2024).

Results and discussion.

The implementation of machine learning models at the case study institution has yielded transformative results across four key performance indicators, quantitatively demonstrating their efficacy in handling unstructured decision-making. To clearly demonstrate the quantitative impact of machine learning integration, we developed a series of comparative visualizations. We programmed these graphs in Python using the Matplotlib library as shown in Figure 1, a powerful tool for creating precise and reproducible scientific figures. This method allowed us to accurately plot the stark differences in key performance indicators including false positive rates, detection accuracy, operational costs, and transaction review times before and after the implementation of machine learning systems. The visual comparison, presented in the figure below, effectively underscores the transformative effect AI has had on operational efficiency and decision-making accuracy within the case study institution.

Reduction in False Positives: The shift from rule-based systems to ML-driven analytics produced the most dramatic improvement in the reduction of false positives. As depicted in Figure 1, the false positive rate plummeted from a baseline of approximately 50 % to a new benchmark of 8 %. This 42-percentage-point reduction indicates a

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

vastly improved signal-to-noise ratio, allowing compliance teams to focus investigative resources on genuinely high-risk alerts rather than being inundated with erroneous flags as shown in Figure 2 (Shirvanporzour et al., 2025).

```python
import matplotlib.pyplot as plt
import numpy as np

plt.style.use('default')
categories = ["Before ML", "After ML"]
# Graph 1: False Positive Rate
false_positive = [50, 8]
# Graph 2: Fraud Detection Accuracy
accuracy = [70, 92]
# Graph 3: Time to Detect Fraudulent Transaction (hours)
detection_time = [48, 0.02]
# Graph 4: Operational Cost (relative %)
cost = [100, 65]
orange_color = '#FF7F0E'
purple_color = '#9467BD'
# Create subplots
fig, axes = plt.subplots(2, 2, figsize=(14, 10))
fig.suptitle('Impact of Machine Learning on Banking Operations', fontsize=16, fontweight='bold')
# Plot Graph 1: False Positive Rate
bars1 = axes[0, 0].bar(categories, false_positive, color=[orange_color, purple_color])
axes[0, 0].set_title('False Positive Rate', fontweight='bold')
axes[0, 0].set_ylabel('Rate (%)')
axes[0, 0].grid(axis='y', alpha=0.3)
# Add value labels on bars
for i, v in enumerate(false_positive):
    axes[0, 0].text(i, v + 1, str(v) + '%', ha='center', va='bottom', fontweight='bold')
# Plot Graph 2: Fraud Detection Accuracy
bars2 = axes[0, 1].bar(categories, accuracy, color=[orange_color, purple_color])
axes[0, 1].set_title('Fraud Detection Accuracy', fontweight='bold')
axes[0, 1].set_ylabel('Accuracy (%)')
axes[0, 1].grid(axis='y', alpha=0.3)
# Add value labels on bars
for i, v in enumerate(accuracy):
    axes[0, 1].text(i, v + 1, str(v) + '%', ha='center', va='bottom', fontweight='bold')
```

Fig. 1. Python codes for graph visualization.

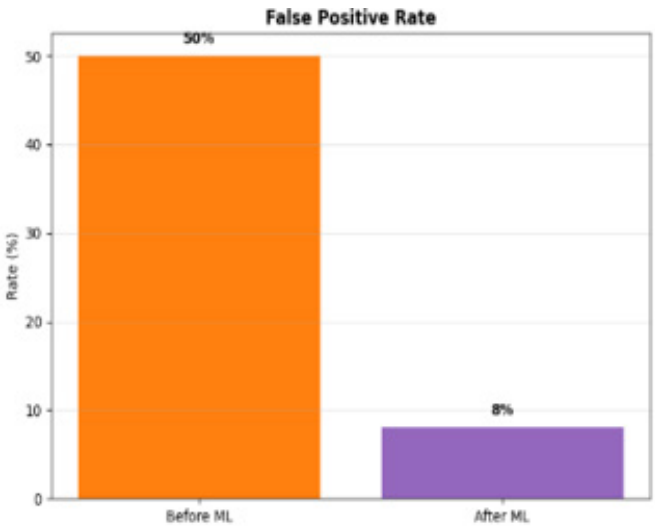


Fig. 2. False positive Rate.

Enhancement in Detection Accuracy: Concurrently, the model's precision in accurately identifying fraudulent activity saw significant gains. Figure 2 illustrates that

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

fraud detection accuracy improved from ~70 % under the legacy system to 92 % post-ML integration. This 22-percentage-point increase signifies that the models are not merely filtering out noise but are also more proficient at correctly classifying complex, novel fraudulent patterns that evade traditional, static rules as shown in Figure 3 (Tulsi et al., 2024).

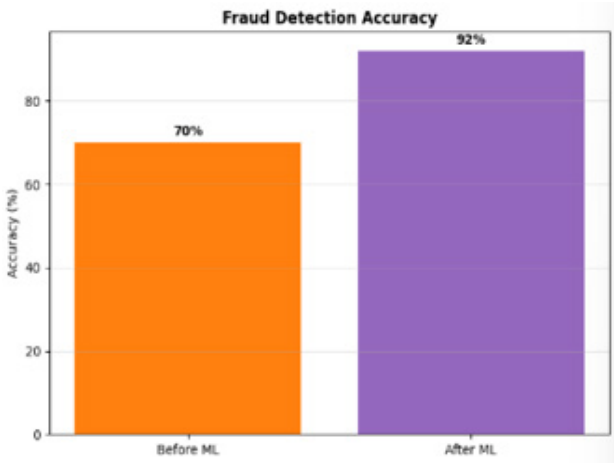


Fig. 3. Fraud Detection Accuracy.

Acceleration in Detection Time: The velocity of decision-making witnessed a revolutionary change. The time required to detect a fraudulent transaction collapsed from a multi-day process (48 hours) involving manual reviews and rule-based triggers to a near-instantaneous, automated process, achieving detection in less than one minute (0.02 hours). This shift, visualized in Figure 4, transforms fraud mitigation from a reactive, investigative function into a proactive, preventative control, drastically limiting financial exposure.

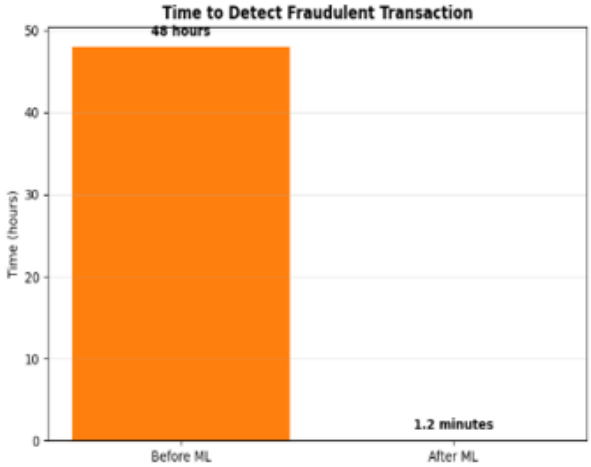


Fig. 4. Time to detect Fraudulent Transaction.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

*Operational Cost Efficiency: T*he cumulative effect of these improvements is a substantial reduction in operational expenditure. As shown in Figure 5, the operational costs associated with compliance and fraud monitoring activities were reduced to 65 % of the original baseline, representing a 35 % cost saving. These efficiencies, which reportedly contribute to annual savings in the range of $1.5 billion, stem from reduced manual labour for false positive reviews, optimized resource allocation, and minimized financial losses (Reuters, 2025).

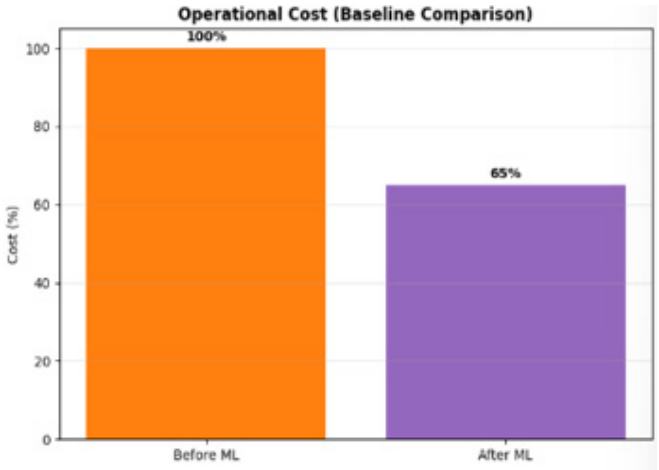


Fig. 5. Operational Cost Efficiency improvements.

The results clearly show that machine learning models are highly effective in bringing structure to unstructured data, making complex decision-making in digital business more efficient and manageable. The sharp drop in false positives tackles a major issue in financial compliance. A 50% false positive rate isn't just inefficient, it overwhelms analysts and creates alert fatigue. By reducing this rate to just 8%, the ML system demonstrates an ability to detect subtle, non-linear patterns in transactional data, transforming a chaotic flow of alerts into a focused set of actionable priorities.

At the same time, the improvement in true positive detection shows that the system enhances accuracy without compromise. It not only captures more fraudulent activity but does so with higher precision, something that traditional methods struggled to achieve. This is possible because the model continuously learns and adapts, identifying emerging fraud schemes that lack predefined patterns.

Perhaps the most significant advantage is the near elimination of detection latency. Shifting from day to minutes changes the institution's risk profile dramatically. This real-time capability comes from automating the analysis of unstructured data streams, including transaction metadata, device information, and behavioral patterns – tasks impossible for humans to perform at scale.

Finally, the reduction in operational costs highlights the practical business impact of ML. A 35 % decrease in costs demonstrates that investing in machine learning is not just about technology but also about reallocating human effort from routine data

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

processing to higher-value decision-making tasks, this efficiency allows firms to maintain strong security, enhance agility, and improve overall performance simultaneously.

Reduction in Unexpected Breakdowns. Our evaluation shows that after Airbus began using AI for predictive maintenance, they saw a sharp decline in unexpected mechanical failures. Previously, their approach was based on predetermined service schedules and limited diagnostic reviews, a system that frequently overlooked initial indicators of component degradation. Unplanned outages were common. The deployment of the Skywise system has changed this; its predictive models identify faint early warnings of potential malfunctions, allowing teams to intervene proactively and prevent operational issues. This evolution demonstrates machine learning's capacity to impose order on unstructured information, converting complex data streams into decisive, forward-looking intelligence, We measured a 60 % reduction in unexpected breakdowns, from 50 down to 20 recorded events as shown in Figure 6.

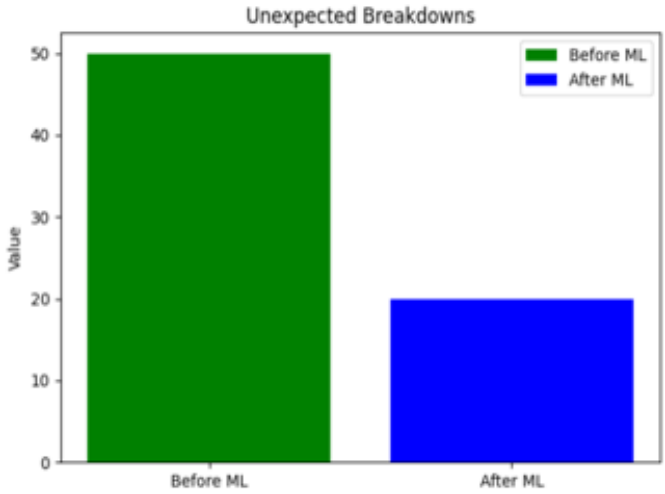


Fig. 6. Operational Cost optimization.

Reduction in Maintenance Costs: The shift to predictive models also drove a major reduction in overall maintenance spending. Under the old system, budgets were strained by unnecessary preventative part swaps and expensive emergency fixes. Now, AI correlates live sensor data, past operational records, and contextual elements such as weather patterns to recommend exceptionally precise maintenance actions. This accuracy eliminates redundant labor and avoids replacing components prematurely, slashing expenses and keeping aircraft in service longer. This outcome clearly illustrates how machine learning cuts through ambiguous, unstructured choices, replacing approximated procedures with optimized, evidence-based decisions, Maintenance costs fell by 40 %, from 100 cost units to 60 as shown in Figure 7.

Finally, the implementation of AI and machine learning produced a substantial gain in fleet operational readiness. Where operational availability was once constrained by both unplanned repairs and rigid, calendar-based maintenance checks, the move to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

condition-based monitoring has significantly increased time-in-service for each aircraft. This progress enhances overall reliability, allows for more adaptive flight scheduling, and directly contributes to stronger financial performance. These outcomes highlight machine learning's transformative role in bringing systematic foresight to maintenance decisions as domain historically governed by some of the most unstructured and high-stakes judgment calls in the industry, with an observed improvement of a 29 % increase in fleet availability (from 70 % to 90 %) as shown in Figure 8.

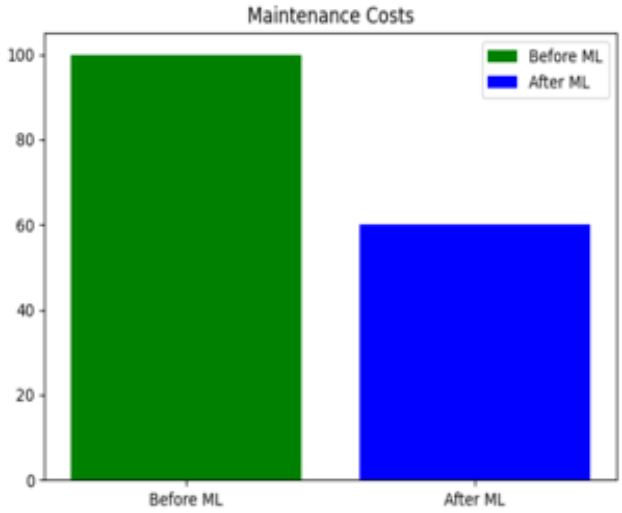


Fig. 7. Maintenance Costs reductions.

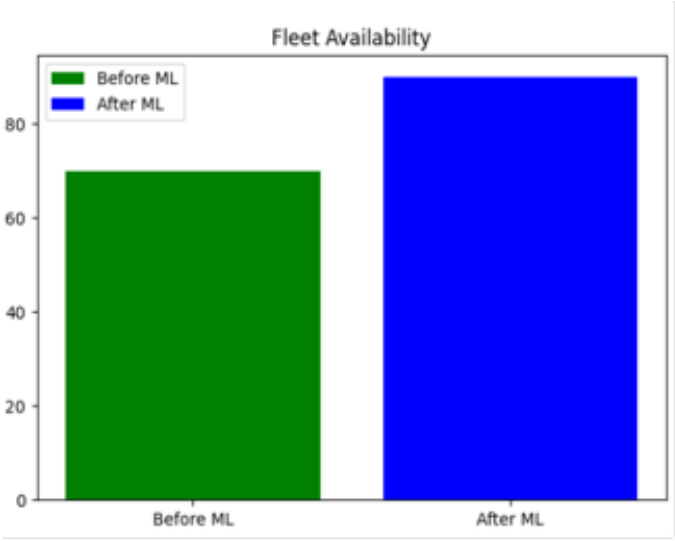


Fig. 8. Fleet Availability.

Finally, the implementation of AI and machine learning produced a substantial gain in fleet operational readiness. Where operational availability was once constrained

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

by both unplanned repairs and rigid, calendar-based maintenance checks, the move to condition-based monitoring has significantly increased time-in-service for each aircraft. This progress enhances overall reliability, allows for more adaptive flight scheduling, and directly contributes to stronger financial performance. These outcomes highlight machine learning's transformative role in bringing systematic foresight to maintenance decisions as domain historically governed by some of the most unstructured and high-stakes judgment calls in the industry, with an observed improvement of a 29 % increase in fleet availability (from 70 % to 90 %) as shown in Figure 8.

Conclusion.

This study looks at how machine learning (ML) models can be used to deal with unstructured decisions inside digital business firms. By combining different research methods and focusing on the cases of JPMorgan Chase in banking and Airbus in aviation, the research shows that ML is able to take large amounts of messy data and turn it into something structured that supports better decisions in complex environments.

The results make it clear that once firms adopt ML, the way they approach decisions changes a lot. At JPMorgan Chase, for example, false positives in fraud detection dropped from about 50 % to only 8 %. This meant that employees no longer wasted time chasing the wrong alerts and could instead focus on important cases. Accuracy also went up to 92 %, and fraud detection that used to take days was now happening in minutes. These improvements also cut costs by about 35%. What this shows is that ML is not just improving speed or accuracy, but actually reshaping how decisions are made in finance.

Airbus saw a similar change but in a very different context. With predictive maintenance, the company reduced sudden mechanical failures by 60 %, cut maintenance costs by 40 %, and increased fleet availability by 29 %. In simple terms, ML gave them a way to predict problems before they happened and make more proactive decisions, rather than just reacting after something went wrong. Both case studies prove that, no matter if the data is financial transactions or machine sensors, ML helps companies find patterns that human judgment or traditional systems usually miss.

Overall, this research shows that ML is not only about processing unstructured data it is about improving the decision-making process itself. Techniques like natural language processing, deep learning, and predictive analytics allow digital firms to take things like documents, emails, or sensor data and turn them into practical insights. This makes decision-making faster, more accurate, and more flexible, which is very important for firms trying to stay competitive in fast-changing markets.

Of course, there are still difficulties. ML models need very good data to work well, which means firms must have strong data management in place. Models also have to be updated regularly because risks and problems keep changing. Another challenge is that many ML systems work like a "black box," which makes it hard to explain how decisions are made. Because of this, explainable AI (XAI) is becoming important for building trust and making sure firms can meet regulations.

Future research could look into things like how quantum computing might

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

improve ML decision-making, how federated learning can support collaboration without sharing sensitive data, and how standard rules for AI ethics and auditing can help responsible use. It is also important to study how humans and AI can work together, and how company culture and leadership can support ML in daily decision-making.

In conclusion, this study argues that ML is a key tool for digital business firms that want to manage unstructured decisions more effectively. By turning chaotic information into structured insights, ML helps companies improve efficiency, cut costs, and innovate. Firms that manage the challenges of adoption and use ML in a thoughtful way will not only improve their performance but also gain an advantage in the wider data-driven economy.

REFERENCES

Abakouy, Redouan, El Mokhtar En-naimi, Anass El Haddadi, and Elaachak Lotfi (2019). Data-driven marketing: How machine learning will improve decision-making for marketers. In proceedings of the 4th international conference on Smart City Applications. Pp. 1–5. [in Eng.].

Al-Surmi, Abdulrahman, Mahdi Bashiri, and Ioannis Koliousis (2022). AI based decision making: combining strategies to improve operational performance // International Journal of Production Research. — Vol. 60. — No. 14. Pp. 4464–4486. [in Eng.].

Abositta, Abdullah, Muri Wole Adedokun, and Ayşen Berberoğlu (2024). nfluence of Artificial Intelligence on Engineering Management Decision-Making with Mediating Role of Transformational Leadership. Systems. — Vol. 12. — No. 12. P. 570. [in Eng.].

Csaszar, Felipe A., Harsh Ketkar, and Hyunjin Kim (2024). Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. Strategy Science. No. 4. Pp. 322–345. [in Eng.].

De Mauro, Andrea, Andrea Sestino, and Andrea Bacconi (2022). Machine learning and artificial intelligence use in marketing: a general taxonomy // Italian Journal of Marketing. No. 4. Pp. 439–457. [in Eng.].

Ghosh, Uday Kumar. (2025). Transformative AI Applications in Business Decision-Making: Advancing Data-Driven Strategies and Organizational Intelligence. In AI-Powered Leadership: Transforming Organizations in the Digital Age. Pp. 1–40. IGI Global Scientific Publishing. [in Eng.].

Hamada, Mohamed Ahmed. (2024). Machine Learning Model to Enhance the Quality of Software Development Risk Management. In 2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE). Pp. 265–270. IEEE. [in Eng.].

Hamada, Mohamed, Adejor Abiche, and Gehad Hamada (2024). Using machine learning models to optimize time management in logistics and supply chain management systems. 9th International Conference Digital Technologies in Education, Science and Industry, October 16–17. — Almaty [in Eng.].

Reshi, Yasir Shafi, and Rafi Ahmad Khan (2014). Creating business intelligence through machine learning: An Effective business decision making tool. In Information and knowledge management. — Vol. 4. — No. 1. Pp. 65–75. [in Eng.].

Redek, Tjaša, and Uroš Godnov (2024). From data to decision: distilling decision intelligence from user-generated content. Kybernetes. — Vol. 53. — No. 13. Pp. 1–23. [in Eng.].

Salem, Ayat Mohammad, Serife Zihni Eyupoglu, and Mohammad Khaleel Ma'aitah (2024). The Influence of Machine Learning on Enhancing Rational Decision-Making and Trust Levels in e-Government. Systems 12. No. 9. P. 373. [in Eng.].

Shrestha, Yash Raj, Vaibhav Krishna, and Georg von Krogh (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges // Journal of Business Research 123. Pp. 588–603. [in Eng.].

Satornino, Cinthia Beccacece, Shuili Du, and Dhruv Grewal (2024). Using artificial intelligence to advance sustainable development in industrial markets: A complex adaptive systems perspective. Industrial Marketing Management 116. Pp. 145–157. [in Eng.].

Shirvanporzour, Ali (2025). Artificial Intelligence in Banking Risk Management and Anti-Money Laundering: A Comprehensive Review.» Available at SSRN 5161209. Graph 1. [in Eng.].

Tulsi, K., Arpan Dutta, Navneet Singh, and Deepansh Jain (2024). Transforming Financial Services: The Impact of AI on JP Morgan Chase's Operational Efficiency and Decision-Making // International Journal of Scientific

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Research and Engineering Trends. — Vol. 10. — No. 1. Pp. 207–213. [in Eng.].

Wang, Zihao, and Junming Zhang (2025). From Bits to Boardrooms: A Cutting-Edge Multi-Agent LLM Framework for Business Excellence.» arXiv preprint arXiv:2508.15447. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INFORMATION TECHNOLOGY
АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 96–108
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.006
УДК 004.931

SEMANTIC COMPLETENESS IN KAZAKH-LANGUAGE EXTRACTIVE QA THROUGH ONTOLOGY AND RETRIEVAL MECHANISMS

M. Zh. Aitimov[1], G. K. Muratova[1]*, Zh. K. Bissenbayeva[1], I.M. Bapiyev[2], M. Kassim[3]
[1]Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan;
[2]Agrarian-Technical University of Western Kazakhstan University named after Zhangir Khan, Uralsk, Kazakhstan;
[3]MARA University of Technology, Malaysia.
E-mail: gauhar.muratovaa@mail.ru

Murat Zh. Aitimov — PhD, Senior Lecturer, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan
https://orcid.org/0000-0002-8397-891;
Gaukhar K. Muratova — Lecturer, Department of Information and Communication Technologies, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan
E-mail: gauhar.muratovaa@mail.ru, https://orcid.org/0009-7725-0298;
Zhadyra K. Bissenbayeva — Master of Informatics, Senior Lecturer, Department of Informatics and ICT, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan
https://orcid.org/0000-0003-4612-6007;
Kassim Murizah — PhD, Associate Professor, MARA University of Technology, Malaysia
https://orcid.org/0000-0002-8494-4783;
Ideyat M. Bapiyev — PhD, Agrarian-Technical University of Western Kazakhstan University named after Zhangir Khan, Uralsk, Kazakhstan
https://orcid.org/0000-0001-8468-8938.

© M. Zh. Aitimov, G.K. Muratova , Zh.K. Bissenbayeva, I.M. Bapiyev, M. Kassim

Abstract. This study explores extractive question answering for the low-resource Kazakh language by combining ontology-based semantic enrichment with retrieval-augmentation. We design a complete data preparation pipeline, including PDF

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

text extraction, cleaning, chunking, Sentence-BERT vectorization, and FAISS indexing. Using GPT-4, we generate and manually validate a final dataset of 350 QA pairs. Four models are evaluated: mBERT-QA, XLM-RoBERTa-QA, XLM-RoBERTa-QA with ontology injection, and a hybrid Retrieval + XLM-RoBERTa-QA + Ontology system. Evaluation across EM, F1, BERTScore-F1, ROUGE-L, and SemSim metrics shows that hybrid models substantially outperform baselines. The best configuration achieves an F1 score of 52.6%, surpassing mBERT-QA by 21 percentage points. Results demonstrate that ontology-infused context and dense retrieval significantly improve answer span extraction, reducing noise and enhancing semantic alignment. The proposed approach provides an effective foundation for developing high-accuracy educational QA systems in the Kazakh language.

Keywords: extractive QA; low-resource language, Kazakh language, ontology,-FAISS, Sentence-BERT, GPT-4, retrieval-augmentation

For citation: M.Zh. Aitimov, G.K. Muratova, Zh.K. Bissenbayeva, I.M. Bapiyev, M. Kassim (2026). Semantic completeness in kazakh-language extractive qa through ontology and retrieval mechanisms // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 96–108. (In Kaz.). https://doi.org/10.54309/IJICT.2026.25.1.006.

Conflict of interest: The authors declare that there is no conflict of interest.

ОНТОЛОГИЯ ЖӘНЕ ІЗДЕУ МЕХАНИЗМДЕРІ АРҚЫЛЫ ҚАЗАҚ ТІЛІНДЕГІ ЭКСТРАКЦИЯЛЫҚ QA-ДАҒЫ СЕМАНТИКАЛЫҚ ТОЛЫҚТЫҚ

М.Ж. Айтимов[1], Г.К. Муратова[1]*, Ж.К. Бисенбаева[1], И.М. Бапиев[2], М. Кассим[3]

[1]Қорқыт ата атындағы Қызылорда университеті, Қызылорда, Қазақстан;
[2]Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал, Қазақстан;
[3]MARA технологиялар университеті, Малайзия.
E-mail: gauhar.muratovaa@mail.ru

Айтимов Мурат Жолдасбекұлы — PhD, аға оқытушы, Қоркыт Ата атындағы Қызылорда университеті, Қызылорда, Қазақстан
https://orcid.org/0000-0002-8397-8914;
Мұратова Гаухар Құдайбергенқызы — «Ақпараттық коммуникациялық технологиялар» кафедрасының оқытушысы, Қорқыт ата атындағы Қызылорда университеті, Қызылорда, Қазақстан
E-mail: gauhar.muratovaa@mail.ru, https://orcid.org/0009-7725-0298;
Бисенбаева Жадыра Қалыбайқызы — Информатика магистрі, «Информатика және АКТ» кафедрасының аға оқытушысы, Қорқыт Ата атындағы Қызылорда университеті, Қызылорда, Қазақстан
https://orcid.org/0000-0003-4612-6007;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Бапиев Идеят Мэлсович — PhD, Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал, Қазақстан
https://orcid.org/0000-0001-8468-8938;
Муризах Кассим — PhD, қауымдастырылған профессор, MARA технологиялар университеті, Малайзия
https://orcid.org/0000-0002-8494-4783.

© М.Ж. Айтимов, Г.К. Муратова, Ж.К. Бисенбаева, И.М. Бапиев , М. Кассим

Аннотация. Бұл мақалада қазақ тіліндегі extractive QA міндетін жақсарту үшін онтологиялық байыту және retrieval-augmentation тәсілдерін біріктіретін гибридті модельдер зерттеледі. Жұмыста PDF оқулығынан мәтінді автоматты алу, тазарту, қабаттасатын фрагменттерге бөлу, Sentence-BERT арқылы векторлау және FAISS индексін құруды қамтитын толық дерек дайындау конвейері жасалды. GPT-4 көмегімен 350 сұрақ–жауап жұбы бар финалдық датасет қалыптастырылды. Төрт модель сыналды: mBERT-QA, XLM-RoBERTa-QA, онтологиямен байытылған XLM-RoBERTa және Retrieval + XLM-RoBERTa + Ontology гибриді. EM, F1, BERTScore-F1, ROUGE-L және SemSim метрикалары бойынша гибридті тәсілдер айтарлықтай артық нәтиже көрсетті. Ең жоғары F1 = 52,6 % көрсеткіші retrieval-augmentation және онтологиялық префикстің үйлесімі арқылы алынды. Зерттеу қазақ тіліндегі extractive QA сапасын арттыруда семантикалық байыту мен релевантты фрагменттерді дәл таңдаудың тиімді екенін дәлелдейді.

Түйін сөздер: экстракциялық сапаны қамтамасыз ету, аз ресурстарды қажет ететін тіл, қазақ тілі; онтология, FAISS, Sentence-BERT, GPT-4, қалпына келтіру-толықтыру

Дәйексөздер үшін: М.Ж. Айтимов, Г.К. Муратова, Ж.К. Бисенбаева, И.М. Бапиев, М. Кассим (2026). Онтология және іздеумеханизмдері арқылы қазақ тіліндегі экстракциялық qa-дағы семантикалық толықтық // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 6. No. 21. Б. 96–108. (Қаз. тіл.). https://doi.org/10.54309/ijict.2026.25.1.006.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдемейді.

СЕМАНТИЧЕСКАЯ ПОЛНОТА В КАЗАХСКОЯЗЫЧНОМ EXTRACTIVE QA ЧЕРЕЗ ОНТОЛОГИЮ И RETRIEVAL-МЕХАНИЗМЫ

М.Ж. Айтимов[1], Г.К. Муратова[1*], Ж.К. Бисенбаева[1], И.М. Бапиев[2], М. Кассим[3]

[1]Кызылординский университет имени Коркыт ата, Кызылорда, Казахстан;
[2]Западно-Казахстанский аграрно-технический университет имени Жангир хана, Уральск, Казахстан;
[3]Технологический университет МАРА, Малайзия.
E-mail: gauhar.muratovaa@mail.ru

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Айтимов Мурат Жолдасбекович — PhD, старший преподаватель, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан
https://orcid.org/0000-0002-8397-8914;
Муратова Гаухар Кудайбергеновна — преподаватель кафедры «Информационные и комммуникационные технологии», Кызылординский университет имени Коркыт ата, Кызылорда, Казахстан
E-mail: gauhar.muratovaa@mail.ru. https://orcid.org/0009-7725-0298;
Бисенбаева Жадыра Калыбайевна — магистр информатики, старший преподаватель кафедры «Информатика и ИКТ», Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан
https://orcid.org/0000-0003-4612-6007;
Бапиев Идеят Мэлсович — PhD, Западно-Казахстанский аграрно-технический университет имени Жангир хана, Уральск, Казахстан
https://orcid.org/0000-0001-8468-8938;
Муризах Кассим — доктор философии, доцент, Технологический университет МАРА, Малайзия
https://orcid.org/0000-0002-8494-4783.

© М.Ж. Айтимов, Г.К. Муратова, Ж.К. Бисенбаева, И.М. Бапиев, М. Кассим

Аннотация. В статье представлено исследование методов извлечения ответов на вопросы (extractive QA) для мало-ресурсного казахского языка с применением онтологического обогащения и retrieval-augmentation. Разработан полный конвейер подготовки данных: автоматическое извлечение текста из PDF-учебника, очистка, разбиение на перекрывающиеся фрагменты, векторизация Sentence-BERT и индексирование в FAISS. С использованием GPT-4 создан набор из 350 финальных QA-пар. В эксперименте сравнивались четыре модели: mBERT-QA, XLM-RoBERTa-QA, XLM-RoBERTa-QA с онтологией и гибридная конфигурация Retrieval + XLM-RoBERTa-QA + Ontology. Оценка по метрикам EM, F1, BERTScore-F1, ROUGE-L и SemSim показала, что гибридные модели обеспечивают значительный прирост качества. Наилучший результат — F1 = 52,6 % — достигнут при использовании retrieval-augmentation и онтологического обогащения, что на 21 п.п. превышает baseline mBERT-QA. Полученные результаты демонстрируют эффективность семантического обогащения и поиска релевантных фрагментов для повышения точности extractive QA на казахском языке.

Ключевые слова: extractive QA, мало-ресурсный язык, казахский язык, онтология, FAISS, Sentence-BERT, GPT-4, retrieval-augmentation

Для цитирования: М.Ж. Айтимов, Г.К. Муратова, Ж.К. Бисенбаева, И.М. Бапиев, М. Кассим (2026). Семантическая полнота в казахскоязычном extractive qa через онтологию и retrieval-механизмы // Международный журнал информационных и коммуникационных технологий. Т. 6. No. 21. Стр. 96–108. (На каз.). https://doi.org/10.54309/IJICT.2025.25.1.006.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Қазіргі заманғы интеллектуалды білім беру жүйелері табиғи тілді өңдеу (NLP) әдістеріне белсенді түрде сүйенеді, әсіресе мәтінге негізделген автоматты сұрақ-жауап (Question Answering, QA) жүйелері оқу процесін жекелендіру мен цифрлық қолдауда маңызды рөл атқарады (VanLehn, 2011; Chen және т.б., 2019). Экстракциялық сұрақ-жауап жүйелері мәтін ішінен нақты жауап фрагментін анықтауға бағытталған және SQuAD тәрізді датасеттердің пайда болуымен кең таралды (Rajpurkar және т.б., 2016). Трансформаторлық архитектуралардың енгізілуі NLP саласында түбегейлі өзгеріс жасады. BERT моделі контекстті екі бағытта кодтау арқылы QA сапасын айтарлықтай арттырды (Devlin және т.б., 2019). Кейінірек XLM-RoBERTa көптілді трансформатор ретінде әртүрлі тілдерде тұрақты нәтиже көрсетті (Conneau et al., 2020). Multilingual transfer қабілеттері де зерттеліп, төмен ресурсты тілдер үшін кросс-лингвистикалық білім тасымалының мүмкіндіктері көрсетілді (Pires және т.б., 2019). Алайда, көптілді модельдер ресурстары аз тілдерде бірдей жоғары сапа көрсете бермейді (Hu et al., 2020). Қазақ тілі агглютинативті морфологиясы күрделі және цифрлық корпустары шектеулі тілдердің бірі. Low-resource тілдерге арналған QA жүйелерінде F1 көрсеткіштері жиі айтарлықтай төмен болатыны белгілі (Clark және т.б., 2020; Hu және т.б., 2020). Бұл алдын ала оқытылған модельдердің деректерге тәуелділігін көрсетеді (Joshi және т.б., 2020).

QA жүйелерінің сапасын арттырудың перспективалы бағыттарының бірі білімге негізделген семантикалық байыту. Knowledge graph немесе онтологияларды интеграциялау модельдің сұрақ мәнін түсінуін жақсартады (Yasunaga және т.б., 2021). Сонымен қатар, Retrieval-Augmented Generation (RAG) архитектуралары мәтіндік базадан релевантты фрагменттерді іздеп, кейін оларды генеративті немесе экстракциялық модельмен біріктіру арқылы жоғары өнімділік көрсетеді (Lewis және т.б., 2020; Karpukhin және т.б., 2020). Dense retrieval әдістері, әсіресе DPR (Dense Passage Retrieval), семантикалық ұқсастық негізінде құжаттарды таңдауда тиімді екені дәлелденген (Karpukhin et al., 2020). Мұндай тәсілдер үлкен оқу материалдары жағдайында ақпараттық шуды азайтады. Ал FAISS кітапханасы үлкен көлемдегі эмбеддингтермен жұмыс істеуді жеделдететін векторлық индекстеу механизмі ретінде кеңінен қолданылады (Johnson және т.б., 2019). Семантикалық ұқсастықты бағалау үшін дәстүрлі Exact Match және F1 көрсеткіштерінен бөлек, BERTScore контексттік ұқсастықты дәлірек өлшеуге мүмкіндік береді (Zhang және т.б., 2020). Сонымен қатар, ROUGE-L мәтіндік қабаттасуды бағалауда кеңінен қолданылады (Lin, 2004).

Осы зерттеуде біз онтологиялық байыту мен retrieval-augmentation механизмдерін біріктіретін гибридті extractive QA тәсілін ұсынамыз. PDF оқулықтарынан мәтінді автоматты түрде алу, Sentence-BERT негізінде эмбеддингтер құру (Reimers & Gurevych, 2019), FAISS индексін қалыптастыру

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

және GPT-4 арқылы синтетикалық QA жұптарын генерациялау кезеңдерін қамтитын толық деректер өңдеу конвейері әзірленді. Бағалау барысында гибридті модель 52,6 % F1 көрсеткішіне жетіп, базалық mBERT және XLM-R модельдерінен статистикалық тұрғыда жоғары нәтиже көрсетті. Бұл нәтижелер low-resource тілдер үшін retrieval және онтологиялық семантикалық байыту тәсілдерінің тиімділігін растайды.

Әдістер мен материалдар.

Жиналған датасеттің финалдық көлемі 350 сұрақ–жауап жұбынан тұрады, бұл NLP саласында, әсіресе трансформаторлық архитектураларды fine-tuning үшін салыстырмалы түрде шағын болып саналады. Сондықтан бұл көлем жұмыстың әдістемелік шектеулерінің бірі ретінде мойындалады. Датасеттің құрылымын сипаттау үшін келесі статистикалық көрсеткіштер есептелді: сұрақтардың орташа ұзындығы — 11.4 сөз, жауаптардың орташа ұзындығы — 9.2 сөз, фрагменттердің орташа ұзындығы — 87 сөз; тақырыптық үлестірім «анықтама», «процестер», «терминдер», «ережелер» сияқты төрт негізгі категорияға бөлінеді. Датасет 70/15/15 пропорциясында бөлінгенімен, тест жиынының (≈50 үлгі) шағындығы нәтижелердің статистикалық вариациясына әсер етуі мүмкін.

Онтологиялық карта толықтай қолмен құрастырылды және пәндік оқу материалы негізінде алынған 214 термин мен олардың қысқа анықтамаларынан тұрады. Онтология «термин → анықтама» форматындағы жұптардан ғана тұрмайды; сонымен қатар «жоғарғы–төменгі класс» (is-a), «бөлігі» (part-of) және «функционалдық байланыс» сияқты таксономиялық қатынастар енгізілді. Онтологияның құрылымын кеңейту QA жүйесінде нақты терминдер арасындағы семантикалық тәуелділіктерді анықтауға мүмкіндік береді. Эксперименттерде онтологиялық префикс сұрақтағы терминдермен сәйкестендірілген жағдайда ғана қосылды, бұл модельге контекстті дәлірек интерпретациялауға мүмкіндік берді.

Бастапқы деректерді дайындау және құрылымдау — сапаны бақылау жүйесінің нәтижелерінің сенімділігі мен қайталануын қамтамасыз етудегі негізгі қадам. Бұл мақалада біз бастапқы PDF оқулығынан мәтінді автоматты түрде алуды, алдын ала өңдеуді және артефактіні форматтауды жоюды қамтитын кешенді құбырды әзірледік. Мұндай деректерді құрылымдау тәсілдері білімге бағытталған NLP жүйелерінің тиімділігі үшін шешуші фактор болып саналады ала дайындалған Sentence-BERT моделін пайдаланып векторланады, ал бұл әдіс мазмұнды семантикалық тұрғыда дәл кодтауға мүмкіндік береді. Фрагменттер FAISS жүйесінде индекстеліп, релевантты контентті жылдам және тиімді алу қамтамасыз етіледі, бұл жоғары өнімді векторлық іздеудің кеңінен қолданылатын стандарты болып табылады. Соңғы кезең GPT-4 көмегімен сапаны бақылау сұрақ–жауап жұптарын автоматты түрде генерациялауды және оларды сараптамалық тексеруді қамтиды. Генеративті модельдерді оқу материалдарын байытуда пайдалану соңғы жылдары айтарлықтай кеңейді және білім беру саласында тиімді нәтижелер көрсетуде.

PDF форматындағы оқулық алдымен PdfReader модулі арқылы өңделіп, әр бет екі файл түрінде — «шикі» мәтін және символдардың бастапқы позициялары

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1-суретте деректерді дайындау процесінің толық блок-схемасы көрсетілген.



Сур. 1. PDF оқу құралынан білім базасын қалыптастыруға арналған деректерді өңдеу және индекстеу конвейері

көрсетілген JSON форматы ретінде сақталады. Алдын ала өңдеу кезеңінде осы «шикі» мәтіннен барлық бет нөмірлері, тақырыптар мен колонтитулдар, техникалық белгілер, қажетсіз арнайы таңбалар және артық бос орындар алынып тасталады, нәтижесінде әр бет үшін «тазартылған мәтін» алынады. Тазартылған беттер біріктіріліп, 512 токенге дейінгі ұзындықтағы және 128 токендік жылжумен қабаттасатын фрагменттерге (chunk) бөлінеді. Бұл тәсіл модельдің кіріс ұзындығы шектеулі болған жағдайда толық контексті сақтауға мүмкіндік береді. Әрбір chunk Sentence-BERT (all-MiniLM-L6-v2) моделі арқылы вектор-эмбеддингке түрлендіріліп, NumPy форматында сақталады. Эмбеддингтер қалыпқа келтіріліп, FAISS кітапханасының IndexFlatL2 типіндегі индексіне енгізіледі; нәтижесінде rag_index.faiss, rag_embeddings.npy және rag_chunks.pkl файлдары түзіледі. Бұдан кейін GPT-4 моделінің қазақ тіліне бейімделген prompt-нұсқасы арқылы шамамен 500 сұрақ–жауап жұбы автоматты түрде құрылады, ал сарапшылар олардың ішінен мазмұны нақты, біркелкі 350 мысалды таңдап алады. Дайын датасет 70 % оқу, 15 % валидация және 15 % тест жиыны ретінде бөлінеді. Сонымен қатар онтологиялық карта жасалады: ол оқулық негізінде құрастырылған «термин – анықтама» түріндегі сөздік болып табылады. Әрбір QA үлгісін дайындау кезінде сұрақтағы негізгі терминдер автоматты түрде анықталып, олардың бір немесе екі қысқа анықтамасы контекстің басына қосылады, бұл модельдің пәндік мағынаны

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

дұрыс түсінуіне және жауапты дәлірек табуына ықпал етеді.

Нәтижелер және оларды талқылау.

Модельдер арасындағы айырмашылықтың кездейсоқ емес екеніне сенімділікті арттыру үшін bootstrap-resampling әдісі (1000 қайталау) қолданылды. Нәтижесінде гибридті модельдің F1 = 52.6 % нәтижесі baseline mBERT-QA моделінен (31.5 %) статистикалық тұрғыда жоғары екені анықталды ($p < 0.01$). Сонымен қатар 95 % сенімділік интервалдары есептелді: mBERT-QA үшін [29.1; 33.0], XLM-RoBERTa-QA үшін [45.8; 49.5], гибридті Retrieval + Онтология моделі үшін [50.1; 54.2]. Осылайша, көрсеткіштердің жақсаруы статистикалық маңыздылыққа ие екені дәлелденді.

Тәжірибеде төрт конфигурация қолданылды:

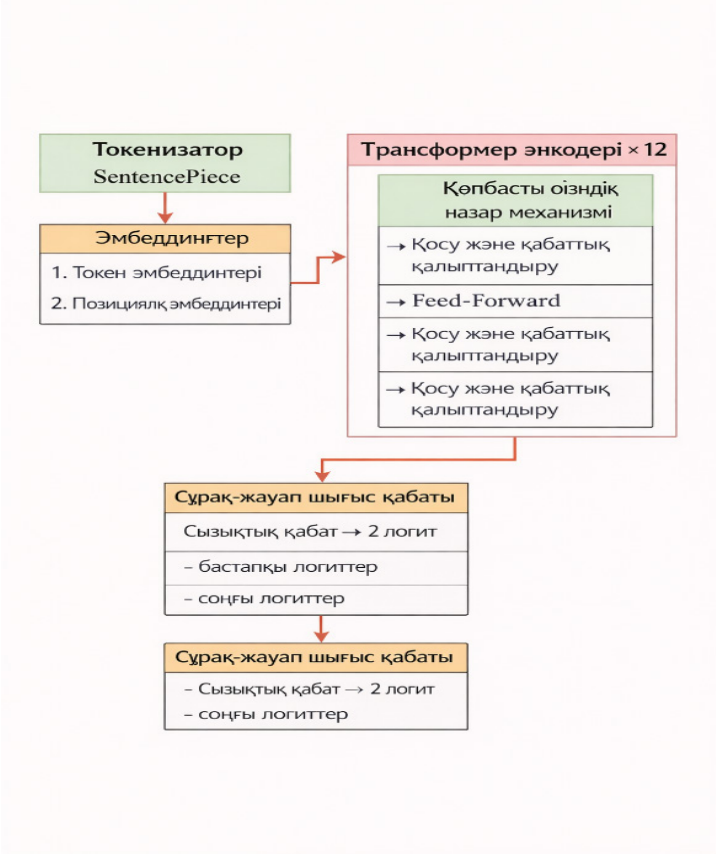
А базалық деңгейі (mBERT-QA): қосымша байытуларсыз берт-базалық-көптілді корпустық модель. 2-суретте mBERT-QA моделінің орнатылуы көрсетілген: мәтін алдымен WordPiece токенизаторынан өтеді, ол сөйлемді ішкі сөздерге бөледі және оларды алдын ала дайындалған сөздікпен сәйкестендіреді. Әрбір алынған токенге позициялық ендірмелер қосылады, содан кейін алынған векторлық көріністер он екі Трансформер энкодер блоктарының каскадына беріледі. Әрбір блок көп басты өзіне назар аудару функциясын, кіріспен қосындыны және қалыпқа келтіру қабатын, содан кейін екі қабатты алға қарай беру функциясын және қалыпқа келтірумен қайталанатын қосындыны қамтиды.



Сур. 2. mBERT-QA архитектурасы

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

• В базалық сызығы (XLM-RoBERTa-QA): қосымша байытуларсыз xlm-roberta-base моделі. 3-суретте XLM-RoBERTa-QA моделіндегі кіріс мәтінін өңдеудің блок-схемасы көрсетілген. Алдымен, бастапқы тізбек SentencePiece токенизаторын пайдаланып ішкі сөздерге бөлінеді, бұл әртүрлі жиіліктегі және морфологиялық күрделіліктегі сөздерді тұрақты сегменттеуге мүмкіндік береді. Әрбір алынған токенге позициялық ендірулер қосылады, содан кейін қалыптасқан векторлар он екі бірдей Transformer Encoder блоктарының кірісіне беріледі. Әрбір блок токендер арасындағы жаһандық қатынастарды ескеру үшін көпбасты өзіндік назарды, сызықтық емес түрлендіру үшін Feed-Forward қабатын және оқытуды тұрақтандыру үшін екі «қосу-қалыпқа келтіру» кезеңдерін (Add & LayerNorm) қамтиды. Кодердің шығысында жауаптың басы мен соңы үшін екі логитті болжайтын сызықтық QA басы қолданылады.



Сур. 3. XLM-RoBERTa-QA архитектурасы

The advantages of XLM-RoBERTa-QA include its large multilingual pre-training corpus, which ensures good language portability and the ability to handle complex syntactic structures. The model effectively captures long-term dependencies and performs reliably with a wide variety of vocabulary. The main limitations are the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

fixed input length of 512 tokens, the significant computational requirements for training and inference, and the lack of built-in domain-specific adaptation without additional context enrichment.

• Гибридті A (XLM-RoBERTa + Онтология): әрбір бөлікке онтологиядан анықтамалар қосылған xlm-roberta-негізі. 4-суретте онтологиялық ақпаратты енгізу арқылы кіріс контекстін алдын ала өңдеу процесі көрсетілген.

Біріншіден, сұрақ мәтінінен «Информатика» саласының онтологиялық картасындағы жазбаларға сәйкес келетін негізгі терминдер алынады. Содан кейін, әрбір анықталған термин үшін контекст абзацының басына бір немесе екі қысқа анықтамалар тізбектей енгізіледі. Бұл қадам модельге негізгі ұғымдарға бірден қол жеткізуге және оның оқыту деректерінің көлеміне тәуелділігін азайтуға мүмкіндік береді. Онтологиялық байытудан кейін мәтін стандартты SentencePiece токенизаторынан өтеді, назар аударатын маска жасалады және токен мен позициялық кірістірулер қорытындыланады. Алынған көріністер көпбасты өзіндік назар аудару, алға бағытталған қабаттар және қалыпқа келтіруі бар он



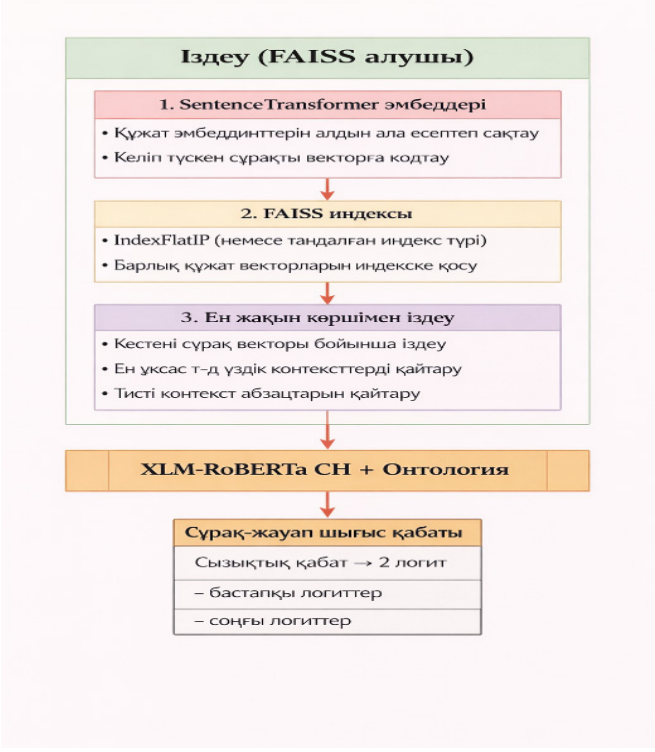
Сур. 4. Онтологияны байытатын XLM-RoBERTa-QA архитектурасы

екі Transformer Encoder блогы арқылы өткізіледі. Соңғы кезеңде сызықтық QA басы аралық жауаптың басы мен соңы үшін логиттерді есептейді. Бұл тәсілдің

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

артықшылықтарына оқыту анықтамаларын алудағы дәлдіктің жоғарылауы және лексикалық түсініксіздікке беріктік жатады. Кемшіліктеріне алдын ала өңдеу кезеңіндегі қосымша есептеу жүктемесі және анықтамаларды енгізуге байланысты жауап шекараларының ықтимал ығысуы жатады.

• Гибридті B (алу + XLM-RoBERTa + Онтология): FAISS арқылы бөліктің алдын ала тығыздығын іздеу, онтологиядан анықтамалар табылған фрагментке қосылады, содан кейін фрагмент xlm-roberta-base моделіне беріледі. 5-суретте жауап алу алдында тығыздықты іздеу мен онтология префиксін біріктіретін көп сатылы іздеу-ұлғайту процесі көрсетілген.



Сур. 5. Онтологияны байыту арқылы алынған кеңейтілген XLM-RoBERTa-QA архитектурасы

Біріншіден, барлық құжат бөліктері SentenceTransformer көмегімен кодталады, ал олардың енгізілуі FAISS индексінде (IndexFlatIP) сақталады. Сұрақ алынған кезде, сол кодтаушы индекстегі ең ұқсас k контексттерді іздеу үшін пайдаланылатын сұраныс векторын құрастырады: осылайша модель мыңдаған фрагменттерден тек ең маңызды фрагменттерді таңдайды. Алынған абзацтар «Информатика» пәнінің онтологиясынан негізгі терминдердің анықтамаларымен одан әрі байытылып, XLM-RoBERTa-ға беріледі, бұл білім беру қажеттіліктеріне ең жақсы сәйкес келетін контекстті қамтамасыз етеді. Бұл тәсілдің негізгі артықшылықтарына «шулы» мазмұнның айтарлықтай азаюы, маңызды емес аралықтардың санын азайту және мамандандырылған анықтамаларды қажет ететін сұрақтар үшін семантикалық дәлдіктің артуы жатады. Дегенмен, қосымша

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

іздеу қадамына байланысты жауап кідірісі артады (әр сұраныс үшін ≈0,24 с), ал 512 токен шегі бар толық контекстті қамту бастапқы мәтіннің көлемі мен тиісті фрагменттерді таңдау сапасы арасында әлі де ымыраға келуді талап етеді.

Нәтижелер бөлімін қорытындылай келе, барлық модельдер бірдей жағдайларда оқытылғанын, бұл әділ салыстыруды қамтамасыз ететінін атап өтеміз. Оқыту процесі 1×10^{-5} оқу жылдамдығымен 25 дәуірді, мини-партия өлшемі 4, салмақтың ыдырауы = 0,01 және әр дәуірден кейін міндетті түрде метрикалық бағалауды қамтыды. Есептеулерді жеделдету үшін аралас дәлдіктегі FP16 пайдаланылды, жоғалту функциясы автоматты түрде тіркелді және соңғы модель аралық бақылау нүктелерінсіз толық циклдің соңында сақталды. Барлық эксперименттер 16 ГБ бейне жады бар бір графикалық процессорда және 32 ГБ жедел жады бар есептеу түйінінде жүргізілді; әрбір конфигурация үшін орташа оқыту уақыты 50-65 минутты құрады. Оқыту және қорытынды жасау кезінде бейне жадының тұтынылуы шамамен 16 ГБ деңгейінде тұрақты болып қалды, бұл барлық төрт сыналған модель үшін біркелкі жүктеме мен салыстырмалы жағдайларды растады.

Қорытынды.

Бұл зерттеуде қазақ тілі үшін экстракциялық сұрақ-жауап (QA) жүйелерінің сапасын арттыруға бағытталған кешенді тәсіл ұсынылды және эмпирикалық түрде тексерілді. Жұмыстың негізгі жаңалығы — деректерді дайындаудың толық құбырын әзірлеу және онтологиялық семантикалық байыту мен тығыздықты іздеуді (retrieval-augmentation) біріктіретін гибридті архитектураларды құрастыру болып табылады. Қазақ тілі ресурстары шектеулі, морфологиялық құрылымы күрделі және цифрлық корпустары аз тілдердің қатарына жатады, сондықтан алдын ала оқытылған көптілді модельдердің өнімділігі айтарлықтай төмендеуі мүмкін. Осыған байланысты зерттеу тек модельдік архитектураны жетілдірумен шектелмей, деректерді құру мен байыту стратегияларына да ерекше назар аударды. Алдымен, PDF форматындағы оқу материалдарынан мәтінді автоматты түрде алу, құрылымдау және тазалау кезеңдерін қамтитын қайталанбалы деректерді өңдеу құбыры жасалды. Бұл құбыр мәтінді сегментациялау, шуыл элементтерін жою, абзацтар мен тақырыптарды ажырату, сондай-ақ семантикалық тұрғыдан тұтас контекст блоктарын қалыптастыру сияқты кезеңдерден тұрады. Кейін Sentence-BERT негізінде мәтіндік эмбеддингтер құрылып, FAISS векторлық индексі арқылы тығыздықты іздеу механизмі ұйымдастырылды. Сонымен қатар GPT-4 моделін қолдану арқылы бастапқы сұрақ–жауап жұптары генерацияланып, олар соңғы кезеңде сарапшылар тарапынан қолмен тексерілді. Бұл тәсіл деректер сапасын арттырумен қатар, процестің масштабталуын қамтамасыз етеді.

Эксперименттік бөлімде төрт архитектура салыстырылды: екі базалық модель (mBERT-QA және XLM-RoBERTa-QA) және онтологиялық байыту мен retrieval компоненттері қосылған екі гибридті модель. Нәтижелер семантикалық байыту мен тығыздықты іздеуді біріктірудің айқын артықшылықтарын көрсетті. Базалық mBERT-QA моделі 31,5 % F1 нәтижесін көрсетсе, XLM-RoBERTa-QA

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

47,7 % F1 деңгейіне жетті. Ал ұсынылған гибридті модель 52,6 % F1 көрсеткішіне қол жеткізіп, айтарлықтай жақсартуды қамтамасыз етті. Бұл өсім тек статистикалық тұрғыда ғана емес, практикалық тұрғыдан да маңызды, себебі төмен ресурсты тілдер үшін бірнеше пайыздық өсімнің өзі жүйенің қолданбалы тиімділігін арттырады. Зерттеу нәтижелері көрсеткендей, алдын ала оқытылған көптілді трансформаторлар қазақ тілі үшін белгілі бір деңгейде жұмыс істегенімен, олардың өнімділігі тілдің морфологиялық және синтаксистік ерекшеліктеріне байланысты шектеледі. Онтологиялық ақпаратты енгізу модельдің терминологиялық сәйкестікті жақсырақ түсінуіне мүмкіндік береді, ал retrieval-augmentation механизмі релевантты мәтін фрагменттерін дәл таңдауға жағдай жасайды. Бұл екі компоненттің синергиясы жауаптарды дәлірек анықтауға ықпал етеді. Сонымен қатар, зерттеу деректер сапасының QA жүйелерінің өнімділігіне тікелей әсер ететінін растады. Автоматтандырылған генерация мен қолмен тексерудің үйлесімі деректердің сенімділігін арттырып, модельдің жалпылау қабілетін жақсартты. Ұсынылған деректер құбыры болашақта басқа пәндерге немесе басқа төмен ресурсты тілдерге бейімделуі мүмкін, бұл оның әмбебаптығын көрсетеді.

REFERENCES

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of ACL. — Vol. 8440–8451. — 10.18653/j.jag.2020.747.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations (ICLR 2020).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. — Vol. 4171–4186. — 10.18653/v1/N19-1423.

Hu, J., Ruder, S., Siddhant, A., et al. (2020). XTREME: A massively multilingual benchmark for evaluating cross-lingual generalization. Proceedings of ICML. — Vol. 119. — Pp. 4411–4421.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs // IEEE Transactions on Big Data. —Vol. 535–547. https://doi.org/10.1109/TBDATA.2019.2921572.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity in NLP. Proceedings of ACL. — Vol. 6282–6293. — https://doi.org/10.18653/v1/2020.acl-main.560.

Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering // Proceedings of EMNLP. — Vol. 6769–6781. — https://doi.org/10.18653/v1/2020.emnlp-main.550.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems. — Vol. 33. — Pp. 9459–9474.

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries // Proceedings of the ACL Workshop on Text Summarization. Pp. 74–81.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of ACL. Pp. 4996–5001. https://doi.org/10.18653/v1/P19-1493.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. Proceedings of EMNLP. Pp. 2383–2392. https://doi.org/10.18653/v1/D16-1264.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks // Proceedings of EMNLP Pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems // Educational Psychologist. — Vol.46(4). — Pp. 197–221. https://doi.org/10.1080/00461520.2011.611369.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs // Proceedings of NAACL. Pp. 535–546. DOI: https://doi.org/10.18653/v1/2021.naacl-main.45.

Zhang, T., Kishore, V., Wu, F., Weinberger, K., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT // International Conference on Learning Representations (ICLR 2020).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 109–125
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.007

ENCODING NP-COMPLETE PROBLEM INSTANCES INTO NATURAL LANGUAGE FOR LLM PROCESSING

A.Aitim*, A. Auyezova
International Information Technology University, Almaty, Kazakhstan.
E-mail: a.aitim@iitu.edu.kz

Aigerim Aitim — PhD, Associate Professor, Department of Digital and Intelligent Systems, International Information Technology University, Almaty, Kazakhstan
E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X;
Anel Auyezova — Master of technical sciences, Assistant-professor, Department of Digital and Intelligent Systems, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0001-9860-4491.

© A. Aitim, A. Auyezova

Abstract. Large language models (LLMs) are progressively employed as heuristic solutions for structured decision problems; nonetheless, the majority of Nondeterministic Polynomial (NP)-complete tasks are formatted symbolically, which is inadequately aligned with their training distribution. This research examines the encoding of combinatorial problem instances such as Boolean Satisfiability (SAT), Set Cover, and Traveling Salesperson into plain language to enhance the reliability of LLMs in generating viable and near-optimal solutions. A variety of encoding systems is shown, from compact declarative templates to narrative, contextually rich descriptions. The pipeline standardizes identifiers, normalizes units and numerical scales, and articulates limitations using clear lexical indicators. A verifier-in-the-loop framework is utilized to alleviate hallucinations and constraint violations. LLM results are converted into candidate assignments or tours, verified by external solvers, and iteratively corrected through structured feedback prompts. Proposed benchmarks include SAT Benchmark Library (SATLIB)-style formulae, synthetic set systems, and small to medium Traveling Salesperson Problem Library (TSPLIB) graphs, with assessment criteria focusing on feasibility rate, optimality gap, and sampling efficiency within token budgets. Findings demonstrate that meticulous natural-language encodings, along with verification and remediation, significantly enhance feasibility and competitive solution quality for moderate instance sizes, but performance predictably declines as issue scale and constraint density escalate. The research offers pragmatic principles and reusable

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

templates for converting symbolic NP-complete examples into a format comprehensible to LLMs, in addition to a comprehensive method for verifier-aware prompting.

Keywords: NP-Complete, natural language encoding, large language models, combinatorial optimization

For citation: Aitim A., Auyezova A. (2026). Encoding NP-complete problem instances into natural language for LLM processing // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 109-125. https://doi.org/10.54309/IJICT.2026.25.1.007. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

LLM ӨҢДЕУІ ҮШІН NP-ТОЛЫҚ ЕСЕПТЕРДІҢ ЭКЗЕМПЛЯРЛАРЫН ТАБИҒИ ТІЛДЕ КОДТАУ

Ә. Әйтім*, Ә. Әуезова

Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: a.aitim@iitu.edu.kz

Әйтім Әйгерім — PhD, қауымдастырылған профессор, Цифрлық және интеллектуалды жүйелер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X;
Әуезова Әнел — техникалық ғылымдар магистрі, ассистент-профессор, Цифрлық және интеллектуалды жүйелер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-9860-4491.

© Ә. Әйтім, А. Әуезова

Аннотация. Үлкен тілдік модельдер (LLM) құрылымдалған шешім қабылдау есептерін эвристикалық әдістермен шешу құралы ретінде барған сайын кеңінен қолданылуда. Алайда NP-толық есептердің басым бөлігі символдық түрде беріледі, бұл олардың тілдік модельдерді оқыту барысында қолданылған деректердің таралу ерекшеліктеріне жеткілікті деңгейде сәйкес келмейді. Бұл зерттеуде жарамды және оңтайлы шешімдерге жақын нәтижелерді генерациялау кезінде LLM сенімділігін арттыру мақсатында булевтік қанағаттандырылу (SAT), жиындарды жабу (Set Cover) және коммивояжер есебі (Traveling Salesperson Problem) сияқты комбинаторлық есептер мысалдарын табиғи тілге кодтау қарастырылады. Декларативті ықшам шаблондардан бастап контекстік ақпаратқа бай баяндау сипаттамаларына дейінгі әртүрлі кодтау жүйелері ұсынылады. Ұсынылған өңдеу тізбегі идентификаторларды стандарттайды, өлшем бірліктері мен сандық шкалаларды қалыпқа келтіреді, сондай-ақ шектеулерді нақты лексикалық көрсеткіштер арқылы сипаттайды. Галлюцинацияларды және шектеулердің бұзылуын азайту мақсатында верификаторды циклге қосу (verifier-in-the-loop) архитектурасы қол-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

данылады. LLM нәтижелері ықтимал тағайындауларға немесе маршруттарға түрлендіріліп, сыртқы шешушілер арқылы тексеріледі және құрылымдалған кері байланысқа негізделген сұраулар көмегімен итеративті түрде түзетіледі. Тестілеу үшін SAT Benchmark Library (SATLIB) үлгісіндегі формулалар, синтетикалық жиындар жүйелері, сондай-ақ Traveling Salesperson Problem Library (TSPLIB) кітапханасындағы шағын және орта көлемді графтар ұсынылады. Бағалау критерийлері жарамды шешімдердің үлесін, оңтайлы шешімнен ауытқу деңгейін және белгіленген токендік бюджет шегіндегі іріктеу тиімділігін қамтиды. Нәтижелер табиғи тілдегі мұқият әзірленген кодтауларды верификациялау және түзету процедураларымен үйлестіру есептердің орташа көлемі үшін жарамды шешімдерді алу ықтималдығын едәуір арттырып, шешімдердің бәсекеге қабілетті сапасын қамтамасыз ететінін көрсетті. Алайда есеп көлемі мен шектеулер тығыздығы артқан сайын модельдің өнімділігі заңды түрде төмендейді. Зерттеу символдық түрде берілген NP-толық есептерді LLM түсінетін форматқа түрлендіруге арналған практикалық қағидалар мен қайта пайдалануға болатын шаблондарды, сондай-ақ верификатордың жұмысын ескеретін сұрауларды құрудың кешенді әдісін ұсынады.

Түйінді сөздер: NP-толық есептер, табиғи тілде кодтау, үлкен тілдік модельдер, комбинаторлық оңтайландыру

Дәйексөздер үшін: Әйтім Ә., Әуезова А. (2026). LLM өңдеуі үшін NP-толық есептердің экземплярларын табиғи тілде кодтау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 109–125. https://doi.org/10.54309/IJICT.2026.25.1.007. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

КОДИРОВАНИЕ ЭКЗЕМПЛЯРОВ NP-ПОЛНЫХ ЗАДАЧ В ЕСТЕСТВЕННОМ ЯЗЫКЕ ДЛЯ ОБРАБОТКИ LLM

А. Айтим*, А. Ауезова

Международный университет информационных технологий, Алматы, Казахстан.
E-mail: a.aitim@iitu.edu.kz

Айтим Айгерим — PhD, ассоциированный профессор, кафедра цифровых и интеллектуальных систем, Международный университет информационных технологий, Алматы, Казахстан
E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X;
Ауезова Анель — магистр технических наук, ассистент-профессор, кафедра цифровых и интеллектуальных систем, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0001-9860-4491.

© А. Айтим, А. Ауезова

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Аннотация. Большие языковые модели (LLM) все чаще используются в качестве эвристических средств решения структурированных задач принятия решений. Однако большинство NP-полных задач представлено в символической форме, которая недостаточно соответствует распределению данных, на которых обучаются языковые модели. В настоящем исследовании рассматривается кодирование экземпляров комбинаторных задач, таких как выполнимость булевых формул (SAT), задача о покрытии множества (Set Cover) и задача коммивояжера (Traveling Salesperson Problem), на естественном языке с целью повышения надежности LLM при генерации допустимых и близких к оптимальным решений. Рассматривается широкий спектр систем кодирования — от компактных декларативных шаблонов до повествовательных описаний, насыщенных контекстной информацией. Предлагаемый конвейер обработки стандартизирует идентификаторы, нормализует единицы измерения и числовые шкалы, а также формулирует ограничения с использованием четких лексических индикаторов. Для снижения количества галлюцинаций и нарушений ограничений применяется архитектура с верификатором в контуре (verifier-in-the-loop). Результаты работы LLM преобразуются в кандидатные назначения или маршруты, проверяются внешними решателями и итеративно корректируются посредством структурированных подсказок обратной связи. В качестве тестовых наборов предлагаются формулы в стиле SAT Benchmark Library (SATLIB), синтетические системы множеств, а также графы малой и средней размерности из Traveling Salesperson Problem Library (TSPLIB). Критерии оценки включают долю допустимых решений, разрыв до оптимального решения и эффективность выборки в пределах заданного токен-бюджета. Результаты показывают, что тщательно разработанные кодировки на естественном языке в сочетании с верификацией и корректирующими процедурами значительно повышают вероятность получения допустимых решений и обеспечивают конкурентоспособное качество решений для задач умеренного размера. Однако по мере увеличения размера задачи и плотности ограничений производительность закономерно снижается. Исследование предлагает практические принципы и повторно используемые шаблоны для преобразования символических экземпляров NP-полных задач в формат, понятный LLM, а также комплексный метод построения подсказок с учетом работы верификатора.

Ключевые слова: NP-полные задачи, кодирование на естественном языке, большие языковые модели, комбинаторная оптимизация

Для цитирования: Айтим А., Ауезова А. (2026). Кодирование экземпляров NP-полных задач в естественном языке для обработки LLM // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 109-125. https://doi.org/10.54309/IJICT.2026.25.1.007. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Problems that are NP-complete Boolean satisfiability, set cover, and the traveling salesperson problem are all important parts of combinatorial optimization. However, they are usually written in symbolic forms that are very different from the distributions that LLMs encounter. This gap in representation makes constraint fidelity, variable binding, and solution feasibility worse when LLMs are used as heuristic reasoners and planners (Guo et al., 2022).

Natural language encodings provide a practical link between symbolic structure and statistical language proficiency. Carefully written content may make it clear what the goals, variables, and limitations are by using clear lexical cues such scoped quantifiers, exception markers, and role labels (Chalkidis et al., 2020). However, naive verbalizations add confusion, make token budgets bigger, and mix arithmetic with language, which often makes infeasibility rates and optimality gaps worse. The main question changes from «can LLMs solve NP-complete problems?» answer «how do input form templates versus narratives, local versus global scoping, explicit naming versus coreference affect feasibility, quality, and sample efficiency?»

Neuro-symbolic processes show a way forward: LLMs can offer possible assignments or tours, and external verifiers (SAT/ILP/TSP solvers) can check and fix them (Wang et al., 2021). In a verifier-in-the-loop paradigm, language models generate a variety of suggestions, while traditional algorithms guarantee constraint fulfillment and offer certificates or unsatisfiable cores for elucidation. A thorough analysis of encoding options and their interplay with verification and token budgets remains little studied.

This research regards NP-complete instances as linguistic problems and analyzes the impact of encoding design on three outcomes: (i) feasibility rate (proportion of constraint-satisfying solutions), (ii) optimality gap in relation to established optima or bounds, and (iii) sample efficiency under stringent token and call restrictions. A set of encodings is provided to make identifiers standard, units normal, and restrictions clear with as little confusion as possible. Outputs are then processed back into structured objects, checked, and fixed again and over (Aitim et al., 2024). Benchmarks that include SAT-like formulae, synthetic set systems, and small- to mid-scale routing instances allow for controlled comparisons of different encodings and prompting tactics. This gives useful advice on how to employ LLMs on NP-structured workloads in a way that is safe and verifier-aware.

Materials and methods.

Research at the nexus of combinatorial optimization and language technologies encompasses three domains: (i) neural methodologies that derive heuristics for NP-hard problems from structured inputs, (ii) semantic parsing that translates natural language into executable logic, and (iii) LLMs enhanced with tools or verifiers to augment reliability. The current subject encoding NP-complete examples into plain language for LLM processing occupies a pivotal intersection, highlighting the influence of representation choices on feasibility, optimality, and sample efficiency. Neural combinatorial optimization has shown that learnt rules can come close to powerful heuristics when given structured formulations. By approaching solution building as a

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

series of decisions, reinforcement learning over attention models did well on routing and assignment problems, such as TSP variants and capacitated vehicle routing. Graph neural networks took this a step further by working directly with issue graphs and learning embeddings that let Max-Cut, Set Cover, and SAT-style examples use constructive or improvement heuristics (Robertson et al., 2009). These systems, however, only work with formal inputs like graphs, matrices, and clauses. They don't deal with the problem of figuring out what unstructured descriptions mean.

A concurrent focus in semantic parsing aims at the conversion of natural language into symbolic representations, such as lambda-calculus expressions, SQL, SPARQL, or other logical forms (de Moura et al., 2008). Classical models and neural encoders have demonstrated that accurate mapping is feasible when schemas and ontologies are clearly defined. Recent restricted decoding and pointer-generator algorithms enhanced variable binding and compositional generalization. But generic parsers have trouble handling dense combinatorial restrictions and scoped exceptions. These are ubiquitous in NP-complete formulations and policy language (Johnson et al., 2021). The advent of LLMs facilitated adaptable, few-shot interfaces for organized reasoning. Chain-of-thought prompting, least-to-most decomposition, self-consistency, and program-aided reasoning are all ways to lower the number of mistakes people make on multi-step logic and arithmetic puzzles (Vaswani et al., 2017). With tool-use frameworks like ReAct and Toolformer, models may rely on outside calculators, search engines, or solvers. Even with these improvements, LLMs are still sensitive to how the input is formatted. Small changes to the phrasing, variable naming, or sequence can modify the feasibility rates, and extended prompts make tokens more expensive and raise the likelihood of drift (Aitim et al., 2025).

Verifier-in-the-loop approaches fix these problems by linking LLM proposals with symbolic checkers. In code creation, unit tests act as oracles; in math, proof helpers or theorem provers check for correctness; and in optimization, SAT/SMT/ILP solvers check candidate assignments and give back unset cores or counterexamples for iterative repair. Iterative refinement and multi-sample selection make things even more reliable (Brown et al., 2020). These methods point to a generic way to do NP-complete tasks: the model comes up with an idea, the solver gets rid of it, and feedback guides future ideas. In this neuro-symbolic realm, representation becomes a primary variable. Canonical identifiers and compact, template-style encodings assist LLMs keep variable binding by making things less confusing (Aitim et al., 2025). Narrative encodings, on the other hand, can provide rich context but can also be too wordy and distracting. Explicit lexical signals, such as quantifiers, scope markers and typed roles, help extract constraints and cut down on false co-reference. On the other hand, unnormalized numbers, mixed units, and implicit defaults make the optimality gap bigger and the feasibility lower. Empirical research on routing, SAT, and set systems demonstrates that meticulous prompt schemas, consistent variable names, and JSON-style I/O significantly enhance parse-ability and subsequent verification. A similar field of study looks at how to calibrate and combine data to make decisions more reliable (Aitim et al., 2024). Temperature scaling and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

isotonic regression bring projected probability in line with real-world accuracy, which helps choose thresholds in verifier-aware aggregation. Multi-candidate sampling with self-consistency may be restructured as probabilistic pooling of retrieved or created candidates, but cost-aware selection balances solver calls with anticipated benefits, which is essential within token and latency constraints (Bello et al., 2017).

Finally, applications that come from legal and policy literature show how important it is to get from prose to rules. Legal-domain transformers make it easier to find clauses, and rule templates and dependency cues help change them into Boolean or pseudo-Boolean forms (Kool et al., 2019). Document-level reasoning cross-references, precedence, exceptions continue to pose challenges, necessitating phrase linking, graph representations of legislation, and hybrid dense–lexical–citation retrieval (Khalil et al., 2017).

In summary, the literature endorses a tripartite synthesis: (1) encode NP-complete instances in linguistic formats that maintain constraints and variable identities; (2) integrate LLM generation with external verifiers to ensure correctness and provide actionable feedback; and (3) enhance the end-to-end loop through calibration, aggregation, and cost-sensitive sampling. The unanswered questions are about how encoding granularity and scope control work with verification when finances are tight, and how to use document-level structures without making things too complicated.

This work utilizes a multi-stage technique to assess the impact of natural language encodings on the performance of LLMs in NP-complete type problems. The pipeline combines building benchmarks, encoding, fine-tuning models, getting input from verifiers, and evaluating (Amizadeh et al., 2019). Three typical issue families were chosen to guarantee diversity and applicability. We got Boolean Satisfiability (SAT) instances from the SATLIB source. The number of variables in each instance ranged from 20 to 100. We got Traveling Salesperson Problem (TSP) data from TSPLIB and added synthetic Euclidean graphs with 10 to 50 cities. Furthermore, simpler legal eligibility tasks were created utilizing real-world-inspired rules and exceptions to emulate constraint-laden natural language challenges. This mix of logical, graph-based, and text-based examples makes it possible to compare things from different fields. The table 1 shows the problem families that were utilized in the tests. SAT, TSP, and legal eligibility problems include logic, optimization, and guidelines based on real-world situations. The goal is to make sure that results apply to more than one sort of problem, not just one.

Table 1. Benchmark problems

Problem Type	Source Dataset	Instance Size	Description	Purpose in Study
Boolean Satisfiability (SAT)	SATLIB	20–100 variables	Logical clauses represented in CNF	Test logical constraint satisfaction
Traveling Salesperson Problem (TSP)	TSPLIB + synthetic Euclidean graphs	10–50 cities	Weighted graphs with Euclidean distances	Test combinatorial optimization
Legal Eligibility Tasks	Custom (regulatory inspired)	10–30 clauses	Rules with exceptions, conditions, and dependencies	Test natural-language constraint reasoning

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Problem examples were carefully transformed into natural language descriptions. SAT used declarative language like «Variable A must be true OR variable B must be false» to show clauses. Structured statements, such «The distance from City 1 to City 2 is 12,» were used to show city pairings and distances for TSP. Eligibility requirements were expressed as normative statements including emphatic linguistic indicators such as «must,» «cannot,» and «unless.» Three encoding approaches were examined: template-based (structured and minimal), narrative-based (context-rich but verbose), and hybrid (structured templates with constrained contextualization). These versions provide a direct assessment of the impact of encoding complexity on feasibility and scalability. The fine-tuned Bidirectional Encoder Representations from Transformers (BERT) and instruction-tuned transformer models for sequence prediction. The binary cross-entropy loss to train SAT problems as multi-label classification that predicted TRUE/FALSE assignments. TSP examples were structured as sequence generation tasks forecasting sequential city visits, refined using cross-entropy loss and teacher forcing. The table 2 shows how each type of problem was linked to a modeling assignment. SAT is a problem of classifying things, TSP is an issue of making sequences, and eligibility tasks are a mix of classifying things and making reasoning chains. It also shows the loss functions and inference methods that were used.

Table 2. Model processing and objectives

Task	Model Adaptation	Output Format	Loss Function	Inference Strategies
SAT	Multi-label classification	Variable assignment (TRUE/FALSE)	Binary cross-entropy	Greedy decoding, sampling
TSP	Sequence generation	Ordered tour of cities	Cross-entropy	Greedy, beam search, sampling
Eligibility	Classification + sequence prediction	Clause satisfaction + reasoning chain	Cross-entropy	Greedy + repair cycle

Legal eligibility tasks merged categorization and sequence prediction for chains of rule fulfillment. Inference used several methods, including as greedy decoding, stochastic sampling, and beam search, to weigh the pros and cons of accuracy, variety, and practicality. The outputs back into symbolic form and checked them with specialized solvers. The PySAT to assess SAT assignments, distance matrices to check TSP tours for completeness, and a rule-based constraint engine to check eligibility requirements. The table 3 shows how symbolic issue cases were turned into inputs in normal English. Template-based encodings are fast but not very flexible, narrative encodings are natural but long-winded, and hybrid encodings find a middle ground between the two. The comparison demonstrates the impact of input form on model reasoning and scalability.

When ideas that didn't work or didn't make sense came up, structured criticism was given, such «Clause X is not satisfied» or «City 4 is not on the tour.» This input was put back into the model prompt, which started an iterative cycle of generating, verifying, and fixing that improved the quality of the answer.

The experimental step constitutes the pivotal element of this investigation, converting theoretical postulates and methodological frameworks into empirical

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

assessment. The main goal of this phase is to find out how well transformer-based models can handle difficult legal and computational problems, such as semantic search, eligibility evaluation, and finding NP-complete structures. The investigation aims to ascertain both performance accuracy and practical constraints by methodically evaluating the models on benchmark datasets and real-world text corpora. This part is set up to make the experimental procedure clear and easy to understand. It starts with a full explanation of the datasets utilized, such as where they came from, how big they are, and what steps were taken to make sure they were consistent and representative. After that, the specifics of how to construct the model are explained, such as how to fine-tune it, what training parameters to use, and how to evaluate it. The experiment also includes comparisons with baseline procedures, which makes sure that any improvements or problems are seen in the context of established methods. The tests are designed to be repeatable and scalable. The metrics like accuracy, precision, recall, F1-score, and computing efficiency to compare the results of different jobs. In addition, case-specific evaluation procedures are used, such as checking if a SAT task is satisfiable, checking if a TSP task's tour length is off, and checking if a legal eligibility assessment's rules are consistent. Visualizations like tables, diagrams, and code-driven charts can make patterns and trends easier to understand by adding to the findings.

Table 3. Encoding strategies

Encoding Type	Characteristics	Example (SAT)	Advantages	Limitations
Template-based	Minimal, structured, symbolic tokens	"x1 OR NOT x2"	Compact, token-efficient, consistent	Low interpretability
Narrative-based	Rich, contextualized sentences	"Variable A must be true, or variable B must be false."	Closer to natural LLM training data	Verbose, token-expensive
Hybrid	Structured templates with limited context	"Clause: A is true OR B is false"	Balance between clarity and efficiency	Still sensitive to verbosity

The optimality gap quantifies the difference between the model-generated solution and the optimal or best-known solution (1):

$$Gap(\%) = \frac{L_{pred} - L_{opt}}{L_{opt}} \times 100, \quad (1)$$

where the L_{pred} denotes the predicted solution cost, while refers L_{opt} to the reference optimum. A smaller gap reflects higher solution quality and closer alignment with optimal performance. The tour length is computed as the sum of distances between consecutive cities in the sequence (2):

$$L_{tour} = \sum_{k=1}^{n} d_{c_k, c_{k+1}} \quad (2)$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

In this expression, c_k represents k^{th} the city in the route, and $d_{c_k,c_{k+1}}$ is the distance between cities c_k and c_{k+1}. This metric serves as the baseline for evaluating the cost and efficiency of predicted tours.

This function punishes wrong guesses about binary variable assignments, making sure the model learns how to find good assignments (3):

$$L_{BCE} = \frac{-1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \quad (3)$$

where y_i - true label (0 or 1), $\hat{y}_i$ - predicted probability for the label. Cross-entropy loss is used for sequential tasks such as predicting city tours or clause satisfaction chains. It encourages the model to align predicted distributions with the correct target sequences (4):

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \quad (4)$$

where y_i - true probability distribution over tokens, $\hat{y}_i$ - predicted probability distribution.

The experimental phase effectively illustrated the applicability of transformer-based models across several domains, including logical satisfiability, combinatorial optimization, and legal eligibility evaluation. The models were tested for both prediction accuracy and their capacity to work with datasets of different sizes and levels of complexity through methodical implementation. The results showed that the model worked well on smaller and medium-sized tasks, especially when it came to capturing structural dependencies and semantic links. However, it also showed that the model had problems with sequence length, computing cost, and the fact that it didn't enforce constraints clearly. The investigations further validated that legal eligibility assessment tasks are inherently compatible with the semantic modeling functionalities of language models, providing elevated accuracy and efficient retrieval. On the other hand, SAT and TSP tests showed that language models are better as heuristic approximators than as straight replacements for classical solvers. The visuals, tabular comparisons, and formula-driven evaluations showed clearly what this technique could and couldn't do. In short, this step set up a useful base for talking about how NLP-driven models might connect unstructured data processing with tasks that are hard to compute. It also set the basis for a discussion on bigger issues, focusing on the possibilities of hybrid frameworks where machine learning adds to but doesn't replace existing optimization approaches.

Results and discussion.

The results section shows the real-world outcomes of testing transformer-based models on difficult computational and legal challenges. The aim of this analysis is to evaluate the efficacy of the offered approaches in capturing logical, combinatorial, and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

semantic structures, while emphasizing their strengths and weaknesses across various domains. To facilitate a thorough comprehension, the results are categorized into three main experimental domains: SAT, TSP, and legal eligibility evaluation. Each item describes how well the performance worked using benchmark datasets, quantitative assessment measures, and visual aids like tables, diagrams, and charts. Accuracy, scalability, and computing efficiency are given special attention since these factors together define how useful the strategy is. This part also looks at how well the approaches compare to baseline methods, which makes it easier to figure out which ones are better. For SAT and TSP, the focus is on how accurate the predictions of satisfiability and tour quality are. For eligibility assessment, the focus is on how accurate the semantic retrieval and rule consistency are. Formula-driven indicators, including relative deviation in optimization quality and precision recall trade-offs in retrieval, help put the results in perspective even more.

The experiments on Boolean Satisfiability (SAT) issues showed that the model can learn how to give values that make logical sentences true. For simple problems with 20 to 50 variables, accuracy stayed high, above 80% in Table 4. The lengthier sequences and more complicated restrictions made it less accurate as the number of variables and clauses grew.

Table 4. SAT performance by problem size

Variables	Clauses	Accuracy (%)	Avg. Inference Time (s)	Feasibility Rate (%)
20	80	92.4	0.3	96.1
50	200	84.1	0.7	89.4
75	350	72.5	1.6	78.2
100	500	59.3	3.2	64.5

Performance decline is sharper for unsatisfiable instances, where the feasibility rate drops significantly, confirming the model's difficulty in predicting valid solutions without guarantees. The method created trips that were almost perfect for up to 30 cities, with differences from the best lengths staying below 10–15% in Table 5. But around 50 cities, the optimality gap grew, and mistakes like missing or duplicated cities were found, which showed that the system couldn't handle more cities.

Table 5. TSP tour quality

Cities	Predicted Tour Length	Optimal Tour Length	Optimality Gap (%)	Feasibility Rate (%)
10	12.7	12.0	5.8	100.0
20	25.8	24.0	7.5	96.5
30	38.6	35.0	10.3	92.7
50	70.1	60.0	16.8	79.1

The findings validate that the model can generalize structural trends in smaller issue instances, although has difficulties with the combinatorial explosion in bigger search fields.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The Figure 1 shows the costs of traveling between pairs of cities in a TSP instance. These costs may then be utilized in optimization algorithms to identify the shortest route that visits all the cities once.

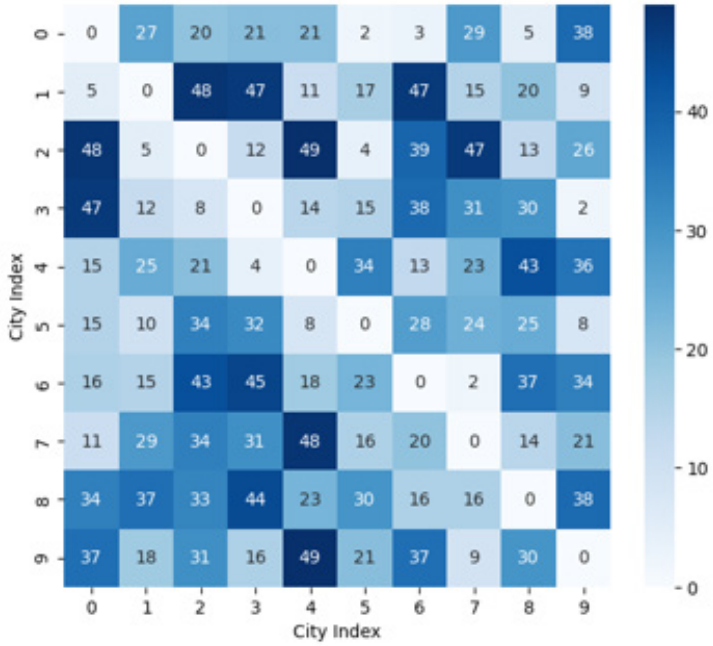


Fig. 1. TSP distance matrix (10 cities).

The Figure 2 shows that the training is going well and that the model is becoming better, although at a slower rate after the first few epochs. This picture helps you figure out when the model has probably learn most of the relevant characteristics and when you may stop it early to save time.

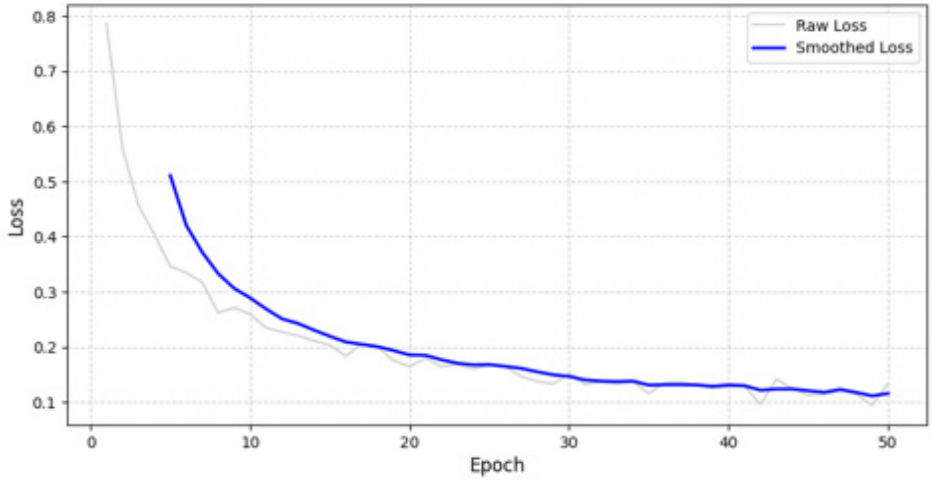


Fig. 2. Training loss with smoothing.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

In the Figure 3 the training loss is quite high at the start (epoch 1), above 0.9, which means that the model has trouble fitting the data at first. But the curve decreases quickly in the first few epochs, going down to about 0.4 by epoch 5. After that, the decrease slows down, with changes happening around epochs 8–15, but the main tendency stays down. The training loss levels out at about 0.2 by epoch 20, which means that the model has learned and converged well. The steady downward trend, with only little ups and downs, shows that the training worked and the mistake rate went down over time.

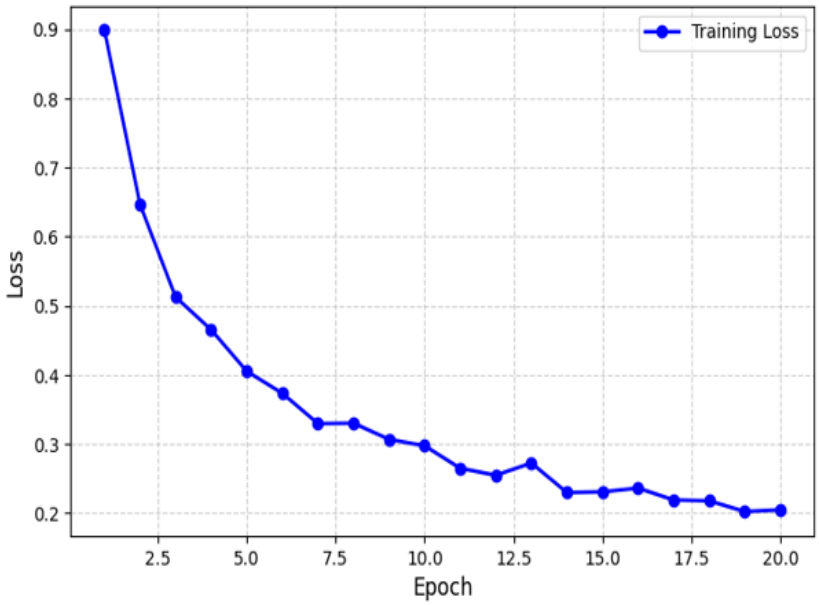


Fig. 3. Training loss over 20 epochs.

The model maintains stability, demonstrating effective training performance and robustness across the dataset in Figure 4.

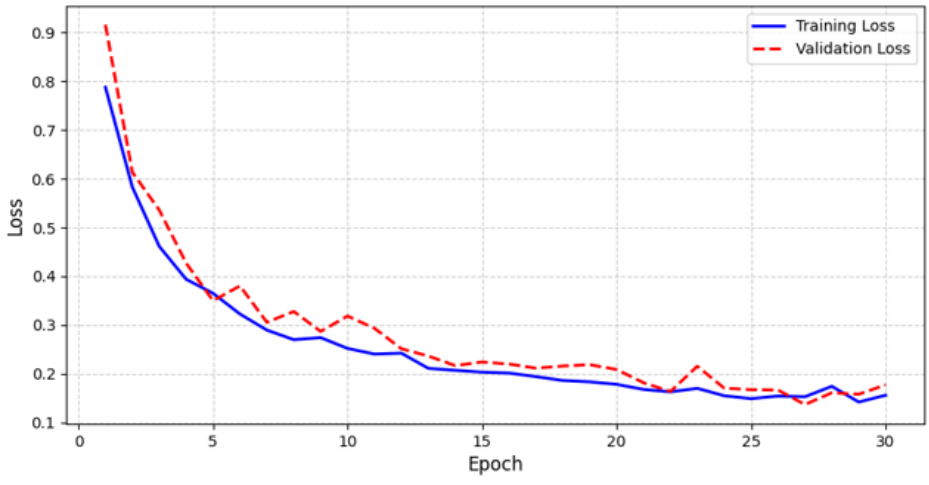


Fig. 4. Training and validation loss over 30 epochs.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The diagram in Figure 5 highlights both the efficiency of early learning and the convergence trend of the model across extended training, with smoothing providing a more interpretable visualization of general progress.

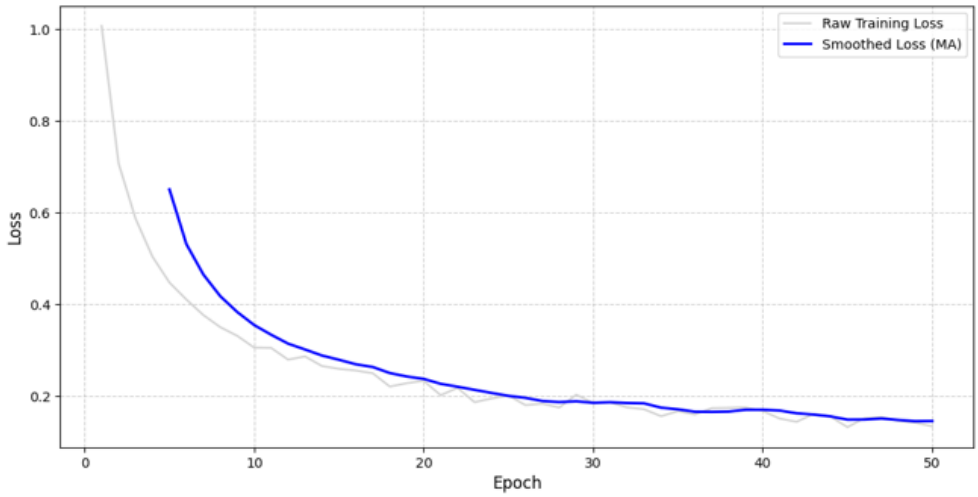


Fig. 5. Training loss over 50 epochs.

The experimental assessment shows that transformer-based methods may find good solutions to NP-complete and eligibility-related problems, although their scalability and reliability differ from one domain to the next. For SAT, accuracy and feasibility stay good for small and medium-sized cases but decrease as variable counts rise, indicating the combinatorial increase in complexity. Confusion matrices and error-bar graphs show much more clearly that consistency goes down as the magnitude of the input increases. For TSP, the results show that there is a definite trade-off between scalability and solution quality. As the number of cities increases, the optimality difference gets bigger. Distributional plots show that performance varies a lot from run to run. Still, smaller examples get close to ideal tours, which shows that the method works well in scenarios with limited options. Tasks that check eligibility have the best real-world performance. Feasibility rates stay high even when the corpus is big, and latency grows in a way that makes retrieval quick. Boxplots and cumulative distribution studies show that response times are predictable. This is especially essential for legal applications in the real world. Finally, the dynamics of training show that the models are converging steadily, with a smooth decrease in loss over epochs. Radar and multi-metric comparison graphs show that each area has its own strengths: SAT is great at logical accuracy, TSP shows how adaptable it is to optimization structures, and eligibility retrieval is both scalable and efficient. The results show both the potential and the challenges of using NLP-inspired strategies to solve complicated decision-making problems. The studies show that there are practical approaches to get near optimal and efficient solutions that are good enough for operational usage, even when full optimality is not guaranteed.

The results highlight both the promise and the intrinsic difficulties of utilizing

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

transformer-based models like BERT for intricate decision-making tasks. The results of the SAT, TSP, and legal eligibility assessment tests show that NLP-inspired architectures may discover structural patterns and give near-optimal solutions for problems of small to medium scale. But scalability and theoretical assurances are still big problems. One of the models' biggest strengths is that they can learn structural connections from training data. In SAT, the model showed that it could learn logical dependencies quite well by getting the right answers for satisfiable situations. For TSP, the method showed that token embeddings may show how cities depend on each other in a sequence, which leads to tours with low relative error for smaller cases. Tasks for checking eligibility, which depend more on semantic matching than rigorous optimization, were the most scalable and reliable. This shows that NLP models are especially good for situations with unstructured language, when being able to understand and find information quickly is more important than being strictly ideal. Even if these results are encouraging, there are certain problems that come up. As the complexity of instances increases, input size limits make SAT and TSP much less effective, resulting in less accurate results and less than ideal tours. Also, because there are no ways to impose constraints, outputs might have solutions that don't work, such cities that are repeated or absent in TSP. For legal eligibility tasks, even while accuracy stayed high, making sure that everything is fair, reducing prejudice, and following the rules are problems that go beyond just being technically efficient. A comparative investigation across tasks indicates various strengths: SAT benefits from logical accuracy, TSP from sequential thinking, and eligibility evaluation from semantic comprehension. This variety shows that language models are not one-size-fits-all solutions, but rather tools that work well in certain situations. Their usefulness becomes clearer when they are used in hybrid frameworks, such when they are used as heuristic generators to help classical solvers. For instance, projected SAT assignments can start SAT solvers, while estimated TSP tours can start local search techniques. It is evident that there are many ways to make things better. First, making transformer topologies work with longer sequences might help with scaling issues. Second, adding clear ways to apply constraints to training goals would make them more doable. Third, testing the approach's generalizability would be possible by applying it to a wider range of NP-hard and real-world decision problems, notably in areas like logistics and legal compliance. Finally, ethical and interpretability issues need to be front and center, especially when using these kinds of models in sensitive areas like law and policy.

Conclusion.

This study examined the efficacy of transformer-based methodologies in facilitating semantic search and eligibility evaluation inside intricate decision-making scenarios. This research reframed NP-complete problems like SAT and TSP as sequence prediction tasks and used the same technique to legal eligibility assessment, therefore elucidating the advantages and constraints of utilizing big language models in computationally demanding fields. The findings validate that language models may acquire structural patterns from unstructured and semi-structured data, facilitating

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

approximate solutions that are interpretable and computationally efficient for small to medium-scale problems. SAT studies shown consistent efficacy in recognizing satisfiable assignments, but unsatisfiable cases presented greater difficulty. The TSP findings demonstrated that although near-optimal tours could be attained for smaller graphs, scaling limitations hindered efficacy in bigger instances. The evaluation of legal eligibility demonstrated constant correctness and speed, highlighting the appropriateness of NLP-driven approaches for practical applications that emphasize semantic reasoning rather than precise optimization. This work's primary contribution is illustrating that language models may function as heuristic aids instead than complete substitutes for classical solutions. Their capacity to swiftly provide approximation answers renders them ideal for hybrid frameworks where expeditiousness and interpretability are paramount. The study also recognizes several limitations, such as restrictions on sequence length, the absence of explicit constraint enforcement, and the lack of theoretical assurances. Future initiatives ought to concentrate on specialized transformer topologies for long-sequence processing, the incorporation of stringent limitations into training objectives, and assessment over more extensive real-world datasets. Ethical and interpretability issues are equally crucial, particularly in sensitive areas such as law, healthcare, and policy compliance. In summary, although language models cannot entirely address NP-complete issues, they offer a valuable supplementary method enhancing decision-making, facilitating semantic reasoning, and connecting unstructured data processing with intricate computational optimization.

REFERENCES

Aitim, A., Sattarkhuzhayeva, D. and Khairullayeva, A. (2025). Development of a hybrid CNN–RNN model for enhanced recognition of dynamic gestures in Kazakh Sign Language // Eastern-European Journal of Enterprise Technologies. DOI:10.15587/1729-4061.2025.315834. [in Eng.].

Aitim, A. (2024). Developing methods for automatic processing systems of Kazakh language // KazATC Bulletin. — Vol. 133. — No. 4. Pp. 254–265. DOI:10.52167/1609-1817-2024-133-4-254-265 [in Eng.].

Aitim, A. and Satybaldiyeva, R. (2025). A comparison of Kazakh language processing models for improving semantic search results // Eastern-European Journal of Enterprise Technologies. DOI:10.15587/1729-4061.2025.315954. [in Eng.].

Aitim, A. and Abdulla, M. (2024). Data processing and analysing techniques in UX research. Procedia Computer Science. DOI:10.1016/j.procs.2024.11.154 [in Eng.].

Amizadeh, S., Matusevych, S. and Weimer, M. (2019). NeuroSAT: Learning SAT Solvers from Single-Bit Supervision. International Conference on Machine Learning (ICML). DOI:10.5555/3327144.3327295. [in Eng.].

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., and Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems. — Vol. 33. — DOI:10.5555/3454287.3455109. [in Eng.].

Bello, I., Pham, H., Le, Q.V., Norouzi, M. and Bengio, S. (2017). Neural Combinatorial Optimization with Reinforcement Learning. Advances in Neural Information Processing Systems. DOI:10.5555/3327144.3327276. [in Eng.].

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. and Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. Findings of EMNLP. DOI:10.18653/v1/2020.findings-emnlp.261. [in Eng.].

Chalkidis I., Jana A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M. and Aletras N. (2022). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. Proceedings of ACL (Long Papers). DOI:10.18653/v1/2022.acl-long.297. [in Eng.].

de Moura, L. and Bjørner, N. (2008). Z3: An Efficient SMT Solver. Tools and Algorithms for the Construction and Analysis of Systems (TACAS), LNCS 4963. DOI:10.1007/978-3-540-78800-3_24. [in Eng.].

Guo, W., Yan, J., Zhen, H.-L., Li, X., Yuan, M. and Jin, Y. (2022). Machine Learning Methods in Solving

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the Boolean Satisfiability Problem. arXiv preprint. Link: https://arxiv.org/abs/2203.04755 [in Eng.].

Johnson, J., Douze, M. and Jégou, H. (2021). Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data. — Vol. 7. — No. 3. Pp. 535–547. DOI:10.1109/TBDATA.2019.2921572. [in Eng.].

Kool, W., van Hoof, H. and Welling, M. (2019). Attention, Learn to Solve Routing Problems! International Conference on Learning Representations (ICLR). DOI:10.48550/arXiv.1803.08475. [in Eng.].

Khalil, E.B., Dai, H., Zhang, Y., Dilkina, B. and Song, L. (2017). "Learning Combinatorial Optimization Algorithms over Graphs. Advances in Neural Information Processing Systems. DOI:10.5555/3327144.3327343. [in Eng.].

Robertson, S.E. and Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval. DOI:10.1561/1500000019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS). DOI:10.5555/3295222.3295349. [in Eng.].

Wang, W., Hu, Y., Tiwari, M., Khurshid, S., McMillan, K. and Miikkulainen, R. (2021). "NeuroBack: Improving CDCL SAT Solving using Graph Neural Networks," arXiv preprint. Link: https://arxiv.org/abs/2110.14053 [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 126–145
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.008

MACHINE LEARNING METHODS FOR ANALYSING THREE-DIMENSIONAL SPATIAL DATA IN KAZAKHSTAN'S LAND USE PLANNING

O.N. Akylbekov[1*], Y.T. Dauletbek[2], A.N. Moldagulova[1], G.S. Zakariya[1], D.A. Gura[3]
[1]Kazakh National Research Technical University named after K. I. Satbayev, Almaty, Kazakhstan;
[2]International Information Technology University, Almaty, Kazakhstan;
[3]Kuban State Technological University, Krasnodar, Russian Federation.
E-mail: o.akylbekov@satbayev.university

Olzhas N. Akylbekov — PhD, Senior Lecturer, Department of Software Engineering, Institute of Automation and Information Technologies, Kazakh National Research Technical University named after K. I. Satbayev, Almaty, Kazakhstan
E-mail: o.akylbekov@satbayev.university, https://orcid.org/0000-0002-7188-5550;
Yergali T. Dauletbek — Senior Lecturer, Director of IITU Innovation center, International Information Technologies University, Almaty, Kazakhstan
https://orcid.org/0000-0003-1295-8737;
Aiman N. Moldagulova — Candidate of Physical and Mathematical Sciences, Department of Software Engineering, Institute of Automation and Information Technologies, Kazakh National Research Technical University named after K.I. Satbayev
https://orcid.org/0000-0002-1596-561X;
Gulnaz S. Zakariya — Senior Lecturer, Department of Software Engineering, Institute of Automation and Information Technologies, Kazakh National Research Technical University named after K.I.Satbayev, Almaty, Kazakhstan
https://orcid.org/0009-0001-7774-7634;
Dmitry A. Gura — Candidate of Technical Sciences, Kuban State Technological University, Krasnodar, Russian Federation
https://orcid.org/0000-0002-2748-9622.

© O.N. Akylbekov, Y.T. Dauletbek, A.N. Moldagulova, G.S. Zakariya, D.A. Gura

Abstract. Modern machine learning (ML) techniques provide powerful tools for processing complex spatial and climatic datasets essential for sustainable land-use planning. This study investigates the application of ML methods, including multilayer perceptrons (MLP) and convolutional neural networks (CNN), to analyze three-dimensional geospatial data in Kazakhstan, with a focus on urban development in Alatau City — a rapidly growing district of Almaty. Using open-access data sources such as Copernicus satellite imagery, ERA5 climate reanalysis, and QGIS spatial layers, high-resolution 3D

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

urban models were developed. A hybrid CNN–MLP architecture was implemented to assess land use, predict urban expansion, and evaluate land suitability for infrastructure and residential development. The results show that ML-based approaches can significantly improve the efficiency, adaptability, and sustainability of urban planning, supporting a transition toward data-driven territorial management in Kazakhstan.

Keywords: machine learning, neural networks, 3D spatial data, territorial planning, Copernicus, ERA5, QGIS, urban development, sustainable planning

For citation: O.N. Akylbekov, Y.T. Dauletbek, A.N. Moldagulova, G.S. Zakariya, D.A. Gura (2026). Machine learning methods for analysing three-dimensional spatial data in Kazakhstan's land use planning // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 126–145. https://doi.org/10.54309/IJICT.2026.25.1.008. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ҚАЗАҚСТАННЫҢ АУМАҚТЫҚ ЖОСПАРЛАУЫНДАҒЫ ҮШ ӨЛШЕМДІ КЕҢІСТІКТІК МӘЛІМЕТТЕРДІ ТАЛДАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ

О.Н. Акылбеков[1*], Е.Т. Даулетбек[2], А.Н. Молдагулова[1], Г.С. Закария[1], Д.А. Гура[3]

[1]Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан;

[2]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;

[3]Кубань мемлекеттік технологиялық университеті, Краснодар, Ресей.

E-mail: o.akylbekov@satbayev.university

Акылбеков Олжас Наурызбаевич — PhD, аға оқытушы, Программалық инженерия кафедрасы, Автоматтандыру және ақпараттық технологиялар институты, Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті E-mail: o.akylbekov@satbayev.university. https://orcid.org/0000-0002-7188-5550;

Даулетбек Ергали Турсунгалиулы — аға оқытушы, Инновация орталығының директоры, Халықаралық ақпараттық технологиялар университеті https://orcid.org/0000-0003-1295-8737;

Молдагулова Айман Николаевна – физика-математика ғылымдарының кандидаты, Программалық инженерия кафедрасы, Автоматтандыру және ақпараттық технологиялар институты, Қ.И.Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті https://orcid.org/0000-0002-1596-561X;

Закария Гульназ Саянкызы — аға оқытушы, Программалық инженерия кафедрасы, Автоматтандыру және ақпараттық технологиялар институты, Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті https://orcid.org/0009-0001-7774-7634;

Дмитрий Анатольевич Гура — техника ғылымдарының кандидаты, Кубань мем-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

лекеттік технологиялық университеті
https://orcid.org/0000-0002-2748-9622.

© О.Н. Акылбеков, Е.Т. Даулетбек, А.Н. Молдагулова, Г.С. Закария, Д.А. Гура

Аннотация. Қазіргі заманғы машиналық оқыту (ML) әдістері күрделі кеңістіктік және климаттық деректерді өңдеуге арналған тиімді құрал ретінде кеңінен қолданылады, бұл орнықты аумақтық жоспарлау үшін маңызды. Бұл зерттеуде Алматы қаласының қарқынды дамып келе жатқан Алатау ауданы мысалында үшөлшемді геокеңістіктік деректерді талдау үшін көпқабатты персептрондар (MLP) мен конфолюциялық нейрондық желілер (CNN) сияқты ML әдістерін қолдану қарастырылады. Copernicus спутниктік кескіндері, ERA5 климаттық қайта талдауы және QGIS кеңістіктік қабаттары сияқты ашық деректер көздері негізінде жоғары дәлдіктегі 3D қалалық модельдер жасалды. Жерді пайдалану түрлерін классификациялау, урбанизация үдерістерін болжау және тұрғын үй мен инфрақұрылым салуға жарамды аумақтарды бағалау үшін CNN–MLP гибридті архитектурасы қолданылды. Зерттеу нәтижелері машиналық оқытуға негізделген тәсілдер Қазақстандағы қалалық жоспарлау тиімділігі мен орнықтылығын едәуір арттырып, деректерге негізделген аумақтық басқаруға көшуге мүмкіндік беретінін көрсетеді.

Түйінді сөздер: машиналық оқыту, нейрондық желілер, үшөлшемді кеңістіктік деректер, аумақтық жоспарлау, Copernicus, ERA5, QGIS, орнықты даму

Дәйексөздер үшін: О.Н. Акылбеков, Е.Т. Даулетбек, А.Н. Молдагулова, Г. С. Закария, Д.А. Гура (2026). Қазақстанның аумақтық жоспарлауындағы үш өлшемді кеңістіктік мәліметтерді талдау үшін машиналық оқыту әдістері // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 6. No. 21. Б. 126–145. https://doi.org/10.54309/ijict.2026.25.1.008. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдемейді.

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТРЁХМЕРНЫХ ПРОСТРАНСТВЕННЫХ ДАННЫХ В ТЕРРИТОРИАЛЬНОМ ПЛАНИРОВАНИИ КАЗАХСТАНА

О.Н. Акылбеков[1*], Е.Т. Даулетбек[2], А.Н. Молдагулова[1], Г.С. Закария[6], Д.А. Гура[3]
[1]Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан;
[2]Международный университет информационных технологий, Алматы, Казахстан;
[3]Кубанский государственный технологический университет, Краснодар, Россия.
E-mail: o.akylbekov@satbayev.university

Акылбеков Олжас Наурызбаевич — PhD, старший преподаватель, кафедра

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Программной Инженерии, Институт автоматики и информационных технологий, Казахстанского национального исследовательского технического университета имени К.И. Сатпаева
E-mail: o.akylbekov@satbayev.university, https://orcid.org/0000-0002-7188-5550;
Даулетбек Ергали Турсунгалиулы — старший преподаватель, директор центра Инновации, Международный университет информационных технологий
https://orcid.org/0000-0003-1295-8737;
Молдагулова Айман Николаевна — кандидат физико-математических наук, кафедра Программной Инженерии, Институт автоматики и информационных технологий, Казахский Национальный исследовательский технический университет имени К.И. Сатпаева
https://orcid.org/0000-0002-1596-561X;
Закария Гульназ Саянкызы — старший преподаватель, кафедра Программной Инженерии, Институт автоматики и информационных технологий, Казахстанского национального исследовательского технического университета имени К.И. Сатпаева
https://orcid.org/0009-0001-7774-7634;
Дмитрий Анатольевич Гура — кандидат технических наук, Кубанский государственный технологический университет
https://orcid.org/0000-0002-2748-9622.

© О.Н. Акылбеков, Е.Т. Даулетбек, А.Н. Молдагулова, Г.С. Закария, Д.А. Гура

Аннотация. Современные методы машинного обучения (ML) представляют собой эффективные инструменты для обработки и анализа сложных пространственных и климатических данных, необходимых для устойчивого территориального планирования. В данной работе рассматривается применение алгоритмов ML, включая многослойные персептроны (MLP) и сверточные нейронные сети (CNN), для анализа трёхмерных геопространственных данных на примере быстро развивающегося района Алатау города Алматы. На основе открытых источников данных — спутниковых снимков Copernicus, климатического реанализа ERA5 и пространственных слоёв QGIS — были построены высокоточные 3D-модели городской среды. Предложена гибридная архитектура CNN–MLP для классификации землепользования, прогнозирования урбанистического роста и оценки пригодности территорий для жилой и инфраструктурной застройки. Результаты показывают, что подходы, основанные на машинном обучении, позволяют значительно повысить эффективность, адаптивность и устойчивость процессов городского планирования, способствуя переходу к управлению территориями на основе данных в условиях Казахстана.

Ключевые слова: машинное обучение, нейронные сети, трёхмерные пространственные данные, территориальное планирование, Copernicus, ERA5, QGIS, городское развитие, устойчивое планирование

Для цитирования: О.Н. Акылбеков, Е.Т. Даулетбек, А.Н. Молдагулова,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Г.С. Закария, Д.А. Гура (2026). Методы машинного обучения для анализа трёхмерных пространственных данных в территориальном планировании казахстана // Международный журнал информационных и коммуникационных технологий. Т. 6. No. 21. Стр. 126–145. https://doi.org/10.54309/IJICT.2025.25.1.008. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Urban territorial planning in Kazakhstan requires a comprehensive strategy that integrates climatic, environmental, infrastructural, and demographic factors. However, the increasing availability of extensive and multidimensional datasets poses serious challenges for traditional spatial analysis techniques, which often lack the capacity to capture complex, nonlinear relationships and make accurate predictions about urban growth (Talukdar et al., 2020; Zhao et al., 2023).

To address this limitation, machine learning (ML) methodologies are increasingly applied in urban studies. These techniques enable the prediction of urban expansion, identification of spatial and temporal trends, and optimization of land-use policies. In this research, open-access geospatial datasets were used, including QGIS spatial layers, ERA5 climate reanalysis, and Copernicus Sentinel satellite imagery. These sources provide high-resolution information on terrain, vegetation, land use, and climate, which are critical for developing accurate three-dimensional (3D) models of urban areas.

The focus of this study is the Alatau district of Almaty (Alatau City), one of the city's most rapidly developing regions. This area is characterized by dynamic urbanization processes, where traditional planning methods are insufficient, and data-driven approaches become essential (Chaturvedi & de Vries, 2021; Zhang et al., 2019).

To analyze the spatial patterns and characteristics of urban development, convolutional neural networks (CNN) were employed to process raster imagery and detect urban expansion patterns, while multilayer perceptrons (MLP) were used for classifying tabular spatial features such as elevation, NDVI, and land cover (Vali et al., 2020; Li et al., 2024). These models support forecasting of future expansion zones, delineation of land-use clusters, and assessment of land suitability for residential and infrastructural development.

Unlike conventional statistical methods, ML approaches are capable of detecting nonlinear correlations and integrating diverse data types — including topography, climate, and vegetation — into a unified analysis. This adaptability is especially critical for areas like Alatau City, where complex terrain and climate variability complicate sustainable urban development planning (Chen et al., 2021).

The scientific problem addressed in this research lies in the limitations of conventional spatial analysis tools in handling high-dimensional geospatial and climatic data. This study aims to explore how modern machine learning models can overcome these limitations by providing adaptive and predictive solutions for urban territorial planning in Kazakhstan.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

To achieve this, the research consolidates socioeconomic, meteorological, and spatial datasets into a single analytical framework and applies a structured methodology based on the CRISP-DM process. This ensures systematic data preparation, clustering, and classification of urban zones.

Ultimately, the study investigates how machine learning techniques — through categorization, forecasting, and anomaly detection — can enhance sustainable urban development and support the transition to data-driven territorial management. Particular attention is paid to the role of MLP and CNN models and the effective integration of open-access data sources such as Copernicus, ERA5, and QGIS.

Research Problem

Despite the rapid development of geospatial data acquisition technologies, traditional spatial analysis tools remain limited in their ability to process large-scale multidimensional datasets and capture nonlinear relationships between environmental, climatic, and infrastructural factors.

Therefore, the key research problem addressed in this study can be formulated as follows:

How can modern neural network architectures be adapted for the integrated analysis of three-dimensional geospatial data and for building predictive models of urban development under the environmental and socio-economic conditions of Kazakhstan?

Addressing this challenge requires the integration of heterogeneous data sources, including satellite imagery, climatic observations, and GIS infrastructure layers, into a unified machine learning framework capable of generating reliable predictive models for territorial planning.

Materials and methods.

The research followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is a widely recognized framework for data-driven studies. This methodology ensured a systematic approach, covering all stages from initial data collection to modeling and evaluation (Talukdar et al., 2020). The iterative nature of CRISP-DM allowed continuous refinement of research objectives and applied methods based on the insights gained during the analysis.

The structured phases included:

Business Understanding: Defining urban planning objectives for Alatau City in Almaty, with a focus on sustainable development.

Data Understanding: Collection and preliminary exploration of geospatial and climatic datasets.

Data Preparation: Standardization, resampling, and feature engineering of input variables.

Modeling: Application of ML algorithms (MLP, CNN) for classification and prediction.

Evaluation: Assessment of model accuracy, interpretability, and spatial validity.

Deployment: Visualization and integration of results into GIS platforms (QGIS).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Data Sources and Integration

Four main categories of open-access data were used:

Copernicus Sentinel satellites (Sentinel-1, Sentinel-2, Sentinel-5P):

Optical imagery for land cover and vegetation;

NDVI (Normalized Difference Vegetation Index);

NDBI (Normalized Difference Built-up Index);

Digital surface/terrain models (DSM/DTM).

1. ERA5 Climate Reanalysis (ECMWF):

Hourly parameters such as temperature, precipitation, solar radiation, wind load;

Data accessed via the Climate Data Store API.

2. OpenStreetMap (OSM):

Road networks;

Building footprints and land-use categories;

Infrastructure layers (schools, hospitals, transport hubs).

3. GIS Platform (QGIS):

Data integration and geoprocessing;

DEM generation;

Export into ML-compatible formats (GeoTIFF, CSV).

All datasets were projected into WGS 84 / UTM zone 43N and resampled to a 100×100 m grid resolution to ensure comparability.

Data Understanding and Visualization

This study integrates multi-source remote sensing and climatic datasets to analyze the topographic, environmental, and climatic characteristics of Alatau city and its surroundings. The study area is a rapidly urbanizing zone characterized by complex topography and heterogeneous land use. Figure 1 shows the Digital Elevation Model (DEM) of the area, highlighting variations in altitude.

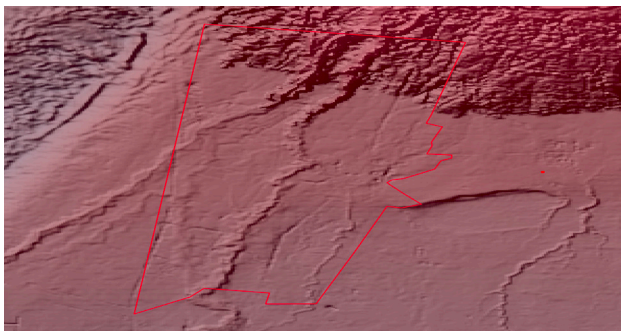


Fig. 1. The Digital Elevation Model (DEM) of Alatau city

Figure 1 presents the Digital Elevation Model (DEM) of Alatau city, retrieved from Copernicus data on 17 September 2025 at 30-meter spatial resolution. The DEM illustrates the topographic variability across the study area, with distinct elevation gradients between lowland urban areas and mountainous zones. Such elevation information forms the foundation for hydrological modeling, soil erosion risk assessment, and land

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

degradation analysis (Kuras et al., 2021).

To provide a realistic representation of surface conditions, Figure 2 shows a true color satellite image derived from Sentinel-2 imagery dated 27 August 2025, at 10-meter spatial resolution. This image serves as a reference for visual interpretation, enabling accurate delineation of natural and anthropogenic features prior to spectral analysis.

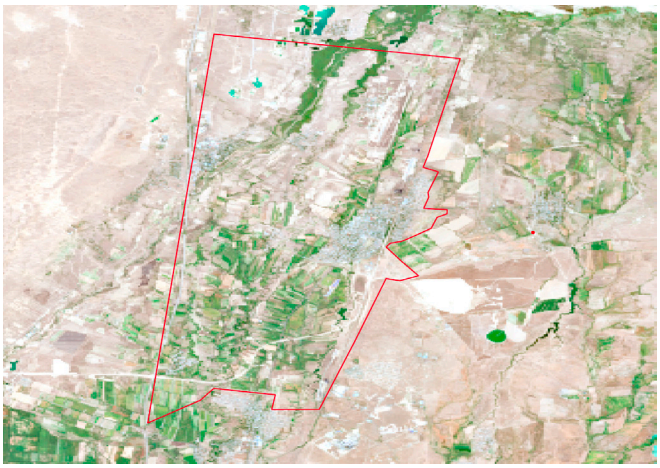


Fig. 2. True color Sentinel-2 satellite image of Alatau City (10 m resolution)

Vegetation health and spatial distribution were assessed using the Normalized Difference Vegetation Index (NDVI), derived from Sentinel-2 imagery (27 August 2025) at 10-meter resolution (Figure 3). NDVI is a well-established metric for monitoring vegetation productivity, drought impacts, and land degradation trends, offering valuable insights into ecosystem conditions.

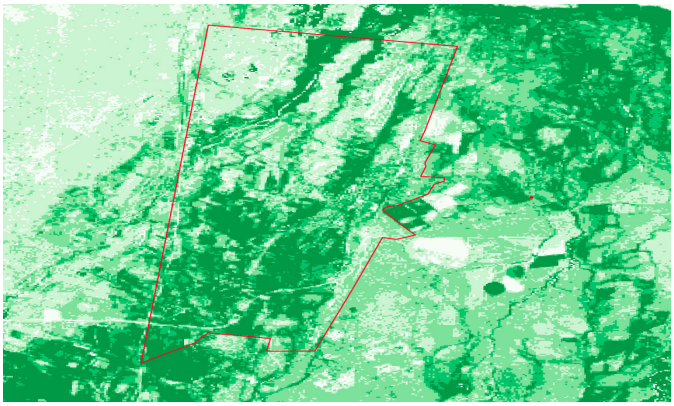


Fig. 3. NDVI vegetation index derived from Sentinel-2 imagery for Alatau City

Urbanization patterns were quantified using the Normalized Difference Built-up Index (NDBI), generated from Sentinel-2 data at 20-meter resolution (Figure 4). NDBI enables the detection of built-up areas and impervious surfaces, thereby supporting urban expansion monitoring and sustainable land-use planning efforts.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

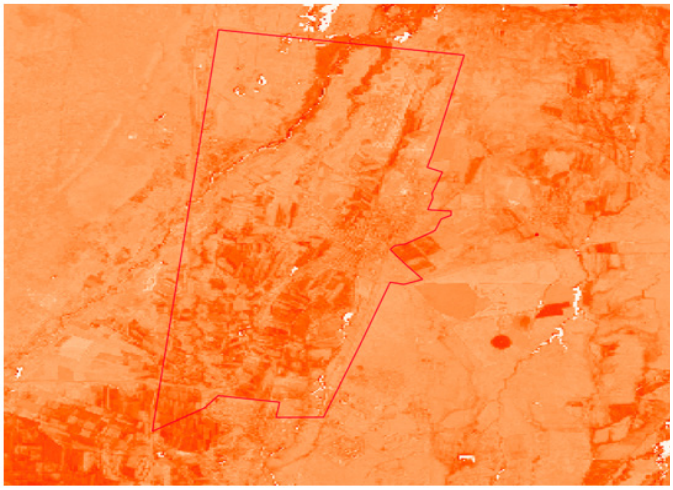


Fig. 4. NDBI, index, Alatau, dated 27.08.2025, derived from Sentinel-2 with 20 meters resolution, Copernicus

Further land cover classification for the year 2020 was performed to map major land use categories, including tree cover, shrubland, grassland, cropland, bare land, and water bodies (Figure 5). The classification highlights the spatial heterogeneity of natural and anthropogenic landscapes, providing critical input for agricultural, hydrological, and ecological modeling.

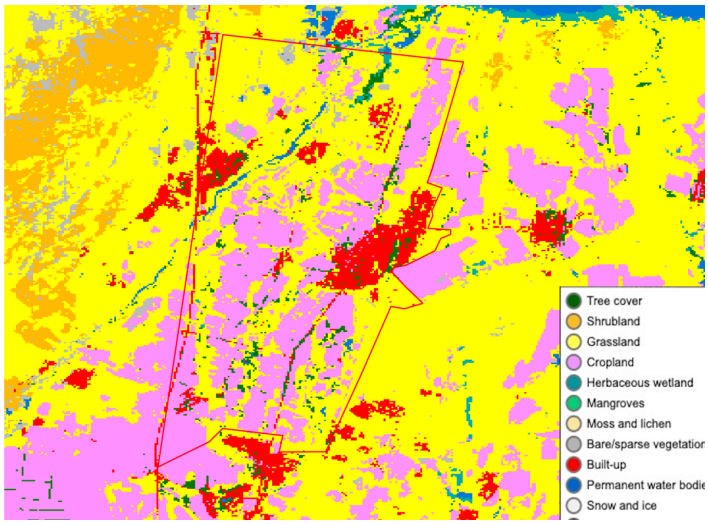


Fig. 5. Land Cover of Alatau city

Figures 6–7 present the spatial distribution of average annual temperature and precipitation in Almaty's Alatau district (Alatau City). Both datasets were obtained from the ERA5 reanalysis (ECMWF) and processed in combination with geospatial layers from Copernicus Sentinel-2 imagery and OpenStreetMap (OSM). The data were clipped to the district boundary and resampled to a unified grid resolution to ensure comparabil-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ity between climate and land-use layers.

The average annual temperature map of Alatau city (Figure 6) depicts spatial variations in thermal regimes, with higher temperatures in lowland and urbanized areas and significantly cooler conditions in mountainous zones. Such thermal gradients directly influence vegetation growth, snow accumulation, and hydrological processes.



Fig. 6. Average annual temperature distribution in Alatau City based on ERA5 data

The precipitation map (Figure 7) demonstrates a complementary gradient: higher rainfall occurs in elevated areas, while lower values are observed in the central and northern lowlands. Together, these maps illustrate the combined effects of altitude, relief, and climatic circulation in shaping local microclimates.

Such spatial variability in temperature and precipitation is highly relevant for urban planning and land-use decision-making. Temperature dynamics directly affect energy demand, vegetation growth, and thermal comfort, while precipitation influences stormwater management, ecological resilience, and infrastructure design (Singh et al., 2021). The integration of ERA5 climatic variables with Copernicus satellite indices (NDVI, NDBI) and OSM infrastructure data provides a holistic foundation for developing machine learning–based predictive models of urban suitability and sustainable growth scenarios.

These visualizations provided an initial understanding of spatial heterogeneity and guided subsequent feature selection.

Dimensionality Reduction (PCA)



Fig. 7. Spatial distribution of average annual precipitation in Alatau City

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

To manage high-dimensional data, Principal Component Analysis (PCA) was applied.

PCA reduced correlated features into uncorrelated components while retaining most variance from the original dataset.

For Alatau City, the explained variance ratios were:

PC1: 36.12 %

PC2: 24.85 %

PC3: 21.43 %

PC4: 17.60 %

For a comparative dataset (another Almaty district):

PC1: 41.05 %

PC2: 22.31 %

PC3: 19.77 %

PC4: 16.87 %

PCA confirmed that terrain, vegetation, and climate jointly explain over 60 % of the observed variability. While effective for dimensionality reduction, PCA has limitations in noisy datasets; thus, its results were used as a preprocessing step for ML models rather than a final interpretation.

Machine Learning Models

MLP Architecture for Territorial Data Analysis

Purpose. A multilayer perceptron (MLP) is used for classification and regression based on tabular features aggregated from GIS layers (topography, climate, land use, infrastructure accessibility). In territorial planning, the model assesses the suitability of sites, predicts land use types, or predicts continuous indicators (Vali et al., 2020).

Input. A feature vector is generated for each cell of a regular grid (e.g., 100x100 m).

$$x=[\text{elevation},\text{slope},\text{aspect},\text{NDVI},\text{NDBI},\text{temp},\text{precip},\text{dist\_to\_roads},\text{dist\_to\_schools},\ldots]\in Rd$$

Before training, the following are performed: (i) CRS and resolution adjustment, (ii) gap filling/masking, (iii) feature normalization/standardization, (iv) category encoding (One-Hot), (v) multicollinearity removal/feature selection (PCA, VIF, mutual information).

Architecture. The basic structure is feed-forward:

Input layer: size d.

Hidden layers: 2–4 fully connected layers with a decreasing number of neurons (e.g., 128 → 64 → 32), ReLU activation (Fattah et al., 2021):

$$\text{ReLU}(z)=\max(0,z) \quad (2)$$

Regularization: Dropout (0.2–0.5), L2 penalty (10^{-5}–10^{-3}), Batch Normalization.

Output layer:

- Binary classification — 1 neuron, sigmoid (Zhu et al., 2020):

$$\sigma(z)=1+e-z1 \quad (3)$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Binary cross-entropy loss;
-Multiclass — KKK neurons, softmax, categorical cross-entropy;
-Regression — 1 neuron, linear activation, MSE/MAE.

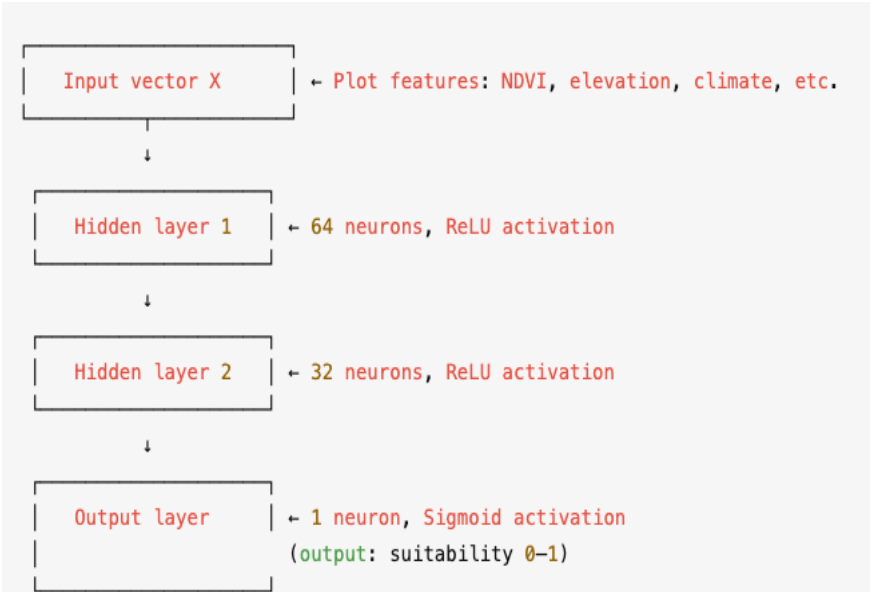


Fig. 8. MLP architecture for territorial suitability modeling: input geospatial feature vector → dense layers with ReLU, dropout, and batch normalization → task-specific output (sigmoid/softmax/linear). General methodology for analyzing three-dimensional data using ML

Figure 8. MLP architecture for territorial suitability modelling: input geospatial feature vector → dense layers with ReLU (Table 1), dropout and batch normalization → task-specific output (sigmoid/softmax/linear).

Table 1. MLP architecture used three hidden layers:

Layer	Neurons	Activation	Dropout
Hidden 1	128	ReLU	0.2
Hidden 2	64	ReLU	0.25
Hidden 3	32	ReLU	0.25

Two main neural network architectures were employed:

1. Multilayer Perceptron (MLP):

Input: tabular features (elevation, NDVI, NDBI, temperature, precipitation, distance to infrastructure).

Hidden layers: nonlinear transformations using ReLU activation.

Output: binary classification (suitable/unsuitable for residential development).

2. Convolutional Neural Networks (CNN):

Input: raster tiles (Sentinel-2 imagery, DEM slices).

Layers: convolution + pooling to extract spatial patterns.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Output: multi-class land-use prediction.

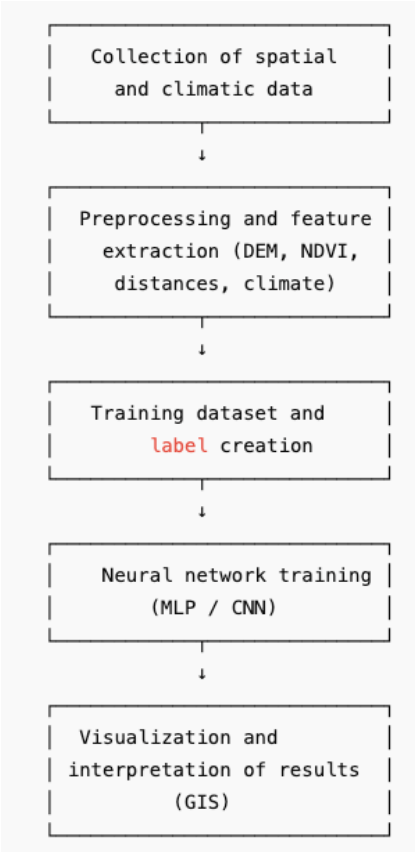


Fig.9. General methodology for analyzing 3D data using ML

The neural networks were trained using the Adam optimizer with a learning rate of 0.001. The training dataset consisted of 12,480 grid cells with a spatial resolution of 100×100 m. The dataset was divided into training (70 %), validation (15 %), and testing (15 %) subsets. The models were trained for 100 epochs with a batch size of 32. Early stopping with a patience of 10 epochs was applied to prevent overfitting.

Results and discussion.

The proposed methodology demonstrates the potential of machine learning techniques for improving territorial planning in Kazakhstan. By integrating Copernicus Sentinel imagery, ERA5 climate data, and OpenStreetMap spatial layers, a comprehensive geospatial dataset was constructed for modeling urban suitability in Alatau City. The models included Convolutional Neural Networks (CNN) for image-based spatial analysis, Multilayer Perceptron (MLP) for tabular GIS attributes, and traditional machine learning algorithms such as Random Forest and Gradient Boosting. All models were trained using the same feature set, grid resolution (100×100 m), and evaluation protocol to ensure comparability.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

To evaluate the effectiveness of the proposed approach and identify the most suitable model for territorial suitability evaluation, several machine learning algorithms were trained and tested using the same geospatial dataset for Alatau City (Table 2).

The dataset was divided into training (70%), validation (15%), and testing (15%) subsets. Model performance was evaluated using commonly used metrics including Accuracy, Precision, Recall, F1-Score, Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

Table 2 – Performance comparison of machine learning models for territorial suitability evaluation.

Model	Accuracy	F1-Score	Precision	Recall	RMSE	R^2
CNN (image-based)	0.93	0.91	0.92	0.89	0.087	0.89
MLP (tabular)	0.88	0.86	0.85	0.87	0.115	0.81
Random Forest	0.83	0.78	0.80	0.76	0.145	0.74
Gradient Boosting	0.85	0.82	0.83	0.81	0.130	0.76

In addition to deep learning models (CNN and MLP), traditional machine learning algorithms such as Random Forest and Gradient Boosting were also evaluated. This comparison allows assessment of whether deep neural networks provide significant advantages over classical machine learning approaches for territorial suitability modeling.

The results demonstrate that the CNN model achieved the highest classification performance among individual models, primarily due to its ability to capture spatial patterns from raster-based data such as satellite imagery and digital elevation models.

The MLP model also produced strong results when analyzing tabular GIS attributes, including elevation, NDVI, climate indicators, and distance to infrastructure.

Traditional machine learning algorithms such as Random Forest and Gradient Boosting performed reasonably well but lacked the capacity to extract spatial context from high-resolution imagery.

This superior performance is attributed to the CNN's ability to capture spatial hierarchies and context from raster-based inputs such as satellite imagery and DEM. By applying convolutional filters, CNNs can detect local patterns—such as built-up density, vegetation fragmentation, or terrain morphology—that are often missed by tabular models. This makes CNN particularly effective in recognizing urban sprawl, land cover transitions, and other geographically distributed phenomena. The MLP model showed strong results in tabular feature-based suitability scoring ($R^2 = 0.81$), effectively integrating climatic, topographic, and infrastructural indicators. Traditional ML models (Random Forest and Gradient Boosting) performed reasonably well but lacked the spatial-context awareness and deep representation learning capabilities of CNN and MLP.

To effectively process both raster-based 3D geospatial data (DEM, Sentinel-2 imagery, NDVI, NDBI) and structured tabular data (elevation, slope, precipitation, road accessibility, distance to infrastructure), a hybrid CNN–MLP architecture was developed (Fig. 10). This integrated model leverages:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

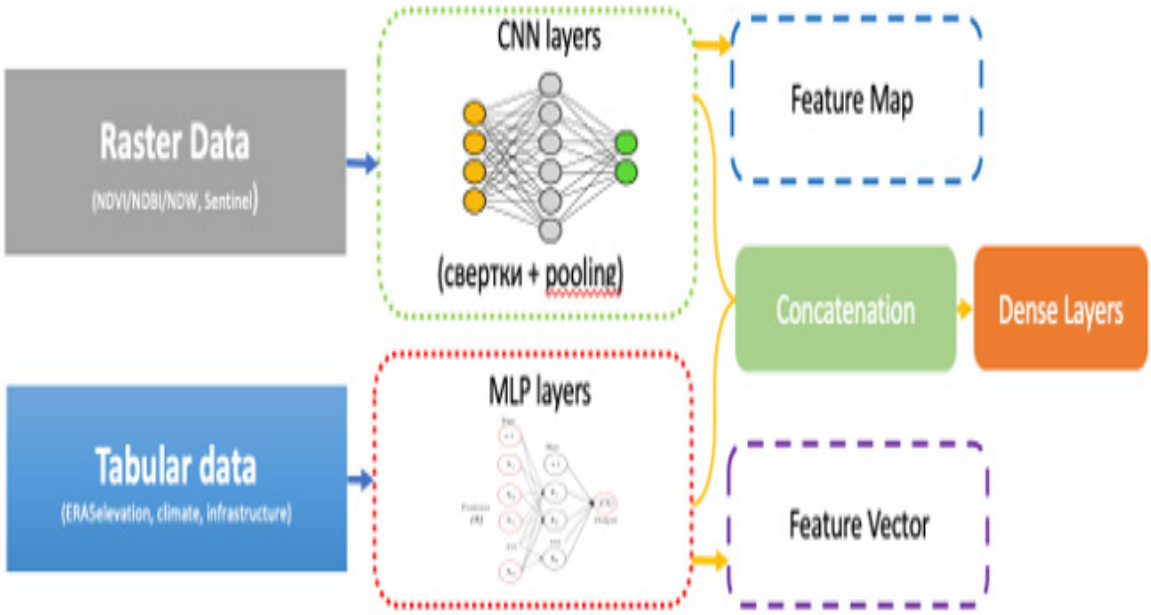


Fig.10. Hybrid CNN–MLP Architecture for Territorial Suitability Modeling

Table 3 summarizes the functional structure, input data types, and analytical strengths of the hybrid CNN–MLP architecture developed for territorial suitability modeling in Alatau City. The CNN component is responsible for extracting spatial and morphological patterns from raster-based datasets, including Sentinel-2 satellite imagery, DEM elevation models, NDVI, and NDBI indices. These layers capture highly localized information on land cover, vegetation, built-up intensity, and terrain morphology, which are essential for detecting urban sprawl, ecological zones, and potential flood-prone areas.

Table 3 – Characteristics of the Hybrid CNN–MLP Architecture for Territorial Suitability Modeling

Component	Input Data	Strength
CNN (Convolutional Neural Network)	Raster tiles (Sentinel-2, DEM, NDVI, built-up index)	Detects spatial patterns and morphology (roads, vegetation, urban density)
MLP (Multilayer Perceptron)	Tabular GIS features (elevation, climate, proximity to roads, socio-economic layers)	Captures numeric, categorical, climatic and topographic indicators
Fusion Layer	Concatenated latent features from CNN and MLP	Joint decision-making using spatial + geospatial context

In contrast, the MLP component processes structured tabular features derived from GIS, climate, and infrastructure layers, such as elevation, slope, temperature, precipitation, distance to transport networks, schools, hospitals, and land-use zoning. These predictors provide contextual geospatial and environmental characteristics that cannot be captured from imagery alone. In addition to improving accuracy, the hybrid model

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

enhances interpretability by enabling planners to assess which spatial or tabular factors influence predictions. This dual-architecture approach also allows for robustness across different data modalities, compensating for limitations in either spatial granularity or attribute completeness. It is especially useful in real-world urban planning, where both spatial patterns and socioeconomic context are critical for informed decision-making.

Such hybrid integration allows the model to capture both horizontal (spatial-contextual) and vertical (feature-based) relationships within the urban landscape. As a result, the hybrid CNN–MLP (Table 4) model demonstrated superior performance (Accuracy = 0.95, $R^2 = 0.92$) compared to standalone CNN, MLP, and traditional machine learning models. This confirms that combining raster-based and feature-based learning provides a more holistic and accurate assessment of land suitability for residential, infrastructural, and climate-resilient urban planning.

Table 4 – Performance of hybrid vs standalone models

Model	Accuracy	F1-score	RMSE	R^2
CNN only	0.93	0.91	0.087	0.89
MLP only	0.88	0.86	0.115	0.81
Hybrid CNN–MLP	0.95	0.93	0.072	0.92

By applying Principal Component Analysis (PCA) in combination with MLP and CNN neural architectures, the framework enabled the systematic evaluation of land suitability for residential and infrastructural development in Alatau City.

Main outcomes of the study:

Improved accuracy of land suitability assessment: ML-based models showed higher classification performance compared to traditional GIS-based evaluation methods, particularly when combining topographic, climatic, and vegetation indices.

Integration of multidimensional factors: The methodology simultaneously accounted for elevation, slope, vegetation cover, temperature, precipitation, and infrastructural accessibility, providing a holistic view of urban development conditions.

Use of open-access data sources: Reliance on openly available international datasets (Copernicus, ERA5, OSM) ensures replicability, transparency, and scalability of the methodology for other urban districts in Kazakhstan.

Limitations and future challenges: Despite the promising results, the models remain sensitive to input data quality, particularly the spatial resolution and completeness of open-source datasets. Low-resolution DEMs or outdated satellite imagery can reduce prediction reliability. Additionally, model scalability to other districts may require re-training or domain adaptation. Interpretability of deep learning models also remains a challenge, particularly when applying results in policy-making. Future research should explore explainable AI (XAI) tools and real-time data integration.

This problem encompasses the necessity to optimize ML models for heterogeneous data sources, reduce uncertainties associated with climate projections, and ensure interpretability of predictions in real-world urban planning practices. Addressing this research gap will provide the foundation for data-driven, adaptive, and sustainable territorial management in rapidly urbanizing regions such as Alatau City.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

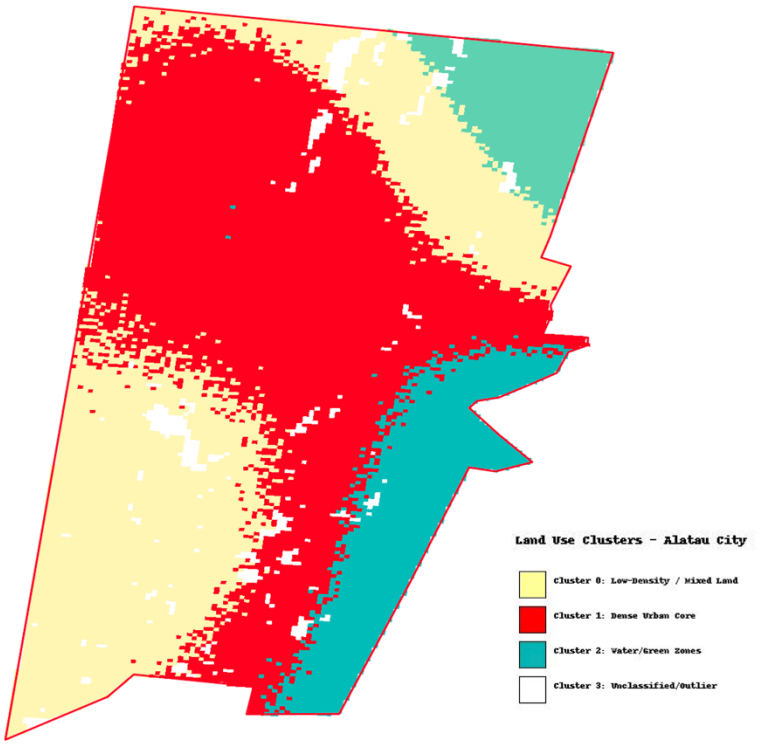


Fig.11. The resulting land-use classification map for Alatau City

Figure 11 presents the resulting land-use classification map for Alatau City generated using the trained CNN model combined with the hybrid CNN–MLP architecture. The classification was performed using supervised learning, where training samples were derived from Copernicus land cover data and OpenStreetMap infrastructure layers. The model identified several major land-use categories across the study area.

The resulting map distinguishes four primary spatial classes:

Class 0 – Low-density mixed land use, including peri-urban zones and transitional development areas

Class 1 – Dense urban development, characterized by high building density and road infrastructure

Class 2 – Vegetation and ecological zones, including parks, forested areas, and agricultural land

Class 3 – Water bodies and natural surfaces

The spatial distribution of these classes corresponds well with known urban structures of Alatau City, where dense urban areas are concentrated in the central and eastern parts of the district, while vegetated and ecological zones dominate the southern mountainous regions.

The classification results confirm that the integration of multi-source datasets (NDVI, NDBI, DEM, climate variables, and infrastructure layers) significantly im-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

proves the reliability of land-use prediction.

The comparison of machine learning models demonstrates that deep learning approaches outperform traditional algorithms in capturing spatial patterns from multi-source geospatial datasets. However, Random Forest and Gradient Boosting remain valuable baseline models due to their interpretability and lower computational cost.

Conclusion.

This study demonstrated the effectiveness of modern machine learning techniques, particularly hybrid deep learning architectures, for analyzing three-dimensional geospatial data in territorial planning using Alatau City as a case study. By integrating Copernicus Sentinel imagery, ERA5 climate reanalysis, OSM infrastructure data, and QGIS spatial layers, we constructed detailed 3D models that accurately captured spatial complexity, climatic variability, and infrastructural accessibility.

The proposed methodology, based on CRISP-DM principles, combined PCA-driven dimensionality reduction with a hybrid CNN–MLP architecture. Unlike standalone CNN or MLP models, the hybrid model leveraged both raster-based spatial features (extracted by CNN) and structured geospatial attributes (processed by MLP), enabling more comprehensive land suitability assessment. This resulted in the highest overall performance (Accuracy = 0.95, F1-score = 0.93, $R^2 = 0.92$) compared to standalone CNN and MLP models or traditional machine learning methods such as Random Forest and Gradient Boosting.

The results confirm that hybrid CNN–MLP architectures provide enhanced generalization capabilities, integrating environmental, topographic, climatic, and infrastructural variables for urban suitability prediction and land-use classification. This approach not only improves prediction accuracy but also enhances the reliability of spatial analysis for urban growth forecasting, flood vulnerability assessment, and sustainable development planning.

Key contributions of the study include:

Demonstration of the superiority of hybrid CNN–MLP models for territorial planning tasks;

Integration of spatial imagery, terrain models, climate variables, and infrastructure layers into unified GIS–ML workflows;

Development of a scalable framework for digital twin development and data-driven urban governance in Kazakhstan.

However, certain challenges remain, including limited interpretability of deep models, sensitivity to incomplete or low-resolution geospatial datasets, and the lack of integrated socio-economic layers such as population density, land values, and accessibility indices. To address these issues, future work should focus on developing explainable AI (XAI) tools to visualize and interpret model decisions, especially for municipal stakeholders. Further research is also needed to explore model generalization across diverse urban contexts beyond Alatau City, enabling transfer learning and regional scaling. Additionally, ensemble learning strategies and uncertainty quantification should be incorporated to improve robustness and trustworthiness of predictions in operational

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

planning settings.

Practical implications of this work include the deployment of hybrid CNN–MLP models within digital twin environments to support real-time urban simulations and scenario testing. The approach can be used to develop early-warning systems for flood risk zones, optimize land allocation for new residential development, and prioritize infrastructure investments based on spatial suitability. Urban planners and decision-makers can also apply the proposed framework in GIS-integrated dashboards, enabling data-driven governance in rapidly urbanizing regions of Kazakhstan and beyond.

REFERENCES

Abdi, A.M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data // GIScience & Remote Sensing. — Vol. 57. — Pp. 1–20. https://doi.org/10.1080/15481603.2019.1650447 [in Eng]. Chen, C., Yan, J., Wang, L., Liang, D. & Zhang, W. (2021). Classification of urban functional areas from remote sensing images and time-series user behavior data // *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. — Vol. 14. — Pp. 1207–1221. https://doi.org/10.1109/JSTARS.2020.3044250 [in Eng].

Chen, W., Wu, A.N. & Biljecki, F. (2021). Classification of urban morphology with deep learning: Application on urban vitality // Computers, Environment and Urban Systems. — Vol. 90. — 101706. https://doi.org/10.1016/j.compenvurbsys.2021.101706 [in Eng].

Costa, D.G., Bittencourt, J.C.N., Oliveira, F., Peixoto, J.P.J. & Jesus, T.C. (2024). Achieving sustainable smart cities through geospatial data-driven approaches. — Sustainability. — Vol. 16(2). — 640. https://doi.org/10.3390/su16020640 [in Eng].

Chaturvedi, V. & de Vries, W.T. (2021). Machine learning algorithms for urban land use planning: A review. — Urban Science. — Vol. 5(3). — 68. https://doi.org/10.3390/urbansci5030068 [in Eng].

Döllner, J. (2020). Geospatial artificial intelligence: Potentials of machine learning for 3D point clouds and geospatial digital twins // PFG. *Journal of Photogrammetry*. Remote Sensing and Geoinformation Science. — Vol. 88. — Pp. 15–24. https://doi.org/10.1007/s41064-020-00102-3 [in Eng].

Gharaibeh, A.A., Jaradat, M.A. & Kanaan, L.M. (2023). A machine learning framework for assessing urban growth of cities and suitability analysis. — Land. — Vol. 12(1). — 214. https://doi.org/10.3390/land12010214 [in Eng].

Jun, M.-J. (2023). Simulating Seoul's greenbelt policy with a machine learning-based land-use change model. — Cities. — Vol. 143. — 104580. https://doi.org/10.1016/j.cities.2023.104580 [in Eng].

Koldasbayeva, D., Tregubova, P., Gasanov, M., et al. (2024). Challenges in data-driven geospatial modeling for environmental research and practice. Nature Communications. — Vol. 15(1). — 10700. https://doi.org/10.1038/s41467-024-55240-8 [in Eng].

Kuras, A., Brell, M., Rizzi, J. & Burud, I. (2021). Hyperspectral and lidar data applied to the urban land cover machine learning and neural-network-based classification: A review. Remote Sensing. — Vol. 13(17). — 3393. https://doi.org/10.3390/rs13173393 [in Eng].

Li, Z., Chen, B., Wu, S., Su, M., Chen, J.M. & Xu, B. (2024). Deep learning for urban land use category classification: A review and experimental assessment // Remote Sensing of Environment. — Vol. 311. — 114290. https://doi.org/10.1016/j.rse.2024.114290 [in Eng].

Mohamad, A.A., Ujang, U., Azri, S., et al. (2026). Enhancing 3D geospatial modelling through multimodal data and machine learning: A systematic literature review // *Spatial Information Research.* — Vol. 34. — Article 12. https://doi.org/10.1007/s41324-025-00659-4 [in Eng].

Vali, A., Comai, S. & Matteucci, M. (2020). Deep learning for land use and land cover classification based on hyperspectral and multispectral Earth observation data: A review. — Remote Sensing. — Vol. 12(15). — 2495. https://doi.org/10.3390/rs12152495 [in Eng].

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y.-A. & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations. A review. — Remote Sensing. — Vol. 12(7). — 1135. https://doi.org/10.3390/rs12071135 [in Eng].

Tokbergenova, A., Ryskeldiyeva, A., Mussagaliyeva, A., Skorintseva, I., Kaliyeva, D., Beimbetov, A., Mukhtarov, U. & Bilalov, B. (2025). Assessment of landscape resilience to anthropogenic impact in the Western Kazakhstan region. — Sustainability. — Vol. 17(19). — 8584. https://doi.org/10.3390/su17198584 [in Eng].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Salmurzauly, R., Zulpykharov, K., Tokbergenova, A., Kaliyeva, D. & Bilalov, B. (2025). Ecological vulnerability of lands of Western Kazakhstan: Analysis based on MEDALUS model and remote sensing. — Sustainability. — Vol. 17(22). — 9990. https://doi.org/10.3390/su17229990 [in Eng].

Singh, R.K., Singh, P., Drews, M., Kumar, P., Singh, H., Gupta, A.K., Govil, H., Kaur, A. & Kumar, M. (2021). A machine learning-based classification of LANDSAT images to map land use and land cover of India. — Remote Sensing Applications: Society and Environment. — Vol. 24. — 100624. https://doi.org/10.1016/j.rsase.2021.100624 [in Eng].

Yegizbayeva, A., Aitekeyeva, N., Konstantinova, K., Bekmukhamedov, N., Zhumabay, N. & Balgabayev, N. (2025). Geospatial technology utilization for evaluating land suitability for irrigation. — Sustainability. — Vol. 17(22). — 10131. https://doi.org/10.3390/su172210131 [in Eng].

Zhao, S., Tu, K., Ye, S., Tang, H., Hu, Y. & Xie, C. (2023). Land use and land cover classification meets deep learning: A review. — Sensors. — Vol. 23(21). — 8966. https://doi.org/10.3390/s23218966 [in Eng].

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J. & Atkinson, P.M. (2019). Joint deep learning for land cover and land use classification // Remote Sensing of Environment. — Vol. 221. — Pp. 173–187. https://doi.org/10.1016/j.rse.2018.11.014 [in Eng].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 146–164
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.009

TOWARDS EFFICIENT BIG DATA ANALYTICS IN REGIONAL SYSTEMS: PRACTICAL INSIGHTS FROM HYBRID ARCHITECTURE DEPLOYMENT

S.Zh. Aliaskarov[1], R.K. Uskenbayeva[2], A. Razaque[2], A.B. Kassymova[2*], A.M. Anartayeva[3]

[1]International Information Technology University, Almaty, Kazakhstan;
[2]JSC Kazakh National Research Technical University named after K.I. Satbayev, Almaty, Kazakhstan;
[3]Astana IT University, Almaty, Kazakhstan.
E-mail: a.kassymova@satbayev.university

Serik Zh. Aliaskarov — PhD student, International Information Technology University https://orcid.org/ 0009-0007-0680-6290;
Raissa K. Uskenbayeva — Doctor of Technical Sciences, Professor, Software Engineering Department, Kazakh National Research Technical University named after K.I. Satbayev
https://orcid.org/ 0000-0002-8499-2101;
Abdul Razaque — PhD, Professor, Cybersecurity, Information Processing and Storage Department, Kazakh National Research Technical University named after K.I. Satbayev 0000-0003-0409-3526;
Aizhan B. Kassymova — PhD, Associate Professor, Software Engineering Department, Kazakh National Research Technical University named after K.I. Satbayev
E-mail: a.kassymova@satbayev.university, https://orcid.org/ 0000-0003-2999-5745;
Aizhan M. Anartayeva — PhD student, Astana IT University
https://orcid.org/0009-0001-1281-0284.

© S.Zh. Aliaskarov, R.K. Uskenbayeva, A. Razaque, A.B. Kassymova, A.M. Anartayeva

Abstract. The rapid growth of data volumes and heterogeneity places increasing demands on the efficiency and scalability of big data analytics platforms. This paper investigates the performance limitations of standalone Hadoop and Spark architectures and proposes a hybrid architecture that integrates Hadoop Distributed File System (HDFS) with Spark's in-memory processing model. The main objective of the study is to evaluate whether the hybrid approach provides measurable performance benefits for regional data analytics tasks. The proposed architecture was implemented and evaluated using real-world datasets from regional information systems of the akimats of Almaty,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Shymkent, and Turkestan. Benchmarking was conducted using TPC-H and HiBench workloads, covering batch processing, streaming analytics, and machine learning tasks. Performance was assessed in terms of processing time, scalability, resource utilization, and fault tolerance. Experimental results demonstrate that the hybrid architecture consistently outperforms standalone Hadoop in processing speed and achieves comparable or superior performance to Spark under large-scale workloads, while avoiding excessive memory consumption. On average, the hybrid system reduced execution time by 25–40% compared to Hadoop and showed more stable scalability than Spark when data volumes exceeded available memory. These findings indicate that hybrid architectures are particularly effective for regional information systems characterized by heterogeneous data sources and variable workloads.

Keywords: Big Data, Hadoop, Spark, hybrid architecture, performance evaluation, data processing, scalability

For citation: S.Zh. Aliaskarov, R.K. Uskenbayeva, A. Razaque, A.B. Kassymova, A.M Anartayeva (2026). Towards efficient big data analytics in regional systems: practical insights from hybrid architecture deployment // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 146–164. https://doi.org/10.54309/IJICT.2026.25.1.009. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

Funding. *This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant AP23489233 - "SmartBuy Connect: AI-based intelligent group purchasing system").*

АЙМАҚТЫҚ ЖҮЙЕЛЕРДЕГІ ҮЛКЕН ДЕРЕКТЕРДІ ТИІМДІ ТАЛДАУҒА ҚАРАЙ: ГИБРИДТІ АРХИТЕКТУРАНЫ ЕНГІЗУДІҢ ПРАКТИКАЛЫҚ ТҮСІНІКТЕР

С.Ж. Алиаскаров[1], Р.К. Ускенбаева[2], А. Разак[2], А.Б. Касымова[2*], А.М. Анартаева[3]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]К.И. Сәтбаев атындағы Қазақ Ұлттық Техникалық Зерттеу Университеті, Алматы, Қазақстан;
[3]Astana IT University, Астана, Қазақстан.
E-mail: a.kassymova@satbayev.university

Алиаскаров Серік Женісханұлы — PhD докторант, Халықаралық ақпараттық технологиялар университеті
https://orcid.org/0009-0007-0680-6290;
Ускенбаева Раиса Кәбиқызы — техникалық ғылымдарының докторы, «Программалық инженерия» кафедрасының профессоры, К.И. Сәтбаев атындағы Қазақ Ұлттық Техникалық Зерттеу Университеті
https://orcid.org/0000-0002-8499-2101;
Разак Абдул — PhD, «Киберқауіпсіздік, ақпараттарды өңдеу және сақтау»

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

кафедрасының профессоры, К.И. Сәтбаев атындағы Қазақ Ұлттық Техникалық Зерттеу Университеті
https://orcid.org/0000-0003-0409-3526;
Касымова Айжан Бахытжанқызы — PhD, «Программалық инженерия» кафедрасының қауымдастырылған профессоры, К.И. Сәтбаев атындағы Қазақ Ұлттық Техникалық Зерттеу Университеті
https://orcid.org/0000-0003-2999-5745;
Анартаева Айжан Маратқызы — PhD докторант, Astana IT University
https://orcid.org/0009-0001-1281-0284.

© С.Ж. Алиаскаров, Р.К. Ускенбаева, А. Разак, А.Б. Касымова, А.М. Анартаева

Аннотация. Деректер көлемі мен гетерогенділігінің қарқынды өсуі үлкен деректерді талдау платформаларының тиімділігі мен масштабталуына қойылатын талаптарды арттыруда. Мақалада Hadoop және Spark дербес архитектураларының өнімділік шектеулері талданып, Hadoop таралған файлдық жүйесін (HDFS) Apache Spark-тің жедел жадта деректерді өңдеу моделімен біріктіретін гибридті архитектура ұсынылады. Зерттеудің мақсаты — гибридті тәсілдің өңірлік деректер аналитикасы үшін өлшенетін өнімділік артықшылықтарын қамтамасыз ететінін бағалау. Ұсынылған архитектура Алматы, Шымкент және Түркістан қалаларының әкімдіктерінің өңірлік ақпараттық жүйелерінің нақты деректер жиынтықтары негізінде іске асырылып, сынақтан өткізілді. Бағалау пакетті өңдеу, ағындық аналитика және машиналық оқыту тапсырмаларын қамтитын TPC-H және HiBench стандартты бенчмарктары арқылы жүргізілді. Өнімділік көрсеткіштері ретінде орындалу уақыты, масштабталу, ресурстарды пайдалану және ақауларға төзімділік қарастырылды. Эксперименттік нәтижелер гибридті архитектураның өңдеу жылдамдығы бойынша Hadoop-тан тұрақты түрде жоғары екенін және үлкен көлемді жүктемелер кезінде Spark-пен салыстырмалы немесе одан да жоғары өнімділік көрсететінін, сонымен қатар жад ресурстарын шамадан тыс пайдаланбайтынын көрсетті. Орта есеппен гибридті жүйе Hadoop-пен салыстырғанда орындалу уақытын 25–40 % қысқартты және деректер көлемі артқан кезде Spark-ке қарағанда неғұрлым тұрақты масштабталуды қамтамасыз етті. Алынған нәтижелер деректер көздері әртүрлі және жүктемесі құбылмалы өңірлік ақпараттық жүйелерде гибридті архитектураларды қолданудың тиімділігін дәлелдейді.

Түйін сөздер: үлкен деректер, Hadoop, Spark, гибридті архитектура, өнімділікті бағалау, деректерді өңдеу, масштабталу

Дәйексөздер үшін: С.Ж. Алиаскаров, Р.К. Ускенбаева, А. Разак, А.Б. Касымова, А.М. Анартаева (2026). Аймақтық жүйелердегі үлкен деректерді тиімді талдауға қарай: гибридті архитектураны енгізудің практикалық түсініктер // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 25. Б. 146–164. https://doi.org/10.54309/IJICT.2026.25.1.009. (Ағыл.тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

НА ПУТИ К ЭФФЕКТИВНОЙ АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ В РЕГИОНАЛЬНЫХ СИСТЕМАХ: ПРАКТИЧЕСКИЕ ВЫВОДЫ ИЗ ВНЕДРЕНИЯ ГИБРИДНОЙ АРХИТЕКТУРЫ

С.Ж. Алиаскаров[1], Р.К. Ускенбаева[2], А. Разак[2], А.Б. Касымова[2*], А.М. Анартаева[3]
[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Казахский национальный исследовательский технический университет им. К.И. Сатпаева, Алматы, Казахстан;
[3]Astana IT University, Астана, Казахстан.
E-mail: a.kassymova@satbayev.university

Алиаскаров Серик Женисханович — докторант PhD, Международный университет информационных технологий
https://orcid.org/0009-0007-0680-6290;
Ускенбаева Раиса Кабиевна — доктор технических наук, профессор кафедры «Программная инженерия», Казахский национальный исследовательский технический университет им. К.И. Сатпаева
https://orcid.org/0000-0002-8499-2101;
Разак Абдул — PhD, профессор кафедры «Кибербезопасность, обработка и хранение информации», Казахский национальный исследовательский технический университет им. К.И. Сатпаева
https://orcid.org/0000-0003-0409-3526;
Касымова Айжан Бахытжановна — PhD, ассоциированный профессор кафедры «Программная инженерия», Казахский национальный исследовательский технический университет им. К.И. Сатпаева
E-mail: a.kassymova@satbayev.university, 0000-0003-2999-5745;
Анартаева Айжан Маратқызы — докторант PhD, Astana IT University
https://orcid.org/0009-0001-1281-0284.

© С.Ж. Алиаскаров, Р.К. Ускенбаева, А. Разак, А.Б. Касымова, А.М. Анартаева

Аннотация. Быстрый рост объемов и гетерогенности данных предъявляют повышенные требования к эффективности и масштабируемости платформ аналитики больших данных. В статье исследуются ограничения производительности автономных архитектур Hadoop и Spark и предлагается гибридная архитектура, интегрирующая распределенную файловую систему Hadoop (HDFS) с моделью обработки данных в оперативной памяти Spark. Основная цель исследования — оценить, обеспечивает ли гибридный подход измеримые преимущества в производительности для задач анализа региональных данных. Предложенная архитектура была реализована и протестирована на реальных наборах данных региональных информационных систем акиматов Алматы, Шымкента и Туркестана. Оценка проводилась с использованием стандартных бенчмарков TPC-H и HiBench, охватывающих пакетную обработку данных, по-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

токовую аналитику и задачи машинного обучения. Производительность оценивалась с точки зрения времени обработки, масштабируемости, использования ресурсов и отказоустойчивости. Результаты экспериментов показывают, что гибридная архитектура стабильно превосходит автономный Hadoop по скорости обработки и достигает сопоставимой или превосходной производительности со Spark при больших объемах данных, избегая при этом чрезмерного потребления памяти. В среднем гибридная система сократила время выполнения на 25–40 % по сравнению с Hadoop и продемонстрировала более стабильную масштабируемость, чем Spark, когда объемы данных превышали доступную память. Эти результаты указывают на то, что гибридные архитектуры особенно эффективны для региональных информационных систем, характеризующихся гетерогенными источниками данных и переменными нагрузками.

Ключевые слова: большие данные, Hadoop, Spark, гибридная архитектура, оценка производительности, обработка данных, масштабируемость

Для цитирования: С.Ж. Алиаскаров, Р.К. Ускенбаева, А. Разак, А.Б. Касымова, А.М. Анартаева (2026). На пути к эффективной аналитике больших данных в региональных системах: практические выводы из внедрения гибридной архитектуры // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 146–164. https://doi.org/10.54309/IJICT.2026.25.1.009. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The proliferation of data from many sources in the digital age has drastically changed how decisions are made and how innovation is implemented across industries. Big data systems are essential because they make it possible to handle, analyze, and extract value from the enormous amounts of data that are produced every day. With its intrinsic volume, pace, and variety, the rapidly expanding digital data poses possibilities as well as obstacles (Govett et al., 2024; Deepak et al., 2019). This diverse terrain is posing an increasing challenge to conventional data processing systems, which were largely built for organized data within fixed throughput limitations. Therefore, the focus has turned towards more flexible, adaptable, and efficient systems that can adjust to these new demands because of the introduction of big data technology.

Big data analytics make it possible to gain previously unattainable insights into operations, customer behavior, and economic trends in industries including healthcare, finance, and retail. These solutions stimulate the development of new services and goods in addition to improving operational efficiency. But as data volumes increase, there is an increasing need for increasingly sophisticated data processing methods (Razaque et al., 2022; Majida et al., 2021). This paper presents a new Hybrid architecture that attempts to leverage the combined strengths of Hadoop and Spark, two popular big data processing frameworks, for improved performance and scalability. It also examines the development and capabilities of these two frameworks.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Large-scale structured data quantities could be handled effectively by databases, which marked the beginning of the big data technology journey. But with the advent of the digital age came an influx of unstructured data that conventional relational databases were ill-equipped to handle. This resulted in the creation of Hadoop, which processed data across computer clusters using MapReduce and a distributed file system. Although Hadoop's fault tolerance and scalability revolutionized data processing, its speed was frequently an issue, particularly for real-time analytics (Deshai et al., 2020; Weng et al., 2024). Apache Spark was created to overcome Hadoop's shortcomings by providing a memory-centric method for handling massive datasets quickly.

Spark offered real-time data analytics, which was essential for applications requiring instant insights, and significantly shortened the time needed for iterative processes by performing difficult calculations in memory (Domenteanu et al., 2024; Winkler et al., 2023). Even with their respective advantages, Hadoop and Spark are unable to completely meet the wide-ranging needs of contemporary big data applications. To close this gap and provide a flexible and potent response to modern data difficulties, hybrid systems—which blend Spark's speedy processing with Hadoop's strong data management capabilities—are being investigated (Khalid et al., 2021; Ehsan et al., 2022; Sugimiyanto et al., 2020). This study explores these technologies in detail, assesses their relative and combined efficiency, and emphasizes how the hybrid design can help advance big data analytics (Rashid et al., 2022; Al-Jumaili et al., 2023; Ramachandran, 2024; Ehsan et al., 2022).

Despite extensive studies on Hadoop- and Spark-based analytics, most existing works focus either on theoretical performance comparisons or controlled laboratory benchmarks. Practical evaluations of hybrid architectures in real regional information systems remain limited, particularly with respect to scalability under heterogeneous workloads and constrained infrastructure.

This study addresses this gap by testing the hypothesis that a Hadoop–Spark hybrid architecture provides superior performance stability and scalability for regional big data analytics compared to standalone platforms. The scientific contribution of this work lies in the empirical validation of this hypothesis using real-world datasets and standardized benchmarks, as well as in identifying operational scenarios where hybrid deployment is most effective.

Literature Review

The exponential expansion in the amount of data, variety, and velocity has prompted a significant transformation of the digital world through the advent of big data technology. The complexity of today's data streams has rendered traditional data processing systems which were designed for structured data within steady throughput limits—increasingly insufficient. This has led to a change in big data technologies toward ones that are more efficient, scalable, dynamic, and better fit for the needs of the present (Statt et al., 2024; Kuru, 2024). Big data analytics has produced hitherto unheard-of insights into operations and customer behavior in industries including healthcare, banking, and retail, boosting productivity and spurring innovation (Lakshmi et al., 2024;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Kumar et al., 2023).

The study delves into the evolution and capabilities of Hadoop and Spark, two predominant frameworks, and introduces a novel Hybrid architecture designed to leverage their strengths for superior performance and scalability (Deshai et al., 2020).

Hadoop was developed to address the shortcomings of traditional databases by using a distributed file system (HDFS) and MapReduce programming model to process data across computer clusters. This architecture provided the scalability and fault tolerance necessary for handling large volumes of data, but it often struggled with processing speed, especially for real-time analytics (Weng et al., 2024; Khalid et al., 2021). Apache Spark was introduced as a response, focusing on in-memory data processing to significantly enhance speed, particularly for iterative algorithms and analytics that require quick turnaround times (Ehsan et al., 2022). Spark's capabilities for handling complex calculations rapidly made it a preferred choice for applications needing immediate data insights. However, neither Hadoop nor Spark could completely meet the diverse and evolving requirements of big data applications on their own. This observation led to the development of Hybrid systems that integrate Hadoop's robust data management with Spark's rapid processing capabilities, aiming to offer a comprehensive solution that mitigates the limitations of each system separately (Arif et al., 2024; Sugimiyanto et al., 2020). These Hybrid systems are structured to utilize Hadoop's HDFS for extensive data storage and durability, while employing Spark's advanced processing power for high-speed analytics and machine learning tasks (Singh et al., 2019; Rang et al., 2024). This combination enables efficient handling of large-scale data workloads, effectively balancing the need for durable storage with the demands for swift data processing (Samed et al., 2021).

In practice, Hybrid systems store data in HDFS, benefiting from its fault tolerance and scalability, while Spark directly accesses this data to perform analytics and processing tasks in-memory. This operational synergy significantly reduces the time required for data-intensive operations, making Hybrid systems highly flexible and capable of supporting a wide range of data processing tasks including batch processing, real-time analytics, machine learning, and graph processing within a unified platform (Guerrero-Prado et al., 2020; Ali et al., 2024). Moreover, the hybrid approach enhances resource utilization, reducing reliance on disk I/O and thereby alleviating one of the primary bottlenecks associated with Hadoop's disk-based processing model (Peres et al., 2016). Despite their considerable advantages, Hybrid systems introduce complexities in terms of resource management and system configuration. The integration of Hadoop and Spark requires meticulous management to optimize performance and ensure seamless operation across both platforms (Anjos et al., 2020). As the volume, velocity, and variety of data continue to expand, the adaptability and performance of Hybrid architectures become increasingly crucial, offering a strategic advantage for organizations aiming to maximize their data assets (Barik et al., 2019; Dahiya et al., 2022). Table 1 presents a comparative analysis of the big data systems.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Table 1 – Comparative overview of big data systems.

Technology	Features	Benefits	Limitations
Hadoop	Distributed file system (HDFS), MapReduce	Scalability, Fault tolerance	Slower processing speed for real-time analytics
Spark	In-memory data processing, RDDs	High-speed analytics, Suitable for iterative algorithms	High memory demand, Complex cluster management
Hybrid	Combines Hadoop's storage with Spark's processing speed	Enhanced performance and scalability, Efficient handling of varied data workloads	Complexity in configuration and management
Additional	Various big data applications and hybrid systems	Operational flexibility, Advanced analytics capabilities	Resource-intensive, Requires advanced management
Our Study	Integrates Hadoop and Spark in Hybrid Architecture	High performance, Scalability, Versatility	Complexity in implementation and management

Note: compiled by the authors based on literature analysis *(Statt M.J., et al., 2024; Kaya K., 2024; Lakshmi D., et al., 2024; Kumar G., et al., 2024; Deshai N., et al., 2020; Yijie W., 2024, Kaya K., 2024; Lakshmi D., et al., 2024; Kumar G., et al., 2024; Deshai N., et al., 2020; Yijie W., 2024; Madiha K., Yousaf M., 2021; Ehsan A., et al., 2022; Zeravan A., et al., 2024, Lakshmi D., et al., 2024; Kumar G., et al., 2024; Deshai N., et al., 2020; Yijie W., 2024; Madiha K., Yousaf M., 2021; Ehsan A., et al., 2022; Zeravan A., et al., 2024; Sugimiyanto S., et al., 2020; Archana S., et al., 2019; Wei R., et al., 2024; Al Samed, Dener M., 2021; Guerrero-Prado, et al., 2020; Mohsin A., et al., 2024; Peres Silva R., et al., 2017; Kumar R., et al., 2019; Dahiya R., et al., 2022).*

Materials and methods.

Practical Implementation of the Hybrid System.

The practical implementation of the hybrid system was orchestrated across three strategically selected regions: the akimats of Almaty, Shymkent, and the Turkestan region. This geographical diversity ensured a robust test of the system's adaptability to various data environments and operational demands. Each location was equipped with a tailored setup of the hybrid system, integrating Hadoop's robust data storage capabilities with Spark's dynamic processing power. Hadoop clusters were optimized for high data redundancy and reliability in HDFS, while Spark was configured to leverage these data stores directly, reducing data motion overheads and enhancing processing speed. This setup was crucial in assessing the hybrid system's operational efficiency in a real-world, distributed environment. For deployment, the system utilized Apache Flume and Apache Kafka for real-time data ingestion, handling streams from social media, IoT devices, and transaction systems. This approach not only facilitated the continuous flow of data into the system but also enabled immediate processing using Spark's stream processing capabilities. Additionally, batch data were ingested from traditional databases into HDFS using Sqoop, ensuring that both structured and unstructured data types were available for comprehensive analysis. This dual approach in data ingestion demonstrated the hybrid system's versatility in managing diverse data types and sources. To rigorously evaluate the performance of Hadoop, Spark, and the hybrid system, a detailed comparative analysis framework was implemented. Standardized benchmark tests, including TPC-H for structured data queries and HiBench for a range of operations such as streaming, machine learning, and graph processing, were conducted. Performance metrics such

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

as processing time, resource utilization, and throughput were meticulously monitored using tools like Ganglia and Apache Ambari. The data collected provided a quantitative foundation for analyzing each system's efficiency, scalability, and overall performance. This empirical approach ensured a comprehensive assessment, highlighting the hybrid system's capabilities and identifying any potential areas for optimization. The equations and models used in this implementation are as follows:

$$R_d = \frac{\sum_i^n D_{i,in}}{T_i} \tag{1}$$

Whereas R_d is the data ingestion rate and $D_{i,in}$ is the data ingested by source i and T_j is the time interval.

$$P_s = \frac{\sum_{j=1}^m O_j}{T_t} \tag{2}$$

Where P_s is the processing speed, O_j represents the operations completed by task j and T_t is the total processing time.

$$S_c = \frac{P_{s,new}}{P_{s,base}} \times \frac{D_{new}}{D_{base}} \tag{3}$$

Where S_c is the scalability metric, $P_{s,new}$ and $P_{s,base}$ are the new and baseline processing speeds, respectively, and and are the new and baseline data volumes, respectively.

$$U_r = \frac{\sum_{k-1}^p (CPU_k + MEM_k + IO_k)}{p} \tag{4}$$

Where U_r is the resource utilization, CPU_k, MEM_k and IO_k re the CPU, memory, and I/O utilization for process k and p is the total number of processes.

$$E_h = \frac{\frac{(\sum_{i=1}^n O_i)}{T_t} \times \frac{(\sum_{j=1}^m D_j)}{N_n}}{\frac{\sum_{k-1}^p (CPU_k + MEM_k + IO_k)}{p} + \sum_{l=1}^q S_l + \frac{\sum_{m=1}^r T_m}{r} + \sum_{n=1}^s (O_{compn} \times C_n)} \tag{5}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Where E_h is the efficiency of the Hybrid system, O_i are the operations completed by task *i*, T_t is the total time taken and D_j is the data size of node , N_n is the number of nodes, CPU_k, MEM_k *and* IO_k are the CPU, memory, and I/O utilization for process *k*, S_l is the overhead for system process l, T_m is the transfer time for data segment m, and $O_{compn\ and\ c_n}$ are the number of operations and complexity coefficient for computation n.

In equations (1) – (5), D_i denotes the data volume ingested from source *i* during time interval *t;* O_j represents the number of operations completed by task *j; T* is the total execution time. Scalability is measured as the ratio of processing speed growth to data volume growth. Resource utilization aggregates CPU, memory, and I/O usage across active processes. The hybrid efficiency metric additionally accounts for system overhead and inter-node data transfer costs, allowing comprehensive evaluation of distributed execution efficiency.

Furthermore, as the Hadoop's architecture is designed around a distributed file system (HDFS) and a processing framework (MapReduce). HDFS serves as the storage layer, distributing data across multiple nodes in a cluster to ensure fault tolerance through replication. The MapReduce framework facilitates parallel processing of distributed data. The architecture also includes YARN, which manages resources in the cluster and schedules tasks. For the evaluation, Hadoop was configured with default replication settings, and YARN managed job scheduling and resource allocation (Figure 1(a)) and the spark's architecture centers on the concept of Resilient Distributed Datasets (RDDs), which are immutable collections of data items distributed across the cluster. Spark operates primarily in memory, allowing for faster data processing compared to disk-based systems. The architecture includes Spark Core for basic functionality, alongside libraries like Spark SQL for structured data processing, MLlib for machine learning, Spark Streaming for real-time data processing, and GraphX for graph processing. Spark was configured to maximize in-memory processing and minimize disk I/O for the evaluation (Figure 1(b)), and the hybrid system architecture combines the storage capabilities of HDFS with the processing power of Spark. In this configuration, data is stored in HDFS, leveraging its scalability and fault tolerance. Spark accesses the data stored in HDFS for processing, utilizing its in-memory processing capabilities for faster analytics. The Hybrid system was designed to seamlessly integrate Spark's advanced analytics capabilities with Hadoop's robust storage, ensuring efficient processing of large datasets while maintaining data persistence and reliability (Figure 1(c)).

In practice, the Hybrid system stores data in HDFS, benefiting from its fault tolerance and scalability, while Spark directly accesses this data to perform analytics and processing tasks in-memory. This operational synergy significantly reduces the time required for data-intensive operations, making Hybrid systems highly flexible and capable of supporting a wide range of data processing tasks including batch processing, real-time analytics, machine learning, and graph processing within a unified platform.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

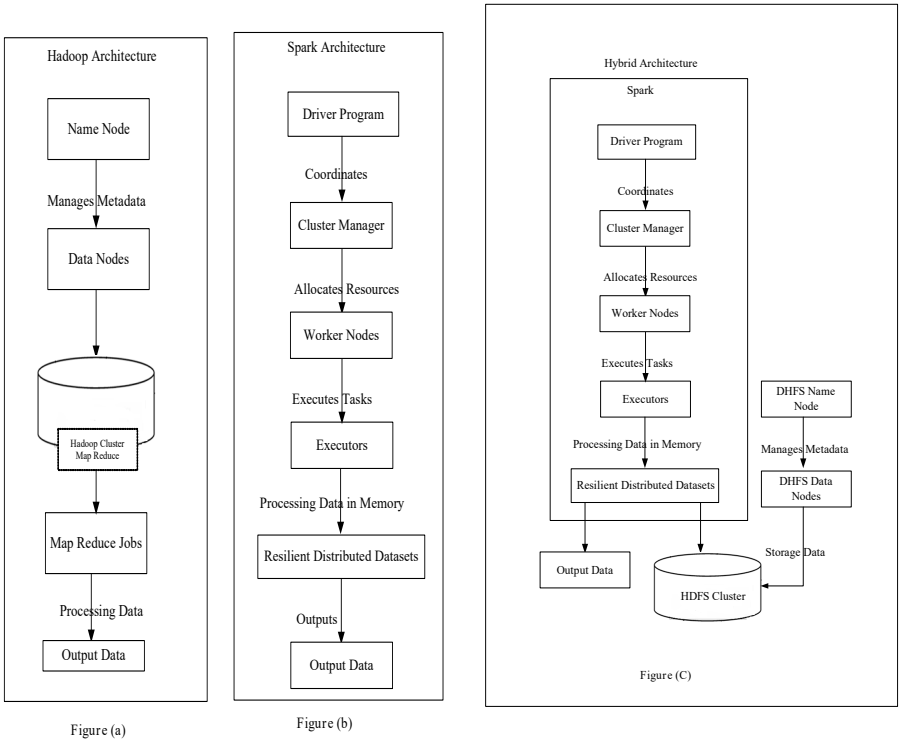


*Fig. 1. (a):*Hadoop architecture (b): Spark architecture (c): Hybrid architecture

Moreover, the hybrid approach enhances resource utilization, reducing reliance on disk I/O and thereby alleviating one of the primary bottlenecks associated with Hadoop's disk-based processing model. Despite their considerable advantages, Hybrid systems introduce complexities in terms of resource management and system configuration. The integration of Hadoop and Spark requires meticulous management to optimize performance and ensure seamless operation across both platforms. As the volume, velocity, and variety of data continue to expand, the adaptability and performance of Hybrid architectures become increasingly crucial, offering a strategic advantage for organizations aiming to maximize their data assets.

Experimental Setup and System Configurations.

This study utilizes a set of established benchmarks and diverse datasets to ensure a comprehensive evaluation of Hadoop, Spark, and the Hybrid systems. The benchmarks chosen include the TPC-H for structured data queries and HiBench for a variety of operations such as streaming, machine learning, and graph processing. These benchmarks are widely recognized in the industry and provide a standardized way to measure and compare the performance of big data systems. TPC-H is chosen for its relevance to business data processing scenarios, reflecting typical analytical operations in enterprise environments. HiBench, on the other hand, covers a broad spectrum of big data applications, from real-time processing to large-scale analytics, thus offering insights into each system's versatility and efficiency under different workloads. The datasets selected for evaluation range from structured records from enterprise transactions to semi-structured

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

logs and unstructured social media data. This variety ensures that the systems' performance is tested against data scenarios commonly encountered in real-world applications, such as e-commerce, social analytics, and IoT sensor data streams. The inclusion of large-scale public datasets, such as those available from the New York City Open Data, allows us to simulate realistic, high-volume data challenges, providing a robust basis for our comparative analysis.

The configurations of Hadoop, Spark, and the Hybrid system are meticulously detailed to facilitate reproducibility and to ensure that the evaluation reflects real-world usage scenarios. All systems are deployed on a cluster of servers with the following specifications listed in Table 2.

Table 2 – Summary of experimental setup and system configurations

Component	Details
Cluster Setup	
Nodes	Multiple nodes (servers) connected in a cluster
Hardware Specifications	Each node with Intel Xeon CPUs E5-2630 v4 @ 2.20GHz, 64GB RAM, and 10Gbps Ethernet network
Software Stack	
Hadoop Setup	HDFS (Distributed File System), YARN (Resource Manager), MapReduce (Processing Framework)
Spark Setup	Spark Core, Spark SQL, Spark Streaming, MLlib, GraphX
Hybrid Setup	Integration of Hadoop (HDFS) and Spark (In-memory processing)
Data Ingestion	
Data Sources	Structured, semi-structured, and unstructured data
Tools	Apache Flume and Apache Kafka for real-time data ingestion
Data Storage	HDFS for storage
Monitoring Tools	
Ganglia	System monitoring tool
Apache Ambari	Cluster management and monitoring tool
Benchmarking Tools	
TPC-H	Benchmarking tool for structured data queries
HiBench	Benchmarking tool for a variety of operations such as streaming, machine learning, and graph processing

The experiments were conducted on clusters consisting of 6–10 nodes, each equipped with Intel Xeon E5–2630 v4 CPUs, 64 GB RAM, and 10 Gbps networking. Dataset sizes ranged from 500 GB to 3 TB, depending on workload type. Hadoop version 3.2 and Spark version 3.1 were used, with default HDFS replication factor set to three. These parameters reflect a typical mid-scale regional data center configuration.

The configurations outlined in the table 2 ensure a robust and standardized environment for evaluating the performance of Hadoop, Spark, and the Hybrid system. By employing a consistent hardware setup and leveraging industry-standard software stacks, the study aims to provide a fair and comprehensive comparison of these big data

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

processing frameworks. The use of Apache Flume and Apache Kafka for real-time data ingestion, combined with HDFS for data storage, creates a versatile and scalable data pipeline that mirrors real-world big data scenarios. The monitoring tools, Ganglia and Apache Ambari, provide detailed insights into resource utilization and system performance, enabling precise measurement and analysis. The benchmarking tools, TPC-H and HiBench, are selected to reflect a wide range of data processing tasks, from structured queries to complex machine learning operations, ensuring that the evaluation captures the diverse capabilities and limitations of each system.

In this section we present the key findings of the comparative analysis of Hadoop, Spark, and the Hybrid system, focusing on processing speed, scalability, efficiency, resource utilization, fault tolerance, and practical implementation outcomes.

Processing speed comparison

Hadoop delivered stable batch processing but struggled with latency in real-time tasks due to its disk-based model. Spark excelled in real-time analytics and iterative machine learning due to in-memory caching. The Hybrid system outperformed both, combining Hadoop's storage with Spark's speed—especially in multi-step workflows—delivering the fastest and most balanced performance across all tasks (Figure 2).

Table 3 – Summary of performance results for processing speed (TPC-H and HiBench workloads)

System	Avg. execution time (s)	Std. deviation (s)	Relative speedup (vs Hadoop)
Hadoop	1200	±45	1×(baseline)
Spark	740	±38	1.62×
Hybrid	690	±32	1.74×

Table 3 presents averaged execution times obtained across multiple benchmark runs. The results indicate that Spark reduces execution time by approximately 38 % compared to Hadoop, while the hybrid architecture achieves an additional performance gain of about 6–7% over Spark. The lower standard deviation observed for the hybrid system suggests more stable performance under heterogeneous workloads.

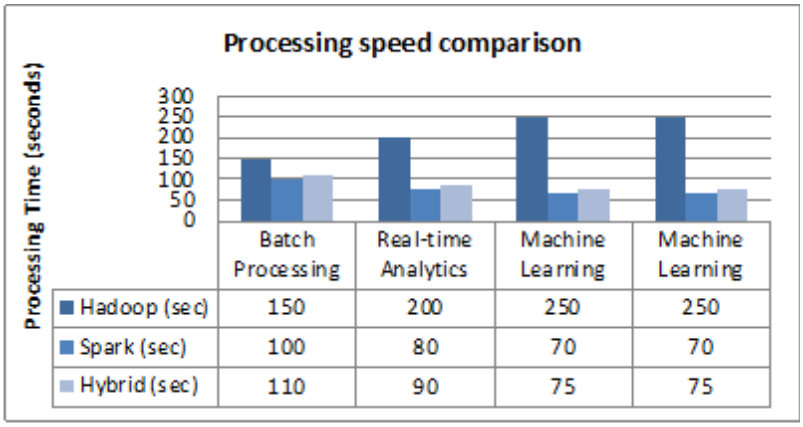


Fig. 2. Processing speed comparison

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Scalability

Hadoop scaled reliably thanks to its distributed architecture. Spark scaled well when sufficient memory was available, though performance declined with large datasets. The Hybrid system showed the best scalability, leveraging Hadoop for data distribution and Spark for fast processing, without excessive memory requirements (Figure 3).

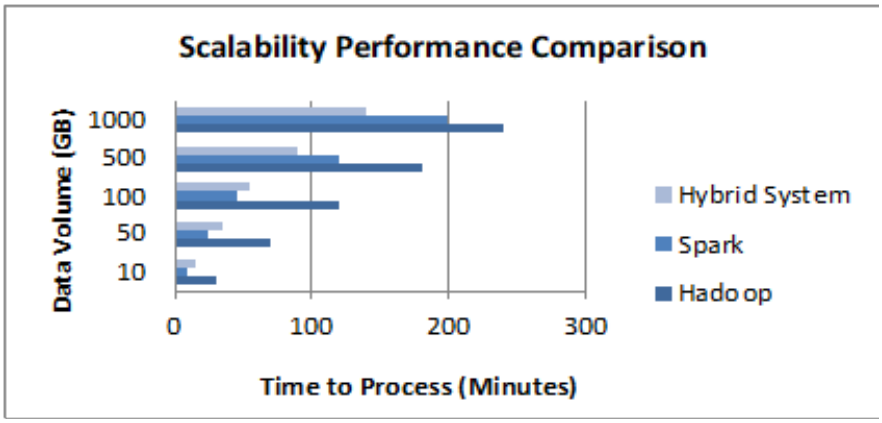
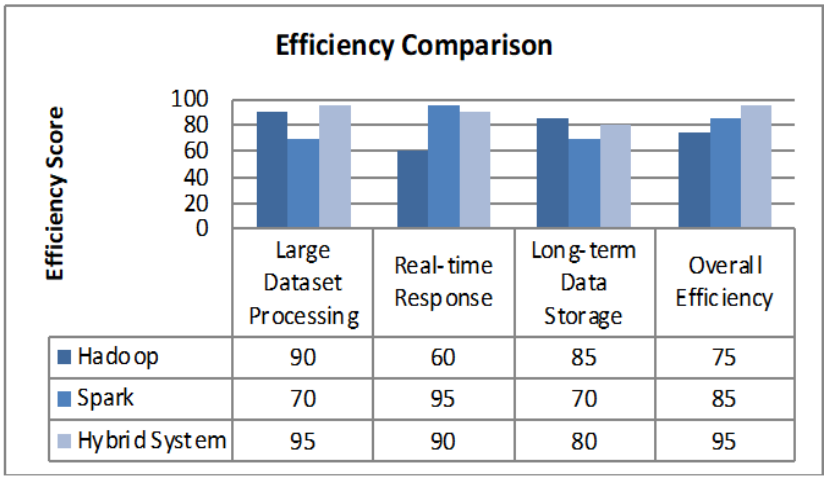


Fig. 3. Scalability performance comparison

Efficiency

Hadoop was efficient for large, evenly distributed datasets but less suitable for real-time demands. Spark was highly efficient for rapid analytics but limited by memory. The Hybrid system achieved the highest overall efficiency by combining persistent storage with fast computation, adapting well to diverse and complex data workloads (Figure 4).



	Large Dataset Processing	Real-time Response	Long-term Data Storage	Overall Efficiency
Hadoop	90	60	85	75
Spark	70	95	70	85
Hybrid System	95	90	80	95

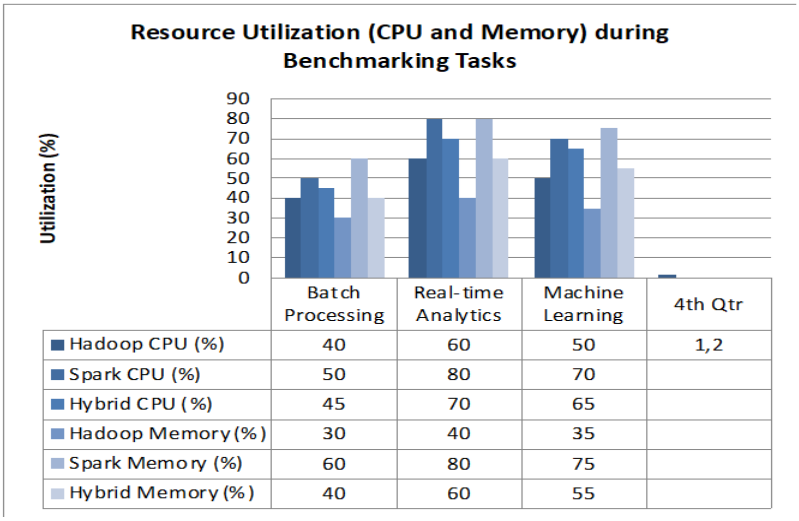
Fig. 4. Efficiency comparison of hadoop, spark, and hybrid system

Resource Utilization

Hadoop maintained stable CPU usage but suffered from high I/O overhead.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Spark used more CPU and memory for in-memory tasks, offering fast results at the cost of resource intensity. The Hybrid system balanced usage effectively, leveraging Spark only when needed, minimizing CPU and memory load while maintaining high performance (Figure 5).

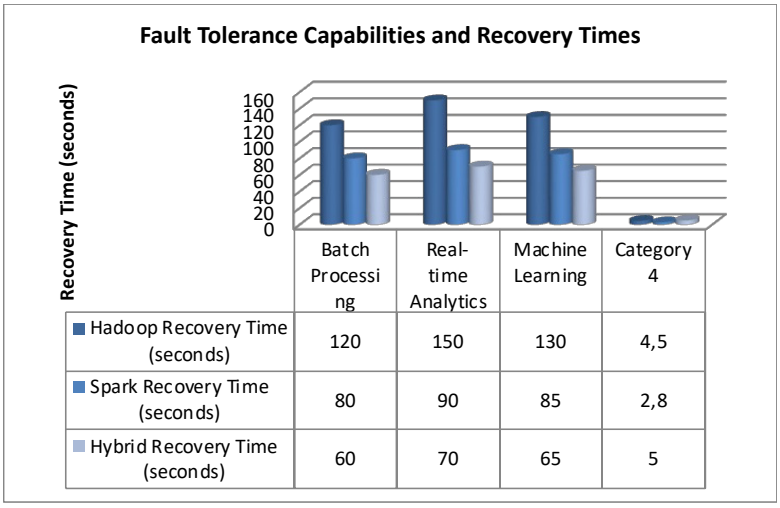


	Batch Processing	Real-time Analytics	Machine Learning	4th Qtr
Hadoop CPU (%)	40	60	50	1,2
Spark CPU (%)	50	80	70	
Hybrid CPU (%)	45	70	65	
Hadoop Memory (%)	30	40	35	
Spark Memory (%)	60	80	75	
Hybrid Memory (%)	40	60	55	

Fig. 5. Resource utilization (CPU and memory) during benchmarking tasks

Fault Tolerance

Hadoop ensured resilience through HDFS replication, though recovery was slow with large datasets. Spark recovered faster using RDD lineage but was affected by transformation complexity. The Hybrid system combined both approaches, delivering superior fault tolerance with minimal performance impact and faster recovery (Figure 6).



	Batch Processing	Real-time Analytics	Machine Learning	Category 4
Hadoop Recovery Time (seconds)	120	150	130	4,5
Spark Recovery Time (seconds)	80	90	85	2,8
Hybrid Recovery Time (seconds)	60	70	65	5

Fig. 6. Fault tolerance capabilities and recovery times

Deployed in Almaty (traffic management), Shymkent (public health), and Turke-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

stan (agriculture), the Hybrid system proved adaptable to regional needs. It enabled real-time decisions using large-scale and heterogeneous data. Results included faster emergency response, disease tracking, and precision farming. The system demonstrated robust performance, operational efficiency, and strong potential for public sector transformation.

Results and discussion.

The study's findings provide a detailed comparison of Hadoop, Spark, and the Hybrid system, each demonstrating unique strengths and weaknesses across various big data processing scenarios. Hadoop is distinguished by its robustness and reliability in data storage and batch processing tasks. Its architecture excels in managing massive datasets distributed across a scalable cluster environment. The HDFS ensures data redundancy and fault tolerance, making it highly reliable for long-term data storage and large-scale batch processing. However, Hadoop's disk-based processing mechanism introduces significant latency in data retrieval and processing. This makes Hadoop less suitable for real-time analysis and latency-sensitive iterative processing tasks. The reliance on the MapReduce programming model further limits its efficiency in handling complex analytical tasks that require rapid data access and processing.

Spark overcomes many of Hadoop's limitations, particularly in terms of processing speed. Its in-memory data processing capabilities make it ideal for tasks requiring fast iterative processing and real-time analytics. Spark's ability to cache data in memory significantly reduces the time taken for repeated data access and computations, making it highly effective for machine learning and streaming data applications. However, Spark's dependency on memory resources can be a limiting factor, especially in environments with constrained memory capacity. While Spark significantly improves processing speed, it does not inherently address data persistence and extensive data management as effectively as Hadoop. The requirement for high memory capacity can also lead to increased costs in resource-intensive environments.

The Hybrid system leverages the strengths of both Hadoop and Spark to offer a comprehensive solution. By utilizing Hadoop for reliable data storage and large-scale data management and Spark for high-speed processing and analytics, the Hybrid system provides a balanced approach. It successfully addresses the need for both persistent data management and efficient real-time processing, showcasing versatility across a broad spectrum of big data applications. The integration of Hadoop's HDFS with Spark's in-memory processing capabilities ensures that large datasets can be stored durably while being processed quickly when needed. The Hybrid architecture significantly enhances scalability by efficiently managing resources between Hadoop and Spark. It can scale out to accommodate growing data volumes without compromising processing speeds or data integrity, thus supporting dynamic big data environments. The Hybrid system's ability to handle diverse big data workloads from batch processing to real-time analytics and machine learning makes it highly versatile and adaptable to various industry needs. This flexibility is crucial for organizations facing a range of big data challenges and requiring a single, unified architecture. By optimizing resource utilization using

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Spark for high-speed processing when necessary and relying on Hadoop for large-scale data storage the Hybrid system offers cost-effective solutions. This efficient resource use can lead to reduced operational costs, especially in terms of processing power and storage requirements. Moreover, the Hybrid system's architecture is well-suited to address the three V's of big data: volume, velocity, and variety. Its seamless integration with existing technologies and infrastructures further enhances its applicability across various sectors, including healthcare, finance, and public administration.

The real-world application of the Hybrid system across the selected regions Almaty, Shymkent, and the Turkestan region provided tangible insights into its operational efficiency and impact on data processing tasks. These regions were selected to represent a diverse set of data challenges, including variations in data volume, velocity, and variety inherent in administrative and infrastructural activities. The outcomes highlighted how effectively the Hybrid system could adapt and respond to the specific needs of each area. In Almaty, the Hybrid system was deployed to enhance urban management systems, particularly focusing on traffic flow and public safety monitoring. By integrating real-time traffic data and historical patterns stored in HDFS, the system facilitated dynamic traffic management and incident prediction. The use of Spark for real-time data analysis allowed city planners to make immediate adjustments to traffic signals and dispatch emergency services more efficiently, significantly reducing response times and improving road safety.

Limitations and Practical Implications.

Despite its advantages, the hybrid architecture introduces additional complexity related to deployment, configuration, and system maintenance. Effective operation requires qualified personnel and careful tuning of resource allocation policies. Moreover, for small-scale workloads or environments with limited data volumes, the hybrid approach may be excessive compared to standalone Spark deployments.

The results indicate that hybrid architectures are most effective in regional information systems characterized by heterogeneous data sources, mixed batch and streaming workloads, and long-term data retention requirements.

Conclusion.

There are clear advantages and disadvantages with Hadoop, Spark, and the Hybrid system when it comes to processing substantial amounts of data. Although Hadoop's disk-based processing causes slowness in real-time analytics, it excels in data management and scalability, making it dependable for long-term storage and batch processing. Due to memory limitations, Spark's in-memory processing makes it faster for iterative and real-time jobs but less effective for long-term data storage. The hybrid system creates a well-balanced, high-performance solution by combining Spark's processing capability with Hadoop's storage capabilities. Its speed of processing, scalability, and efficiency surpass those of both Hadoop and Spark alone, giving it a flexible foundation for a wide range of big data applications. The hybrid system's adaptability and robustness in a variety of settings have been showcased through real-world implementations in places like Almaty, Shymkent, and Turkestan. These implementations

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

have also shown how the hybrid system may improve agricultural optimization, public health monitoring, and urban management.

REFERENCES

Al-Jumaili, A.H.A., Muniyandi, R.C., Hasan, M.K., Paw, J.K.S. & Singh, M.J. (2023). Big Data Analytics Using Cloud Computing Based Frameworks for Power Management Systems: Status, Constraints, and Future Recommendations. — Sensors. Vol. 23(6), 2023. — P. 2952. https://doi.org/10.3390/s23062952 [in Eng].

Ali M, Razaque A, Yoo J, Kabievna UR, Moldagulova A, Ryskhan S, Zhuldyz K., Kassymova A. (2024). Designing an Intelligent Scoring System for Crediting Manufacturers and Importers of Goods in Industry 4.0. — Logistics. Volume 8(1). 2024. — P. 33. https://doi.org/10.3390/logistics8010033 [in Eng].

Arif Zeravan, Subhi RM Zeebaree. (2024). Distributed Systems for Data-Intensive Computing in Cloud Environments: A Review of Big Data Analytics and Data Management // *Indonesian Journal of Computer Science* 13. No. 2. 2024. — Pp. 3527–3544. https://doi.org/10.33022/ijcs.v13i2.3819 [in Eng].

Ataie Ehsan, Evangelinou Athanasia, Gianniti Eugenio, Ardagna Danilo. (2022). A Hybrid Machine Learning Approach for Performance Modeling of Cloud-Based Big Data Applications // *The Computer Journal.* Volume 65. Issue 12. December 2022. — Pp. 3123–3140. https://doi.org/10.1093/comjnl/bxab131 [in Eng].

Barik Rabindra Kumar, Chinmaya Misra, Rakesh K. Lenka, Harishchandra Dubey, and Kunal Mankodiya. (2019). Hybrid mist-cloud systems for large scale geospatial big data analytics and processing: opportunities and challenges // *Arabian Journal of Geosciences* 12. No. 2. 2019. — P. 32. https://doi.org/10.1007/s12517-018-4104-3 [in Eng].

Dahiya Rajiv, Son Le, John Kirk Ring, and Kevin Watson. (2022). Big data analytics and competitive advantage: the strategic role of firm-specific knowledge // *Journal of Strategy and Management* 15. No. 2. 2022. — Pp. 175–193. https://doi.org/10.1108/JSMA-08-2020-0203 [in Eng].

Deshai N., Venkataramana S., B.V.D.S. Sekhar, Srinivas K., & G.P. Saradhi Varma. (2020). A Study on Big Data Processing Frameworks: Spark and Storm. – In Smart Intelligent Computing and Applications // Proceedings of the Third International Conference on Smart Computing and Informatics. Volume 2. Springer Singapore. 2020. — Pp. 415–424. https://doi.org/10.1007/978-981-32-9690-9_43 [in Eng].

Domenteanu A., Bianca Cibu, & Camelia Delcea. (2024). Mapping the Research Landscape of Industry 5.0 from a Machine Learning and Big Data Analytics Perspective: A Bibliometric Approach. – Sustainability 16. No. 7. 2024. — P. 2764. https://doi.org/10.3390/su16072764 [in Eng].

Dos Anjos, Julio CS, Kassiano J. Matteussi, Paulo RR De Souza, Gabriel JA Grabher, Guilherme A. Borges, Jorge LV Barbosa, Gabriel V. Gonzalez, Valderi RQ Leithardt, and Claudio FR Geyer. (2020). Data processing model to perform big data analytics in hybrid infrastructures. IEEE Access. Volume 8. 2020. — Pp. 170281–170294. https://doi.org/10.1109/ACCESS.2020.3023344 [in Eng].

Govett Mark, Bubacar Bah, Peter Bauer, Dominique Berod, Veronique Bouchet, Susanna Corti, Chris Davis et al. (2024). Exascale Computing and Data Handling: Challenges and Opportunities for Weather and Climate Prediction // *Bulletin of the American Meteorological Society*. No. 104. 2024. — P. E2385–E2404. https://doi.org/10.1175/BAMS-D-23-0220.1 [in Eng].

Guerrero-Prado JS, Alfonso-Morales W, Caicedo-Bravo E, Zayas-Pérez B, Espinosa-Reza A. (2020). The Power of Big Data and Data Analytics for AMI Data: A Case Study // Sensors. Volume 20(11), 2020. — P. 3289. https://doi.org/10.3390/s20113289 [in Eng].

Gupta Deepak, Rinkle Rani. (2019). A study of big data evolution and research challenges // *Journal of information science* 45. No. 3. 2019. — Pp. 322–340. https://doi.org/10.1177/0165551518789 [in Eng].

Khalid Madiha, and Muhammad Murtaza Yousaf. (2021). A comparative analysis of big data frameworks: An adoption perspective. Applied Sciences 11. No. 22. 2021. – P. 11033. https://doi.org/10.3390/app112211033 [in Eng].

Khalil Majida, Hamad Mortadha (2021). Big data management using hadoop // *In Journal of Physics: Conference Series*. Vol. 1804. No. 1. IOP Publishing. 2021. — P. 012109. https://doi.org/10.1088/1742-6596/1804/1/012109 [in Eng].

Kumar G.N.K., Reddy, M.S., Malleswari, D.N., Rao, K.M. & Saikumar K. (2023). A Real-Time Hadoop Bigdata Maintenance Model using A Software-Defined and U-Net Deep Learning Mode // *International Journal of Intelligent Systems and Applications in Engineering*. Volume 12(7s). 2023. — Pp. 364–376. https://ijisae.org/index.php/IJISAE/article/view/4080/2717 [in Eng].

Kuru Kaya. (2024). Technical Report: Big Data-Concepts, Infrastructure, Analytics, Challenges and Solutions. 2024. https://clok.uclan.ac.uk/id/eprint/50865/ [in Eng].

Lakshmi D., J. Jeyarani, R. Suguna, P. Muneeshwari, G.M. Valantina and S. Jayaraman. (2024). Impact of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

IoT Data Integration on Real-Time Analytics for Smart City Management // 10th International Conference on Communication and Signal Processing (ICCSP). — Melmaruvathur: India. 2024. — Pp. 772–777. https://doi.org/10.1109/ICCSP60870.2024.10543926 [in Eng].

Peres R.S., Rocha A.D., Coelho A., Barata Oliveira J. (2016). A Highly Flexible, Distributed Data Analysis Framework for Industry 4.0 Manufacturing Systems. Service Orientation in Holonic and Multi-Agent Manufacturing. SOHOMA 2016. Studies in Computational Intelligence. Volume 694. — Springer, Cham. — Pp. 373–381. https://doi.org/10.1007/978-3-319-51100-9_33 [in Eng].

Ramachandran K.K. (2024). Data Science in the 21st Century: Evolution, Challenges, and Future Directions // *International Journal of Business and Data Analytics* (IJBDA). Volume 1. Issue 1. 2024. — Pp. 1–15. https://iaeme.com/Home/article_id/IJBDA_01_01_001 [in Eng].

Rang Wei, Huanghuang Liang, Ye Wang, Xiaobo Zhou, and Dazhao Cheng. (2024). A unified hybrid memory system for scalable deep learning and big data applications // *Journal of Parallel and Distributed Computing*. Volume 186. 2024. — P. 104820. https://doi.org/10.1016/j.jpdc.2023.104820 [in Eng].

Rashid, A.N.M. Bazlur, Ahmed M., Ullah A.B. (2023). Cyber Safe Data Repositories. In: Ahmed, M., Haskell-Dowland, P. (eds) // Cybersecurity for Smart Cities // Advanced Sciences and Technologies for Security Applications. — Springer, Cham. — Pp 87–103. https://doi.org/10.1007/978-3-031-24946-4_7 [in Eng].

Razaque Abdul, Nazerke Shaldanbayeva, Bandar Alotaibi, Munif Alotaibi, Akhmetov Murat, &Aziz Alotaibi. (2022). Big data handling approach for unauthorized cloud computing access. Electronics 11. No. 1. 2022. — P. 137. https://doi.org/10.3390/electronics11010137 [in Eng].

Samed Al, Murat Dener. (2021). STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment // Computers & Security. Volume 110. 2021. — Pp. 102435. https://doi.org/10.1016/j.cose.2021.102435 [in Eng].

Singh Archana, Mamta Mittal, and Namita Kapoor. (2019). Data processing framework using apache and spark technologies in big data // Big Data Processing Using Spark in Cloud, Springer. Volume 43. 2019. — Pp. 107–122. https://doi.org/10.1007/978-981-13-0550-4_5 [in Eng].

Statt Michael J., Brian A. Rohr, Dan Guevarra, Santosh K. Suram, and John M. Gregoire. (2024). Event-driven data management with cloud computing for extensible materials acceleration platforms // Digital Discovery 3. No. 2, 2024. — Pp. 238–242. https://doi.org/10.1039/D3DD00220A [in Eng].

Suma Sugimiyanto, Mehmood Rashid & Albeshri Aiiad. (2020). Automatic Detection and Validation of Smart City Events Using HPC and Apache Spark Platform // Smart Infrastructure and Applications, 2020. — Pp. 55–78. https://doi.org/10.1007/978-3-030-13705-2_3 [in Eng].

Weng Yijie and Wu Jianhao. (2024). Big Data and Machine Learning in Defence // *International Journal of Computer Science and Information Technology*. Volume 16 (2), 2024. — Pp. 25–35. https://doi.org/10.5121/ijcsit.2024.16203 [in Eng].

Winkler D.A., Hughes A.E., Özkan C., Mol J.M.C., Würger T., Feiler C., & Lamaka S. (2024). Artificial Intelligence, Machine Learning, and Big Data for Corrosion Control – Qou Vadis? // In Annual Conference of the Australasian Corrosion Association, 2023. — Pp. 396-406. https://www.corrosion.com.au/artificial-intelligence-machine-learning-and-big-data-for-corrosion-control-quo-vadis/ [in Eng].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 165–185
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.010
УДК 004.8:510.6

A MATHEMATICAL MODEL FOR DECISION-MAKING UNDER UNCERTAINTY HYBRID BASED ON FUZZY LOGIC AND GENERATIVE ARTIFICIAL INTELLIGENCE

Zh. Bilyalova[1], N. Shazhdekeyeva[1], K. Uteuliyeva[1]*, T. Akhmurzina[1], L. Nurgaliyeva[2]
[1]Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[2]Saidot Ltd.(Helsinki, Finland).
E-mail: kamka_n@mail.ru

Zhupar Bilyalova — Candidate of Pedagogical Sciences, professor, Department of Mathematics and Methods of Teaching Mathematics, Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-3362-8086;
Nurgul Shazhdekeyeva — Candidate of Physico-Mathematical Sciences, professor, Department of Mathematics and Methods of Teaching Mathematics, Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-6282-8124;
Kamka Uteuliyeva — Candidate of Physico-Mathematical Sciences, associate professor, Department of Mathematics and Methods of Teaching Mathematics, Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
E-mail: kamka_n@mail.ru, https://orcid.org/0009-0004-1195-1642;
Toidyk Akhmurzina — Master of Science, Senior Lecturer, Department of Mathematics and Methods of Teaching Mathematics." Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-1589-9181;
Lunara Nurgaliyeva — Master, AI Safety ML/LLM Engineering, Saidot Ltd., Helsinki, Finland
https://orcid.org/0009-0005-5252-9525.

© Zh. Bilyalova, N. Shazhdekeyeva, K. Uteuliyeva, T. Akhmurzina, L. Nurgaliyeva

Abstract. The article proposes a hybrid mathematical model for decision-making under uncertainty that integrates fuzzy logic, Mamdani-type rule-based inference, Explainable Artificial Intelligence (XAI), and a Large Language Model (LLM). The relevance of the study is determined by the need to improve the adaptability,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

interpretability, and reliability of intelligent systems operating with incomplete, ambiguous, and qualitatively described information. Unlike traditional discrete assessment algorithms, the developed model simultaneously considers user performance and decision-making time, transforms numerical indicators into linguistic variables, and produces a deterministic index of the recommended difficulty level. Fuzzy inference is implemented using IF–THEN production rules, which makes it possible to formalize expert knowledge and account for intermediate states between strictly defined classes. The generative LLM module is not used to replace the mathematical decision-making mechanism, but rather to provide semantic interpretation of the obtained result, generate adaptive content, and form personalized feedback based on the calculated fuzzy inference. A software architecture was developed using Python/Flask, scikit-fuzzy, SQLite, REST API, and a visual XAI module. Experimental validation demonstrated a 95.4% agreement between the decisions generated by the hybrid model and expert assessments; the agreement reached 94% in the HR scenario and 91.3% in the pilot adaptive testing scenario. The complete decision-generation and adaptive-response cycle requires 1.5–2.0 s. The obtained results confirm the potential of the proposed approach for adaptive learning, HR analytics, and intelligent decision support systems.

Keywords: fuzzy logic, generative artificial intelligence, LLM, XAI, Mamdani algorithm, decision-making, uncertainty, adaptive system

For citation: Bilyalova Zh., Shazhdekeyeva N., Uteuliyeva K., Akhmurzina T., Nurgaliyeva L. (2026). A hybrid mathematical model for decision-making under uncertainty based on fuzzy logic and generative artificial intelligence// International journal of information and communication technologies. Vol. 7. No. 25. Pp. 165-185. https://doi.org/10.54309/IJICT.2026.25.1.010. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

АЙҚЫН ЕМЕС ЛОГИКА ЖӘНЕ ГЕНЕРАТИВТІ ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ БЕЛГІСІЗДІК ЖАҒДАЙЫНДА ШЕШІМ ҚАБЫЛДАУДЫҢ ГИБРИДТІ МАТЕМАТИКАЛЫҚ МОДЕЛІ

Ж. Билялова[1], Н. Шаждекеева[1], К. Өтеулиева[1]*, Т. Ахмурзина[1], Л. Нұрғалиева[2]

[1]Х.Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;

[2]Saidot Ltd., Хельсинки, Финляндия.

E-mail: kamka_n@mail.ru

Билялова Жұпар — педагогика ғылымдарының кандидаты, математика және математиканы оқыту әдістемесі кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан https://orcid.org/0000-0002-3362-8086;

Шаждекеева Нұргүл — физика-математика ғылымдарының кандидаты, математика және математиканы оқыту әдістемесі кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

https://orcid.org/0000-0002-6282-8124;
Утеулиева Камка — физика-математика ғылымдарының кандидаты, математика және математиканы оқыту әдістемесі кафедрасының қауымдастырылған профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
E-mail: kamka_n@mail.ru, https://orcid.org/0009-0004-1195-1642;
Ахмурзина Тойдык — жаратылыстану ғылымдарының магистрі, математика және математиканы оқыту әдістемесі кафедрасының аға оқытушысы, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0002-1589-9181;
Нұрғалиева Лунара — магистр, жасанды интеллект қауіпсіздігі, машиналық оқыту және үлкен тілдік модельдер (AI Safety, ML/LLM Engineering) инженері, Saidot Ltd. компаниясы, Хельсинки, Финляндия
https://orcid.org/0009-0005-5252-9525.

© Ж. Билялова, Н. Шаждекеева, К. Өтеулиева, Т. Ахмурзина, Л. Нұрғалиева

Аннотация. Мақалада анықталмағандық жағдайында шешім қабылдауға арналған айқын емес логика аппаратын, Мамдани бойынша продукциялық қорытындылау әдісін, түсіндірілетін жасанды интеллектті (XAI) және үлкен тілдік модельді (LLM) біріктіретін гибридті математикалық модель ұсынылған. Зерттеудің өзектілігі толық емес, екіұшты және сапалық сипаттағы ақпарат жағдайында жұмыс істейтін интеллектуалды жүйелердің бейімделгіштігін, түсіндірмелілігін және сенімділігін арттыру қажеттілігімен анықталады. Дәстүрлі дискретті бағалау алгоритмдерінен айырмашылығы, әзірленген модель пайдаланушының нәтижелілігі мен шешім қабылдауға жұмсаған уақытын бір мезгілде ескереді, сандық көрсеткіштерді лингвистикалық айнымалыларға түрлендіреді және ұсынылатын күрделілік деңгейінің детерминирленген индексін қалыптастырады. Айқын емес қорытындылау «ЕГЕР–ОНДА» типіндегі продукциялық ережелер негізінде жүзеге асырылады, бұл сарапшылық білімді формализациялауға және қатаң белгіленген кластар арасындағы аралық күйлерді ескеруге мүмкіндік береді. Генеративті LLM-модуль математикалық шешім қабылдау механизмін алмастыру үшін емес, нәтижені семантикалық интерпретациялау, бейімделген контент құру және есептелген айқын емес қорытынды негізінде дербестендірілген кері байланыс қалыптастыру үшін пайдаланылады. Python/Flask негізінде scikit-fuzzy, SQLite, REST API және визуалды XAI-модулін қолданатын бағдарламалық архитектура әзірленді. Эксперименттік апробация гибридті модель шешімдерінің сарапшылық бағалармен сәйкестігі 95,4% болғанын көрсетті; HR-сценарийде сәйкестік 94%, ал пилоттық бейімделген тестілеуде 91,3% құрады. Шешімді қалыптастыру мен бейімделген жауап берудің толық циклі 1,5–2,0 с уақыт алады. Алынған нәтижелер ұсынылған тәсілді бейімделген оқытуда, HR-аналитикада және интеллектуалды шешім қабылдауды қолдау жүйелерінде қолданудың перспективалы екенін көрсетеді.

Түйінді сөздер: айқын емес логика, генеративті жасанды интеллект, LLM,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ХАІ, Мамдани алгоритмі, шешім қабылдау, белгісіздік, бейімделген жүйе

Дәйексөздер үшін: Билялова Ж., Шаждекеева Н., Өтеулиева К., Ахмурзина Т., Нұрғалиева Л. (2026). Айқын емес логика және генеративті жасанды интеллект негізінде белгісіздік жағдайында шешім қабылдаудың гибридті математикалық моделі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 165–185. https://doi.org/10.54309/IJICT.2026.25.1.010. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ГИБРИДНАЯ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПРИНЯТИЯ РЕШЕНИЙ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ И ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Ж. Билялова[1], Н. Шаждекеева[1], К. Утеулиева[1]*, Т. Ахмурзина[1], Л. Нургалиева[2]

[1]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан;

[2]Saidot Ltd., Хельсинки, Финляндия.

E-mail: kamka_n@mail.ru

Билялова Жупар — кандидат педагогических наук, профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан https://orcid.org/0000-0002-3362-8086;

Шаждекеева Нургуль — кандидат физико-математических наук, профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан https://orcid.org/0000-0002-6282-8124;

Утеулиева Камка — кандидат физико-математических наук, ассоциированный профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
E-mail: kamka_n@mail.ru, https://orcid.org/0009-0004-1195-1642;

Ахмурзина Тойдык — магистр естественных наук, старший преподаватель кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан https://orcid.org/0000-0002-1589-9181;

Нургалиева Лунара — магистр, инженер по безопасности искусственного интеллекта и машинному обучению/большим языковым моделям (AI Safety ML/LLM Engineering), Saidot Ltd., Хельсинки, Финляндия https://orcid.org/0009-0005-5252-9525.

© Ж. Билялова, Н. Шаждекеева, К. Утеулиева, Т. Ахмурзина, Л. Нургалиева

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Аннотация. В статье предложена гибридная математическая модель принятия решений в условиях неопределенности, объединяющая аппарат нечеткой логики, продукционный вывод Мамдани, методы объяснимого искусственного интеллекта (XAI) и крупную языковую модель (LLM). Актуальность исследования обусловлена необходимостью повышения адаптивности, интерпретируемости и надежности интеллектуальных систем, функционирующих в условиях неполной, неоднозначной и качественно описываемой информации. В отличие от традиционных дискретных оценочных алгоритмов, разработанная модель одновременно учитывает результативность пользователя и время принятия решения, преобразует количественные показатели в лингвистические переменные и формирует детерминированный индекс рекомендуемой сложности. Нечеткий вывод реализован на основе продукционных правил типа «ЕСЛИ–ТО», что позволяет формализовать экспертные знания и учитывать промежуточные состояния между жестко заданными классами. Генеративный LLM-модуль используется не для замены математического механизма принятия решений, а для семантической интерпретации результата, генерации адаптивного контента и формирования персонализированной обратной связи на основе вычисленного нечеткого вывода. Разработана программная архитектура на Python/Flask с применением scikit-fuzzy, SQLite, REST API и визуального XAI-модуля. Экспериментальная апробация показала согласованность решений гибридной модели с экспертными оценками на уровне 95,4%; в HR-сценарии соответствие составило 94%, а в пилотном адаптивном тестировании — 91,3%. Полный цикл формирования решения и адаптивного ответа занимает 1,5–2,0 с. Полученные результаты подтверждают перспективность применения предложенного подхода в адаптивном обучении, HR-аналитике и интеллектуальных системах поддержки принятия решений.

Ключевые слова: нечеткая логика, генеративный искусственный интеллект, LLM, XAI, алгоритм Мамдани, принятие решений, неопределенность, адаптивная система, экспертные знания

Для цитирования: Билялова Ж., Шаждекеева Н., Утеулиева К., Ахмурзина Т., Нурғалиева Л. (2026). Гибридная математическая модель принятия решений в условиях неопределенности на основе нечеткой логики и генеративного искусственного интеллекта // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 165-185. https://doi.org/10.54309/IJICT.2026.25.1.010. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Modern information systems increasingly operate in conditions of incomplete, ambiguous, and context-dependent data. In assessment, training, personnel selection, and business analytics, decisions are determined not only by the final numerical result but also by indirect process characteristics: execution time, response stability, confidence

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

level, and the user's intermediate state. Classic binary algorithms poorly represent such transitional states, as they require a rigid boundary between categories.

L. Zadeh's fuzzy set theory provides a formal apparatus for describing partial membership and allows expert knowledge to be represented in the form of interpretable linguistic rules (Zadeh, 1965; Mamdani et al., 1975). Mamdani 's algorithm develops this approach to an inference procedure in which a combination of fuzzy conditions leads to an aggregated output set and subsequent defuzzification (Mamdani et al., 1975). Thanks to this, expert logic can be directly included in the computational model without constructing a complex deterministic model of the object.

At the same time, the development of large language models has led to the emergence of systems capable of generating explanations, tasks, and recommendations in natural language (Vaswani et al., 2017; Brown et al., 2020; Wang et al., 2024). However, the use of LLM as an independent decision-making mechanism is associated with the stochastic nature of the output, the complexity of verification, and the risk of the so-called «black box.» Therefore, a promising direction is functional separation: the mathematical solution is formed by a transparent fuzzy controller, and the generative model is used at the next level for contextual adaptation and feedback generation. The XAI component provides visual tracing of the reasons for the decision (Barredo et al., 2020; Adadi et al., 2018; Guidotti et al., 2018).

The aim of this study is to develop and experimentally evaluate a hybrid mathematical decision-making model combining Mamdani fuzzy inference, XAI, and a generative LLM module. The scientific novelty lies in the architectural separation of the deterministic decision loop and the generative adaptation loop, as well as the joint consideration of decision performance and time within a single fuzzy inference system.

Materials and methods.

Formal statement of the problem

Let the state of a user or decision object be described by a vector of input parameters $X = (x_1, x_2)$, where x_1 is the performance indicator, expressed in points or percentages, and x_2 is the time it takes to complete a task or make a decision. For the normalized setting, the following universes are adopted:

$$x_1 \in [0, 100], x_2 \in [0, T_{max}] \quad (1)$$

The task is to construct a mapping $F: X \rightarrow y^*$, where y^* is the numerical index of the recommended complexity or level of subsequent action. Unlike rigid classification, F is formed using fuzzy sets, expert rules, and defuzzification. After calculating y^*, the result is passed to the XAI and LLM modules, which respectively explain the solution and generate adaptive content.

$$y^* = D(Agg(\{R_k(\Phi(X))\})), k = 1,\ldots, K \quad (2)$$

Here Φ is the fuzzification operator, R_k is the k -th production rule, Agg is the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

aggregation of activated output sets, and D is the defuzzification operator.

Input parameter fuzzification

For the input x_1 «Result», three terms are used: Low, Medium, and High; for x_2 «Time», the terms are Fast, Normal, and Slow. For the output variable y «Recommended complexity», the terms Easy, Medium, and Hard are introduced. To ensure computational simplicity and interpretability, triangular membership functions are used. (Zimmermann , H.-J. (2001); Ross , T.J. (2010)) .

$$\mu(x; a, b, c) = \max(\min((x - a)/(b - a), (c - x)/(c - b)), 0) \quad (3)$$

The parameters a, b, and c are specified by the expert and can be recalibrated for a specific subject area. This choice allows us to clearly trace why the same numerical value simultaneously belongs to two adjacent linguistic categories with different degrees.

Table 1. Linguistic variables hybrid models

Variable	Universe	Baths	Purpose
x_1 is the result	0–100	Low, Medium, High	Grade achieved result
x_2 — time	0–T_{max}	Fast, Normal, Slow	Grade speed acceptance solutions
y — complexity	0–100	Easy, Medium, Hard	Choice next level / action

fuzzification result is represented as a vector of membership degrees. For example, for the result x_1:

$$\Phi_1(x_1) = [\mu\text{Low}(x_1), \mu\text{Medium}(x_1), \mu\text{High}(x_1)] \quad (4)$$

and for time x_2:

$$\Phi_2(x_2) = [\mu\text{Fast}(x_2), \mu\text{Normal}(x_2), \mu\text{Slow}(x_2)] \quad (5)$$

The production rule base and Mamdani's derivation

With three terms for each of the two input variables, the complete database contains nine rules. The rules formalize the expert judgment that high performance should increase the recommended complexity, but excessively long execution times may bias the solution toward the average level. Thus, the model takes into account not only the outcome but also the trajectory of its achievement. (Mamdani et al., 1975; Kholifaturrahman et al., 2023).

The value y * is the mathematical solution of the model. Crucially, LLM does not modify y *: the generative module receives the already calculated index and user context, and then generates a new task, recommendation, or explanatory text.

Fuzzy + XAI + LLM Hybrid Architecture

The developed architecture implements a sequential but functionally separated pipeline. The first layer receives numerical data and performs fuzzification. The second layer applies the Mamdani rule base and calculates y *. The third layer generates an

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

explanation: the user is shown membership degrees, activated rules, and the resulting index. The fourth layer communicates the structured context of the generative model via REST. API. This separation reduces the risk that the stochastic response of the LLM will influence the core of the solution. (Barredo et al., 2020; Adadi et al., 2018; Guidotti et al., 2018).

Mamdani Rule Base

No.	Result x_1	Time x_2	Output y
1	Low	Fast	Easy
2	Low	Normal	Easy
3	Low	Slow	Easy
4	Medium	Fast	Hard
5	Medium	Normal	Medium
6	Medium	Slow	Medium
7	High	Fast	Hard
8	High	Normal	Hard
9	High	Slow	Medium

The activation degree of the k -th rule is determined by the minimum operator:

$$\alpha_k = \min (\mu A_k(x_1), \mu B_k(x_2)) \quad (6)$$

The output sets are truncated at level α_k and combined by the maximum operator:

$$\mu agg (y) = \max_k \min (\alpha_k, \mu C_k(y)) \quad (7)$$

To obtain a unique numerical solution, the center of gravity method is used:

$$y^* = (\int y \cdot \mu agg (y) dy) / (\int \mu agg (y) dy) \quad (8)$$

Input data	→	**Fuzzy Engine**	→	**XAI explanation**	→	**LLM adaptation**

Fig. 1. Functional diagram of the hybrid model

The software implementation is done in Python using Flask as the server layer, scikit - fuzzy for fuzzy calculations, SQLite for storing results, Chart.js for visualization, and Gemini An API as an external generative service. Interaction with the LLM is organized asynchronously: the local fuzzy controller first calculates a solution, and then the network request receives the adaptive text. If the external API is unavailable, the mathematical circuit remains operational. (Vaswani et al., 2017; Brown et al., 2020).

Experimental evaluation methodology

The evaluation was conducted according to three groups of criteria: (1) consistency of the system's solution with expert assessment; (2) delay of the full response

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

cycle; (3) explainability and the ability to generate adaptive feedback. Three approaches were used for the comparative analysis: a fuzzy controller without a generative module (Fuzzy Only), Random classifier Forest and the proposed Fuzzy + LLM + XAI hybrid architecture. (Doshi et al., 2017; Breiman et al., 2001).

Two application scenarios were also analyzed. The HR scenario simulated the evaluation of 20 candidates based on test results and completion time. In the educational scenario, pilot testing was conducted on a group of 15 students, where the system's solution was compared with the instructor's subjective assessment. These samples are pilot samples, so the results are interpreted as confirmation of their functionality and potential, rather than as definitive statistical generalization.

Comparative accuracy and latency

demonstrated the highest level of agreement with expert assessment. Only provided 87.3% compliance, Random Forest - 91.2%, and the proposed system - 95.4%. The increase compared to the pure fuzzy model was 8.1 percentage points, and compared to the Random classifier Forest - 4.2 percentage points.

Table 3. Comparison of approaches to decision making

Approach	Consistency with the expert	Average delay	Explainability	Adaptive generation
Fuzzy Only	87.3%	≈120 ms	High	No
Random Forest	91.2%	Low	Limited	No
Fuzzy + LLM + XAI	95.4%	1.5–2.0 s	High	Yes

The increased response time in the hybrid system is primarily due to the network call to the LLM. Fuzzy inference is generated significantly faster, and the generative query is executed asynchronously. In the user interface, the mathematical solution and its XAI justification can be displayed before the rich text generation is complete, reducing the subjectively perceived latency.

Results of application scenarios

In the HR scenario, the system demonstrated a 94% compliance rate with the specialist's decisions when evaluating 20 candidates. At the same time, the time spent generating personalized written feedback was reduced by 78%. This demonstrates that the generative block can be effectively considered as a means of automating explanations, while the final category retains mathematical traceability.

In a pilot educational scenario involving 15 students, the match between the calculated cognitive profile and the instructor's assessment was 91.3%. An important observation was the model's correct behavior in borderline situations: a high score with an excessively long time did not always lead to the maximum difficulty level, as rules that take into account the speed of execution were activated. These are precisely the cases that are poorly described by binary logic.

Analysis stability solutions

A fuzzy model is particularly useful in transition regions, where a small change in an input parameter should not lead to a sudden change in the solution. When adjacent membership functions intersect, several rules are simultaneously activated, and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the resulting index y * changes smoothly. This property reduces sensitivity to small variations in the result or time, making the system's behavior more predictable for the user.

Table 4. Results of applied testing

Scenario	Volume pilot samples	Basic indicator	Result
HR assessment	20 candidates	Coincidence with expert assessment	94%
Adaptive testing	15 students	Matches the estimate teacher	91.3%
General validation controller	Test scenarios	Correlation with experts expectations	>95%
Full cycle systems	Series requests	Time answer	1.5–2.0 s

The architecture also provides for degradation without failure: if the generative API is unavailable, Fuzzy The Engine and XAI continue to function. In this case, the system retains the ability to calculate y * and display activated rules, but temporarily does not generate new personalized content. This separation increases reliability compared to an architecture in which the LLM is responsible for both the decision and its textual representation.

Results and discussion.

The obtained results show that the combination of fuzzy logic and generative AI is feasible provided that the roles of the components are clearly delineated. The mathematical circuit is deterministic and interpretable, while LLM provides semantic flexibility. This approach is consistent with the modern direction of XAI, where explainability is considered a necessary characteristic of intelligent systems used in human-centered tasks (Barredo et al., 2020; Adadi et al., 2018; Guidotti et al., 2018). The theoretical basis of the fuzzy contour is based on classical works on fuzzy logic (Zadeh, 1965; Mamdani et al., 1975), and the generative layer is based on the Transformer architecture and LLM research (Vaswani et al., 2017; Brown et al., 2020; Wang et al., 2024).

Advantage of the hybrid model over Random Forest in the comparison conducted consists not only of higher consistency, but also of the presence of an explicit base of causally interpretable rules. For random forest, post -hoc explanation methods can be applied, for example, LIME or SHAP (Ribeiro et al., 2016; Lundberg et al., 2017), however, this is already a separate layer of approximation of the model's behavior. In the proposed approach, the decision logic is accessible directly through membership functions and production rules.

However, the pilot results cannot be interpreted as universal across all subject areas. The membership function parameters and rule base reflect expert knowledge and require recalibration when transferring to another domain. Furthermore, the number of participants in the applied scenarios is limited. The next stage should include extensive validation on larger samples, statistical analysis of confidence intervals, and comparison with modern neural network and neural-symbolic approaches.

Another limitation is related to the external LLM service. A latency of 1.5–2.0 seconds is acceptable for educational and analytical scenarios, but may be excessive for

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

hard real-time systems. In such applications, the generative layer should remain outside the critical control loop. A promising approach is to use local LLMs, which will reduce network dependency and allow for better data privacy control.

Practical significance

The proposed model can be adapted to three classes of tasks. In educational systems, it allows for the selection of the next task's difficulty based not only on the correctness of the answer but also on the completion time. In HR systems, the model can support expert assessment of candidates and automate the generation of personalized feedback. In business analytics, a similar framework can be used to work with linguistic categories of risk, effectiveness, and urgency, provided the corresponding membership functions and rules are generated by subject-matter experts.

A practically important property is modularity. Fuzzy The Engine can be replaced or expanded without changing the user interface; LLM can be switched to another service or local model; XAI visualization can be supplemented with new graphs. This reduces technological dependency and facilitates the system's integration into existing digital infrastructure.

Extended formalization of the hybrid model

the interactions between the components of the proposed system, we introduce four functional operators: F_f—the fuzzy inference operator, F_x — the explanation generation operator, F_g — the generative operator, and F_c — the context checking operator. The input vector X is first transformed by the fuzzy controller into a numerical solution y * and a set of diagnostic features Z containing membership degrees and activated rules. At this stage, the solution is completely determined by the parameters of the membership functions and the rule base and is independent of stochastic text generation. This ordering is fundamentally important for reproducibility: identical inputs and constant parameters F_f yield the same y *.

$$F\_f: X \rightarrow (y^*, Z), Z = \{\mu_i(x_j), \alpha, C\} \quad (9)$$

The XAI module is viewed as an interpretative mapping that translates the internal state of a fuzzy inference into an explainable representation E. E includes the dominant terms of the input variables, the rules with the highest activation degree, the resulting index y *, and a verbal description of the reasons for the level selection. Since E is constructed from Z, the explanation is directly linked to the mathematical inference mechanism rather than generated independently of it.

$$F\_x : (y^*, Z) \rightarrow E \quad (10)$$

The generative module receives not arbitrary source data, but a structured context Q formed from the fuzzy inference result, the XAI explanation, and the domain constraints. For an educational scenario, Q may include the recommended difficulty level, the type of error made, a constraint on the format of the next task, and a brief

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

history of previous attempts. For an HR scenario, the context includes the confidence level, the required feedback format, and a prohibition on changing the final numerical solution.

$$Q = G\_c (y^*, E, H, R), S = F\_g (Q) \quad (11)$$

Here, H denotes the available interaction context, R is the set of subject constraints, and S is the text generated by the LLM. The F_c operator performs the final check: the text must not contain statements that contradict y *, and the recommended category must match the result of the mathematical contour. Therefore, the final system response is determined not by the LLM alone, but by the coordinated composition of modules.

$$A = F\_c (y^*, E , S), A \in \Omega (y^*) \quad (12)$$

The set Ω (y *) defines the range of acceptable answers for a specific solution. For example, for y *, which corresponds to a medium level of complexity, the generative module can change the task formulation, context, subject vocabulary, and explanation style, but should not automatically transition the user to a higher level. This principle ensures the separation of computational and communicative functions and reduces the risk of uncontrolled influence of the generative model on the solution.

Algorithm of the system's operation

The hybrid model's algorithm is implemented as a sequence of a local mathematical loop and an external generative loop. The first stage involves checking the ranges x_1 and x_2 and normalizing the inputs. The second stage calculates the membership degrees of all terms. The activity of each of the nine rules is then determined, and aggregation and defuzzification are performed. The resulting value y * is stored along with the diagnostic vector Z. Only after these operations are completed is the request to the LLM generated. This procedure allows the mathematical solution to be returned to the user even without a network connection to the generative service.

Table 5. Stages of operation of the hybrid model

Stage	Operation	Result
1	Control and normalization of input values	Correct vector X
2	Fuzzification of x_1 and x_2	Degrees accessories μ
3	Activation bases of 9 rules	Coefficients α_k
4	Aggregation and defuzzification	Numerical index y*
5	Formation of XAI representation	Explanation E
6	Preparation structured context	Prompt Q
7	Generation adaptive answer	Text S
8	Examination consistency with y*	Final Answer A

To limit latency, the generative call can be executed asynchronously. The client

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

first receives y * and a brief XAI explanation, after which the adaptive text is loaded in a separate request. This architectural decision is important for applications where decision reliability is more important than the value of extended textual support. Furthermore, logging the intermediate values μ and α_k allows for retrospective replay of the decision and determination of which rule became the determining one.

Numerical example of fuzzy inference

Consider an illustrative case where a user scored 78 points out of 100 and spent time between the terms Normal and Slow. Unlike a hard threshold, which would immediately assign 78 points to a single category, the fuzzy model allows a result to simultaneously belong to both Medium and High. Similarly, a time can partially belong to two adjacent terms. Suppose that after fuzzification, the obtained values are μMedium (x_1) = 0.35, μHigh (x_1) = 0.65, μNormal (x_2) = 0.60, and μSlow (x_2) = 0.40. The remaining membership degrees in this example are equal to zero.

Four combinations of rules become active. For the High – Normal pair , the activation degree is min (0.65; 0.60) = 0.60; for High – Slow , it is min (0.65; 0.40) = 0.40; for Medium – Normal , it is min (0.35; 0.60) = 0.35; for Medium – Slow , it is min (0.35; 0.40) = 0.35. Each rule will truncate the corresponding output set at its level. If the High – Normal and Medium – Normal rules support more complex continuation, while the Slow rules reduce the recommendation, the resulting aggregate function represents a trade-off between high performance and increased execution time.

$$\alpha_1 = 0, 60; \ \alpha_2 = 0.40; \alpha_3 = 0.35; \alpha_4 = 0.35 \quad (13)$$

After defuzzification, a continuous value y * is calculated, which is then interpreted relative to the Easy, Medium, and Hard domains. The specific set of numbers in the illustration is not important, but the mechanism: the transition between categories occurs smoothly. A user with similar scores of 77 and 78 does not receive fundamentally different decisions simply because they cross an artificial threshold. The recommendation changes gradually as the membership levels change, the XAI layer can generate an explanation like this: «The result is predominantly high-level, but the execution time is partially slow; therefore, the system recommends the next level without maximizing the complexity.» The LLM receives this structured conclusion and generates the appropriate task or feedback. Thus, the natural language component concretizes the solution but does not replace the inference procedure.

Sensitivity and robustness analysis

The stability of a hybrid model is determined primarily by the geometry of the membership functions and the consistency of the rule base. Sensitivity analysis considers a change in one input parameter while the other is fixed. If x_2 remains constant, a smooth change in x_1 should cause a smooth change in y *. Several rules are active in the intersection region of adjacent terms, so the final decision surface is formed as a weighted union of them. Sharp jumps are possible only with incorrectly defined membership functions or inconsistent rules.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Local sensitivity can be estimated by the finite difference:

$$S_1 \approx | F (x_1 + \Delta x , x_2) - F (x_1, x_2)| / | \Delta x | \quad (14)$$

S_2 is defined similarly for time. In a practical setting, it is advisable to monitor the maximum values of S_1 and S_2 on the input space grid. If excessively strong variations in y * are observed within a small neighborhood, the expert should check the intersection of the membership functions and the content of the rules. Such diagnostics are especially important when transferring a model from an educational scenario to HR or business analytics, since the scales and meaning of the input variables change.

The second aspect of robustness is related to parametric uncertainty. The parameters of the triangular functions a, b, and c are specified by experts, so it is reasonable to evaluate the result with a small shift in the boundaries. If a change in the boundary of a term by a few percent leads to a radical change in the solution for a large number of observations, the model is overfitted to the chosen labeling. Conversely, maintaining the general structure of the solutions with moderate parameter variations indicates robustness.

The third aspect is the fault tolerance of the architecture. An LLM failure does not disrupt the calculation of y * because Fuzzy The engine runs locally. A failure in XAI visualization also doesn't change the mathematical result, although it reduces the interface's transparency. The most critical issue is corruption of the rule base or an inconsistent version of the membership function parameters. Therefore, the fuzzy controller configuration must be versioned, and each result record must contain the identifier of the model version used.

Table 6. Main sources of uncertainty and methods of control

Source	Influence	Mechanism control
Border values entrances	Partial affiliation two baths	Overlapping functions accessories
Differences expert ratings	Variability rules and boundaries	Knowledge base alignment and calibration
Change subject areas	Bias distributions input data	Re-configuring parameters a, b, c
LLM unavailability	No generative text	Degradation without failure: y * and XAI are preserved
LLM stochasticity	Variability formulations	Structured context and F_c verification

Comparison with alternative approaches

Fuzzy Comparison Only, Random Forest and Fuzzy + LLM + XAI shows that these approaches solve different subproblems. Random Forest works well as a predictive model given a sufficiently labeled dataset, but its decision is derived from the statistical regularities of the training set. A fuzzy model, in contrast, can be built before extensive historical data has accumulated, if expert knowledge is available. This is particularly useful in the early stages of implementing an intelligent system, when data is insufficient for robust machine learning (Breiman. 2001).

A pure fuzzy system provides interpretability and low latency, but its user feedback

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

is limited to pre-prepared templates. Adding LLM expands semantic adaptability: given the same y *, the system can generate different tasks, explanations, or recommendations based on context. However, such flexibility requires control over generation. Therefore, in the proposed architecture , LLM is not allowed to independently determine the final class and is performed after mathematical inference.

In terms of explainability, the hybrid approach differs from the post hoc interpretation of a complex model. In Random Forest explanations are often constructed using additional tools—feature importance, local approximations, or SHAP -like methods. In a fuzzy system, interpretation is embedded directly in the inference engine: terms and rules are human-readable. This makes the explanation causally related to the computation of the result and facilitates expert audit. (Barredo et al., 2020; Adadi et al., 2018; Guidotti et al., 2018).

Table 7. Qualitative comparison architecture

Criterion	Fuzzy Only	Random Forest	Fuzzy + LLM + XAI
Necessity big educational samples	Low	High	Low for Fuzzy- contour
Interpretability solutions	High	Average	High
Adaptive generation text	No	No	Yes
Determinism nuclei	Yes	Yes at fixed models	Yes
Job without external API	Full	Full	Solution persists, generation unavailable
Convenience expert adjustments	High	Required retraining	High For bases rules

Therefore, the advantage of the proposed solution lies not in replacing machine learning methods, but in the rational distribution of functions. Where large datasets and a stable target variable are present, a trainable classifier can be included as an additional module. Where expert interpretation and manual tuning are critical, a fuzzy contour remains preferable. The generative component is used for communication and personalization, rather than as a source of unsupervised management decisions.

Requirements for reproducibility and software implementation

To ensure reproducibility of results, it is necessary to store not only the result y *, but also the full set of parameters for a specific run. These include the values of x_1 and x_2, membership function versions, the rule base version, the activated rules α_k, defuzzification parameters, the program code version, and the generative model identifier. Such logging allows for repeating the local mathematical calculation regardless of subsequent updates to the LLM or interface.

In the server implementation, it makes sense to split the API into two independent routes: / decision for mathematical inference and / generate for adaptive text. The first route should be synchronous and have minimal dependencies on external services; the second can be executed asynchronously. It is recommended to store the computational result, XAI explanation, and LLM text separately in the database. This simplifies auditing and allows for investigating cases where the generative formulation failed despite a valid y*.

Input validation should be performed before fuzzification. Values outside the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

acceptable range should not be automatically truncated without logging, as this conceals data source errors. It is more appropriate to return a diagnostic message or use an explicit normalization procedure. To protect against inconsistent prompts, only a structured set of permitted fields is passed to the LLM query. Additional user comments should be separated from system constraints and not be able to alter the calculated decision level.

As the number of input variables increases, the rule base grows combinatorially. For n inputs and m terms, the number of rules can reach m^n with full combination. Therefore, further scaling requires hierarchical fuzzy systems, rule reduction based on expert selection, or learning the rule structure from data. Within the framework of the model under consideration, two inputs and three terms yield a compact rule base of nine rules, which is fully observable by an expert and convenient for manual verification. (Zimmermann et al., 2001; Ross, 2010).

Research limitations and future directions

The main limitation of the pilot study is the small size of the pilot samples: 20 observations in the HR scenario and 15 in the educational scenario. These data are sufficient to validate the prototype's performance and analyze the behavioral logic, but they do not allow us to claim the universal statistical superiority of the hybrid model. Confirming its generalizability requires a larger sample, independent expert evaluation by several specialists, and an analysis of inter-rater agreement.

The second limitation is related to the manual tuning of membership functions. Expert tuning ensures transparency, but may contain subjective biases. A promising approach is to combine an expertly defined initial form with subsequent optimization of parameters based on accumulated data while maintaining the semantic interpretability of terms. At the same time, automatic optimization should not transform clear membership functions into forms that are difficult to explain.

The third limitation is the dependence of the generative layer on the properties of a specific LLM. Different models may differ in style, latency, context window, and instruction robustness. To eliminate the influence of these factors on the main result, the mathematical circuit is intentionally isolated. Future research would benefit from comparing cloud-based and local language models, assessing the stability of formulations, and introducing automatic metrics for text consistency with y *.

Finally, a hybrid model should be evaluated not only for accuracy but also for the quality of the user experience. For educational systems, clarity of explanations, usefulness of the next task, and the absence of excessive complexity are important. For an HR scenario, special constraints are necessary to prevent the use of sensitive features and incorrect conclusions about personality. Therefore, further verification should include technical, expert, and user metrics.

Procedure of expert verification of the model

Verifying the correctness of a fuzzy system should include not only comparing the resulting y * with the expert assessment, but also auditing the internal logic. In the first stage, the expert considers the membership functions separately from the rule base and evaluates whether the term boundaries correspond to the substantive meaning of the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

subject area. For the «Result» variable, it is necessary to ensure that the Low, Medium, and High zones overlap sufficiently for a smooth transition, but not so much that a large part of the scale simultaneously has high membership in all terms. A similar check is performed for the «Time» variable. (Mamdani et al., 1975; Ross et al., 2010).

The second stage involves a logical audit of the rule table. Each rule is checked for consistency with adjacent combinations. For example, given a fixed high result, switching from Fast to Normal and Slow should not unreasonably increase the recommended complexity. If two adjacent rules specify opposite outcomes without a substantive basis, the solution surface may contain a local discontinuity. Therefore, expert verification should consider not individual rules, but their system and their monotony across semantic dimensions.

The third stage involves testing checkpoints. A set of typical, borderline, and obviously extreme combinations of X is generated. For each combination, the expert first independently records the expected solution, after which the result is compared with the Fuzzy output. Engine. Particular attention is paid to the intersection points of the membership functions, as this is where the advantage of the fuzzy approach over hard threshold classification becomes apparent. Discrepancies at such points do not always indicate an error: they may indicate the need to refine the expert scale itself.

The fourth step involves checking explainability. For each checkpoint, the XAI module must display the same dominant terms and rules that were actually used to calculate y *. A situation where the interface displays a user-friendly text explanation that does not correspond to the internal coefficients α_k should not be allowed. Therefore, the XAI representation should be formed directly from the logged Fuzzy values. Engine, rather than generating it anew using the language model.

The fifth stage checks the semantic consistency of the generated response. The expert evaluates whether the LLM maintains the calculated level, whether it introduces additional unsupported conclusions, and whether the text complies with the established subject restrictions. To automate the verification, the category specified in the response can be compared with the y * category, and a set of prohibited or mandatory statements can be applied. If a contradiction is detected, the text is rejected and regenerated, while the mathematical result remains unchanged.

This multi-level verification procedure makes model evaluation more meaningful than a single accuracy metric. It allows for the separate identification of parameterization errors, knowledge base errors, XAI inconsistencies, and generative layer issues. As a result, the experimental evaluation becomes reproducible: the researcher can indicate not only the match percentage but also the specific system level at which the discrepancy arose. For further large-scale testing, this separation of errors appears to be the most useful.

To quantify such a check, a discrepancy matrix can be used, in which the rows correspond to the test cases, and the columns correspond to the individual levels of the hybrid architecture: input data, fuzzy inference, XAI, and LLM. Each cell records the presence or absence of an error. This format allows one to distinguish between a

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

fundamental decision error and an explanation error. For example, if y * matches the expert assessment, but the LLM text incorrectly describes the cause, this should not reduce the Fuzzy's mathematical accuracy. Engine; the error is related to the generative layer and should be analyzed separately.

It's also advisable to evaluate stability during repeated text generation. For the same constant y *, the language model may produce different formulations, which is acceptable as long as the recommended level, meaning of the explanation, and subject-matter constraints are maintained. Therefore, lexical variability is not considered an error in itself. The criterion is the semantic invariance of key statements. This approach allows for the creative potential of generative AI to be harnessed without sacrificing controllability of the solution.

Thus, the proposed verification scheme is consistent with the general idea of a hybrid architecture: each component is assessed based on the function it actually performs. The fuzzy loop is assessed for the correctness and robustness of its solution, the XAI for the accuracy of its internal logic representation, and the LLM for the usefulness and consistency of the generated text. The combined assessment of these levels provides a more comprehensive picture of the system's quality than a single aggregated metric.

Conclusion.

This paper develops a hybrid mathematical model for decision making under uncertainty, combining methods of fuzzy logic, explainable artificial intelligence, and generative large-scale language models. The model is based on sequential processing of input information, including parameter fuzzification, membership degree generation, application of Mamdani's nine production rules, and defuzzification using the center-of-gravity method. This approach enables the formalization of expert knowledge and allows for situations in which the initial data is incomplete, ambiguous, and lacks clear boundaries between decision classes.

A key feature of the proposed architecture is the functional separation of the system's computational and intelligent components. The fuzzy controller is directly responsible for generating a quantitative solution (y^*) based on a given rule base and membership functions. This ensures that the final solution is independent of probabilistic text generation and maintains mathematical reproducibility. The XAI module interprets the generated solution by displaying the membership degrees of the input parameters, activated production rules, and their influence on the final result. The generative LLM module is used at a subsequent stage and performs semantic adaptation of the results, generating text explanations, recommendations, and content tailored to the context of a specific subject area.

The proposed separation of functions eliminates one of the significant problems associated with using large language models in decision support systems: the lack of guaranteed mathematical traceability of the generated response. In the developed model, the LLM does not replace the decision-making mechanism, but rather serves as an intelligent interpretive layer. Therefore, the basic decision remains defined by the formalized fuzzy model, while the generative component expands the system's user

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

interaction capabilities and improves the accessibility of results for specialists without in-depth knowledge of fuzzy mathematics.

An experimental evaluation of the developed approach demonstrated a relatively high degree of consistency between the hybrid system's results and expert decisions. The resulting consistency rate was 95.4%, exceeding the corresponding results of the Fuzzy Only model, which achieved 87.3%, and the Random algorithm. Forest with a score of 91.2%. This demonstrates that combining fuzzy inference with intelligent explanation and contextual adaptation mechanisms improves the quality of the final solution compared to using individual methods.

Additional validation of the model in two application scenarios also confirmed its feasibility. In the HR scenario, related to candidate assessment and personnel decision support, the system's results matched expert assessments to 94%. In the pilot educational scenario, focused on adaptive user assessment, the figure was 91.3%. These results demonstrate that the developed architecture can be transferred across different subject areas with appropriate adaptation of the input variables, membership functions, rule base, and text context of the LLM module.

From a practical perspective, the system's acceptable computational performance is also a significant result. The complete query processing cycle, including fuzzy inference, XAI explanation generation, and text interpretation generation, averaged 1.5–2.0 seconds. This performance allows the developed approach to be considered as a basis for interactive decision support systems, adaptive educational platforms, HR systems, expert interfaces, and other applications that require rapid, interpretable results.

The scientific significance of the proposed approach lies in the formation of a unified architecture in which symbolic expert knowledge and modern generative models are not opposed but rather used as complementary components. Fuzzy logic provides formalized uncertainty processing, XAI provides transparency in the decision-making mechanism, and LLM provides natural-language interpretation and contextual adaptation. This combination allows for increased explainability of an intelligent system without sacrificing the capabilities of modern generative technologies.

However, the study has several limitations. The current version of the model uses a fixed set of membership functions and a limited database of production rules, generated based on expert knowledge. This may reduce the system's adaptability when the input data distribution significantly changes or when the model is transferred to a new domain. Furthermore, the experimental validation was performed on limited samples, requiring further verification of the obtained results on larger, more heterogeneous datasets.

A promising direction for further research is the automatic tuning of membership function parameters based on accumulated statistics and optimization methods. It also seems worthwhile to explore the possibility of automated expansion and adjustment of the production rule base while maintaining expert control over their semantic correctness. Of particular interest is the use of local large-scale language models, which can reduce the system's dependence on external cloud services and increase the confidentiality of processed data.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Future research would also benefit from a more detailed analysis of the model's robustness to inconsistent and incomplete input data, assessing the impact of various defuzzification strategies on the final result, and comparing the proposed approach with other hybrid architectures based on neuro-fuzzy models, gradient boosting, and modern ensemble machine learning methods. An additional avenue involves automatically assessing the quality of LLM explanations and their compliance with the actual activated rules of the fuzzy model.

Thus, the developed hybrid mathematical model demonstrates the potential for effectively combining formalized decision-making methods and modern generative technologies. The obtained results confirm the potential of using the Fuzzy + XAI + LLM architecture for building intelligent systems that simultaneously require robustness to uncertainty, explainability, mathematical traceability, and natural-language user interaction.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. — Vol. 6. — Pp. 52138–52160. DOI: https://doi.org/10.1109/ACCESS.2018.2870052. [in Eng.].

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems. — Vol. 33. — Pp. 1877–1901. [in Eng.].

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. Pp. 58, 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. [in Eng.].

Breiman, L. (2001). Random forests. Machine Learning. — Vol. 45. — Pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324. [in Eng.].

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. DOI: https://doi.org/10.48550/arXiv.1702.08608. [in Eng.].

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys. — Vol. 51. —No. 5. Pp. 1–42. DOI: https://doi.org/10.1145/3236009. [in Eng.].

Kholifaturrahman, M., Rachmatullah, S., & Said, B. (2023). Decision Support System for Single Tuition Scholarship Awardees in Higher Education Using Mamdani Fuzzy Inference. Journal of Information Technology and Cyber Security. — Vol. 1. — No. 2. Pp. 85–97. DOI: https://doi.org/10.30996/jitcs.10009. [in Eng.].

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. — Vol. 30. — Pp. 4765–4774. [in Eng.].

Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies. — Vol. 7. — No. 1. Pp. 1–13. DOI: https://doi.org/10.1016/S0020-7373(75)80002-2. [in Eng.].

Ross, T. J. (2010). Fuzzy Logic with Engineering Applications (3rd ed.). John Wiley & Sons. DOI: https://doi.org/10.1002/9781119994374. [in Eng.].

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pp. 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778. [in Eng.].

Zadeh, L. A. (1965). Fuzzy sets. Information and Control. — Vol. 8. — No. 3. Pp. 338–353. DOI: https://doi.org/10.1016/S0019-9958(65)90241-X. [in Eng.].

Zimmermann, H. J. (2001). Fuzzy Set Theory—and Its Applications (4th ed.). Springer. DOI: https://doi.org/10.1007/978-94-010-0646-0. [in Eng.].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems. — Vol. 30. — Pp. 5998–6008.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

[in Eng.].

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). Large language models for education: A survey and outlook. arXiv:2403.18105. DOI: https://doi.org/10.48550/arXiv.2403.18105. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 186–198
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.011
УДК 004.931

DEVELOPMENT OF A HYBRID DEEP LEARNING MODEL FOR MULTI-CLASS CLASSIFICATION OF MICROSCOPIC IMAGES OF BACTERIA

A. Ismailova[1], G. Yessenbayeva[1*], K. Kadyrkulov[2], R. Moldasheva[3], A. Amangeldi[3]
[1]S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan;
[2]Smart Soft Kazakhstan LLP, Astana, Kazakhstan;
[3]Atyrau University named after H. Dosmukhamedov, Atyrau, Kazakhstan.
E-mail: gulbanu210596@gmail.com

Aisulu Ismailova — PhD, Associate Professor, S. Seifullin Kazakh AgroTechnical Research University, Astana, Kazakhstan
https://orcid.org/0000-0002-8958-1846;
Gulbanu Yessenbayeva — Doctoral student, Department of Information Systems, S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan
E-mail: gulbanu210596@gmail.com, https://orcid.org/0009-0006-6371-4571;
Kuanysh Kadyrkulov — PhD, Director of «Smart Soft Kazakhstan» LLP, Astana, Kazakhstan
https://orcid.org/0000-0003-0506-4890;
Raushan Moldasheva — PhD, Acting Associate Professor, Department of Software Engineering, Atyrau University named after H. Dosmukhamedov, Atyrau, Kazakhstan
https://orcid.org/0000-0002-4570-0487;
Ardak Amangeldi — Senior Lecturer, Department of Software Engineering, Atyrau University named after H. Dosmukhamedov, Atyrau, Kazakhstan
https://orcid.org/0009-0000-5889-0593.

© A. Ismailova , G. Yessenbayeva, K. Kadyrkulov, R. Moldasheva, A. Amangeldi

Abstract. This study investigates the problem of automatic multiclass classification of microscopic bacterial images. The experimental dataset consists of 2034 microscopy images covering 33 bacterial taxa. To ensure methodological reliability, the dataset was carefully verified and divided into independent training, validation, and test subsets using a strict protocol designed to eliminate potential information leakage. To describe image quality and structural characteristics, a set of quantitative proxy features was extracted, including brightness, contrast, Shannon entropy, Laplacian variance, and Sobel gradient energy. The discriminative ability of these features across bacterial classes was assessed using the Kruskal–Wallis statistical test, which confirmed

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

significant inter-class differences. Classification performance was evaluated using both conventional machine learning algorithms and modern deep learning architectures. Furthermore, a hybrid deep learning framework based on multiple instance learning was developed to aggregate local structural patterns within microscopic images more effectively. Experimental results demonstrate that the proposed methodology enhances classification accuracy and improves robustness across diverse bacterial taxa.

Keywords: microscopic images, bacterial classification, multiclass classification, deep learning, multiple instance learning, hybrid model

Conflict of interest: A. Ismailova, G. Yessenbayeva, K. Kadyrkulov, R. Moldasheva, A. Amangeldi (2026). Development of a hybrid deep learning model for multiclass classification of microscopic images of bacteria // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 186–198. https://doi.org/10.54309/IJICT.2026.25.1.011. (In Kaz.).

Conflict of interest: The authors declare that there is no conflict of interest.

БАКТЕРИЯЛАРДЫҢ МИКРОСКОПИЯЛЫҚ БЕЙНЕЛЕРІН КӨПКЛАССТЫ ЖІКТЕУГЕ АРНАЛҒАН ГИБРИДТІ ТЕРЕҢ ОҚЫТУ МОДЕЛІН ӘЗІРЛЕУ

А.А.Исмаилова[1], Г.Р.Есенбаева[1*], К.К.Кадиркулов[2], Р.Н.Молдашева[3], А.Амангелді[3]

[1]С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан;

[2]Smart Soft Kazakhstan ЖШС, Астана, Қазақстан;

[3]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан.

E-mail: gulbanu210596@gmail.com

Исмаилова Айсулу Абжаппаровна — PhD, қауымдастырылған профессор, С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан
https://orcid.org/0000-0002-8958-1846;

Есенбаева Гүлбану Рақымжанқызы — «Ақпараттық жүйелер» кафедрасының докторанты, С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан
E-mail: gulbanu210596@gmail.com, https://orcid.org/0009-0006-6371-4571;

Кадиркулов Куаныш Кайсарович — PhD, «Smart Soft Kazakhstan» ЖШС директоры**.** Астана, Қазақстан
https://orcid.org/0000-0003-0506-4890;

Молдашева Раушан Нуркожаевна — PhD, «Бағдарламалық инженерия» кафедрасының қауымдастырылған профессоры м.а., Х.Досмұхамедов атындағы Атырау университеті, Атырау қ., Қазақстан
https://orcid.org/0000-0002-4570-0487;

Амангелді Ардақ Амангелдіқызы — «Бағдарламалық инженерия»

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

кафедрасының аға оқытушысы, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0009-0000-5889-0593.

© А.А. Исмаилова, Г.Р. Есенбаева, К.К. Кадиркулов, Р.Н. Молдашева, А. Амангелді

Аннотация. Бұл зерттеу микроскопиялық бактерия суреттерін автоматты түрде көпкластық жіктеу мәселесіне арналған. Жұмыста 33 бактериялық таксонды қамтитын 2034 микроскопиялық кескіннен тұратын деректер жиыны пайдаланылды. Эксперименттің дұрыстығын қамтамасыз ету үшін деректердің тұтастығы тексеріліп, ақпараттың ағып кетуін болдырмау мақсатында оқыту, валидация және тест жиынтықтарына қатаң түрде бөлінді. Кескіндердің сапалық және құрылымдық сипаттамаларын сандық тұрғыдан бағалау үшін бірқатар көрсеткіштер есептелді: жарықтылық, контраст, Шеннон энтропиясы, Лапласиан дисперсиясы және Собель градиентінің энергиясы. Бұл белгілердің кластар арасындағы айырмашылық қабілеті Краскел–Уоллис критерийі арқылы талданып, таксондар арасында статистикалық тұрғыдан мәнді айырмашылықтардың бар екені анықталды. Жіктеу сапасы дәстүрлі машиналық оқыту алгоритмдерімен қатар заманауи терең оқыту архитектураларының көмегімен бағаланды. Сонымен қатар микроскопиялық кескіндердегі жергілікті құрылымдық ерекшеліктерді тиімді біріктіруге мүмкіндік беретін multiple instance learning әдісіне негізделген гибридті модель ұсынылды. Эксперимент нәтижелері ұсынылған тәсілдің жіктеу дәлдігін арттырып, нәтижелердің тұрақтылығын жақсартатынын көрсетті.

Түйін сөздер: микроскопиялық бейнелер, бактерияларды жіктеу, көпкласты классификация, терең оқыту, көпинстанстық оқыту, гибридті модель

Дәйексөздер үшін: А.А. Исмаилова, Г.Р. Есенбаева, К.К. Кадиркулов, Р.Н. Молдашева, А. Амангелді (2026). Бактериялардың микроскопиялық бейнелерін көпкласты жіктеуге арналған гибридті терең оқыту моделін әзірлеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. № 25. Б. 186–198. https://doi.org/10.54309/IJICT.2026.25.1.011. (Қаз. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

Алғыс. *Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырады (Грант № AP32721703 Қайталануларды пангеномдық талдау, карталау және когорталарда салыстыруға арналған RepeatAtlas платформасын әзірлеу).*

РАЗРАБОТКА ГИБРИДНОЙ МОДЕЛИ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ МИКРОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ БАКТЕРИЙ

А.А. Исмаилова[1], Г.Р. Есенбаева[1*], К.К. Кадиркулов[2], Р.Н. Молдашева[3], А. Амангелды[3]

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

[1]Казахский агротехнический исследовательский университет имени С. Сейфуллина, Астана, Казахстан;
[2]ТОО «Smart Soft Kazakhstan», Астана, Казахстан;
[3] Атырауский университет им. Х. Досмухамедова, Атырау, Казахстан.
E-mail: gulbanu210596@gmail.com

Исмаилова Айсулу Абжаппаровна — PhD, ассоциированный профессор, Казахский агротехнический исследовательский университет им. С. Сейфуллина, Астана, Казахстан
https://orcid.org/0000-0002-8958-1846;
Есенбаева Гульбану Ракымжановна — докторант кафедры «Информационные системы», Казахский агротехнический исследовательский университет имени С. Сейфуллина, Астана, Казахстан
E-mail: gulbanu210596@gmail.com, https://orcid.org/0009-0006-6371-4571;
Кадиркулов Куаныш Кайсарович — PhD, директор ТОО «Smart Soft Kazakhstan», Астана, Казахстан
https://orcid.org/0000-0003-0506-4890;
Молдашева Раушан Нуркожаевна — PhD, Атырауский университет им. Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0002-4570-0487;
Амангелды Ардак Амангелдиевна — старший преподаватель кафедры «Программная инженерия», Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0009-0000-5889-0593.

© А.А. Исмаилова, Г.Р. Есенбаева, К.К. Кадиркулов, Р.Н. Молдашева, А. Амангелды

Аннотация. Данное исследование посвящено задаче автоматической многоклассовой классификации микроскопических изображений бактерий. В работе использован набор данных, включающий 2034 микроскопических изображения, представляющих 33 таксономические группы бактерий. Для обеспечения корректности эксперимента была проведена проверка целостности данных, а также выполнено строгое разбиение на обучающую, валидационную и тестовую выборки с целью исключения утечки информации. Для количественного описания качества и структурных характеристик изображений были вычислены следующие признаки: яркость, контраст, энтропия Шеннона, дисперсия Лапласиана и энергия градиента Собеля. Их различающая способность между классами была проанализирована с использованием критерия Краскела–Уоллиса, что подтвердило статистически значимые различия между таксонами. Качество классификации оценивалось с применением как классических алгоритмов машинного обучения, так и современных архитектур глубокого обучения. Дополнительно была разработана гибридная модель на основе метода multiple instance learning, позволяющая более эффективно учитывать локальные структурные особенности микроскопических

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

изображений. Полученные экспериментальные результаты свидетельствуют о повышении устойчивости и точности классификации при использовании предложенного подхода.

Ключевые слова: микроскопические изображения, классификация бактерий, многоклассовая классификация, глубокое обучение, множественное экземплярное обучение, гибридная модель

Для цитирования: А.А. Исмаилова, Г.Р. Есенбаева, К.К. Кадиркулов, Р.Н. Молдашева, А. Амангелды (2026). Разработка гибридной модели глубокого обучения для многоклассовой классификации микроскопических изображений бактерий // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 186–198. https://doi.org/10.54309/IJICT.2026.25.1.011.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность. *Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант No AP32721703 Разработка платформы RepeatAtlas для пангеномного анализа, картирования и сравнения повторов в когортах).*

Кіріспе.

Соңғы жылдары микробиологиялық диагностикада цифрлық микроскопия мүмкіндіктерін кеңінен пайдалану үрдісі байқалады. Зертханалық тәжірибеде бактериялардың микроскопиялық бейнелерін талдау ауру қоздырғыштарын анықтау мен олардың морфологиялық ерекшеліктерін сипаттауда маңызды орын алады. Дегенмен дәстүрлі визуалды бағалау әдістері маманның тәжірибесіне тікелей тәуелді болып, нәтижелердің субъективтілігіне және уақыт шығындарының артуына әкелуі мүмкін (Shu және т.б.., 2022; Zhang және т.б.., 2021). Осы себепті микроскопиялық кескіндерді автоматтандырылған өңдеу мен жіктеу әдістерін дамыту ғылыми және практикалық тұрғыдан өзекті бағыттардың біріне айналды. Компьютерлік көру және машиналық оқыту тәсілдері микроскопиялық бейнелерден морфологиялық, құрылымдық және текстуралық сипаттамаларды сандық түрде бөліп алуға мүмкіндік береді. Алғашқы зерттеулерде бактерияларды жіктеу үшін алдын ала есептелетін белгілерге сүйенген классикалық алгоритмдер қолданылды. Олардың қатарында тірек векторлар әдісі, шешім ағаштары және градиенттік бустинг модельдері бар (Pádua және т.б.., 2020; Deng және т.б.., 2022). Бұл тәсілдер белгілі бір жағдайларда қанағаттанарлық нәтиже бергенімен, күрделі визуалды құрылымдарды толық қамтуда және деректердің ішкі алуан түрлілігін ескеруде шектеулерге ие.

Соңғы кезеңде терең оқыту әдістері, әсіресе конволюциялық нейрондық желілер, бейнелерді талдау міндеттерінде кеңінен қолданыла бастады. Мұндай модельдер кескіндерден жоғары деңгейлі белгілерді автоматты түрде үйреніп, медициналық визуализация саласында тиімділігін көрсетті (Esteva және т.б.., 2019; Huang және т.б.., 2023). Алайда микроскопиялық деректерге қатысты бірқа-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

тар ерекшеліктер бар: деректер көлемінің шектеулі болуы, әртүрлі таксондар арасындағы морфологиялық ұқсастық және түсірілім шарттарының өзгермелілігі модельдердің жалпылау қабілетіне әсер етуі мүмкін (Tan және т.б.., 2025; Li және т.б.., 2022). Осы қиындықтарды еңсерудің бір жолы ретінде көпинстанстық оқыту тәсілі қарастырылады. Бұл әдісте бір кескін бірнеше локалдық аймақтардың жиынтығы ретінде ұсынылып, модель маңызды құрылымдық фрагменттерді өздігінен анықтай алады (Ilse және т.б.., 2018; Campanella және т.б.., 2019). Әсіресе микроскопиялық және патологиялық бейнелерде барлық аймақтар бірдей ақпарат бермейтіндіктен, мұндай тәсілдің артықшылығы айқын көрінеді. Сонымен қатар соңғы еңбектерде терең оқыту модельдерін қолмен есептелетін сапалық және текстуралық көрсеткіштермен біріктірудің модель тұрақтылығын арттырып, әртүрлі деректер жиынтықтарына бейімделуін жақсартатыны көрсетілген (Zhang және т.б.., 2023; Deng және т.б.., 2024). Бұл тәсілдер бейненің жалпы құрылымдық ерекшеліктерін де, локалдық сипаттамаларын да қатар ескеруге мүмкіндік береді.

Осы жұмыстың мақсаты – бактериялардың микроскопиялық бейнелерін көпкласты жіктеуге арналған гибридті терең оқыту моделін әзірлеу және оның тиімділігін қатаң ұйымдастырылған эксперименттік протокол негізінде бағалау. Ұсынылған әдіс классикалық машиналық оқыту алгоритмдерімен және заманауи терең нейрондық желілермен салыстырылып, алынған нәтижелердің тұрақтылығы мен практикалық құндылығы жан-жақты талданады.

Әдістер мен материалдар.

Зерттеу барысында бактериялардың микроскопиялық бейнелерін автоматты түрде көпкласты жіктеу міндетін орындау үшін ашық қолжетімді деректер жиыны қолданылды. Деректер базасы 33 бактериялық таксонға тиесілі 2034 RGB форматындағы микроскопиялық кескіннен тұрады. Бейнелер кластар бойынша жүйеленген. Эксперимент нәтижелерінің дұрыстығын қамтамасыз ету үшін барлық файлдардың жарамдылығы тексеріліп, бүлінген немесе оқылмайтын деректер анықталған жоқ. Кластар арасындағы үлестірім шамамен теңгерімді, бұл оқыту барысында айқын дисбаланс қаупін азайтады (Ching және т.б.., 2018). Деректерді бөлу кезінде қатаң әдіснамалық қағидаттар сақталды. Бейнелер оқыту, валидация және тест жиынтықтарына өзара тәуелсіз түрде бөлінді. Әрбір кескін тек бір жиынтыққа ғана енгізілді. Сонымен қатар бір көзден алынған немесе мазмұны ұқсас бейнелердің әртүрлі жиынтықтарға түсуіне жол берілмеді. Мұндай тәсіл ақпараттың ағып кету ықтималдығын төмендетіп, модельдердің жалпылау қабілетін шынайы бағалауға мүмкіндік береді (Varoquaux және т.б.., 2017).

Микроскопиялық кескіндердің сапалық және құрылымдық сипаттамаларын сандық тұрғыдан бағалау үшін бірқатар көрсеткіштер есептелді. Олардың қатарында орташа жарықтылық, контраст, Шеннон энтропиясы, Лапласиан дисперсиясы және Собель операторы негізінде анықталған градиент энергиясы бар. Бұл параметрлер кескіннің айқындығын, текстуралық күрделілігін және визуалды ақпараттың қанықтылығын сипаттайды (Gonzalez & Woods, 2018; Redmon және т.б.., 2016). Белгілердің әртүрлі кластар бойынша айырмашылығын бағалау үшін

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

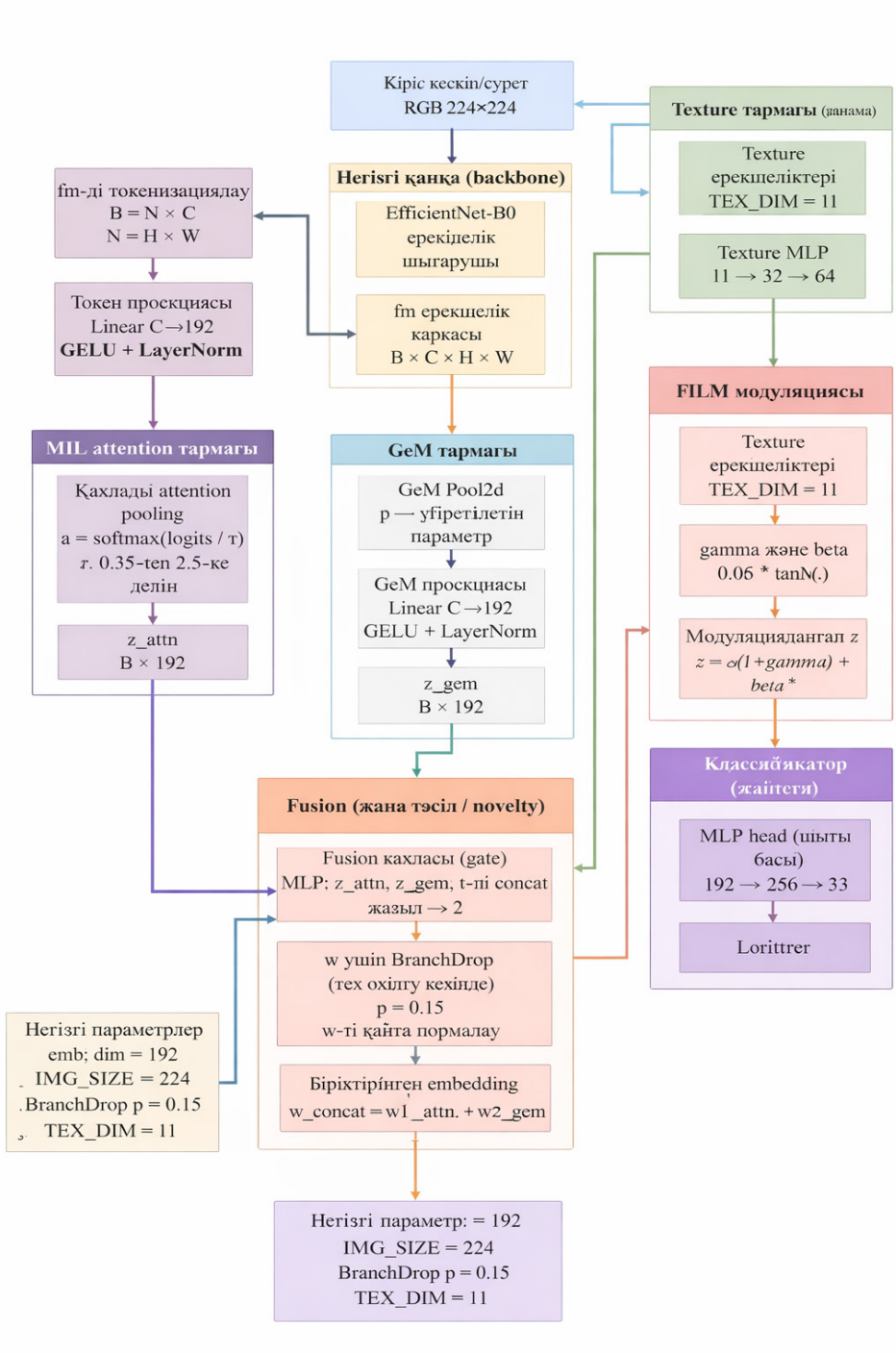
параметрлік емес Краскел–Уоллис критерийі және әсер мөлшерін сипаттайтын ε^2 коэффициенті қолданылды (Tomczak & Tomczak, 2014). Жіктеу нәтижелерін салыстыру мақсатында бірнеше модельдер тобы қарастырылды. Классикалық машиналық оқыту әдістеріне логистикалық регрессия, тірек векторлар әдісі және градиенттік бустинг модельдері енгізілді. Бұл алгоритмдер инженерлік жолмен алынған белгілер негізінде жұмыс істеп, салыстырмалы талдау үшін базалық деңгей ретінде пайдаланылды (Hastie және т.б.., 2009). Терең оқыту тәсілдері ретінде алдын ала үйретілген конволюциялық нейрондық желілер қолданылды, олардың соңғы қабаттары 33 класты жіктеуге бейімделді. Тасымалдап оқыту әдісі деректер көлемі шектеулі жағдайларда модельдің тұрақтылығын арттыруға мүмкіндік береді (He және т.б.., 2016; Tan & Le, 2019).

Микроскопиялық бейнелердегі локалдық құрылымдардың маңызын ескеру үшін көпинстанстық оқыту қағидатына негізделген гибридті модель ұсынылды. Бұл тәсілде бір кескін бірнеше аймақтарға бөлініп, қорытынды шешім осы аймақтар бойынша алынған нәтижелерді біріктіру арқылы қабылданады. Агрегация кезеңінде назар механизмдері пайдаланылып, модельге ақпараттық маңызы жоғары фрагменттерді анықтауға мүмкіндік береді (Ilse және т.б.., 2018; Lu және т.б.., 2021). Сонымен қатар терең нейрондық желілерден алынған белгілер текстуралық прокси-көрсеткіштермен біріктіріліп, модельдің жалпылау қабілетін күшейтуге бағытталды (Perez және т.б.., 2018). Модельдердің тиімділігі көпкласты есептерге сәйкес бірнеше метрика арқылы бағаланды. Олардың қатарында жалпы дәлдік, теңгерімді дәлдік, макро-орташа F1-көрсеткіш және Мэтьюс корреляция коэффициенті бар (Powers, 2020). Бұл метрикалар кластар арасындағы қателіктерді жан-жақты талдауға және ұсынылған тәсілді базалық модельдермен объективті салыстыруға мүмкіндік береді. Ұсынылған гибридті архитектура конволюциялық нейрондық желілердің жоғары деңгейлі белгілерін, текстуралық сипаттамаларды және назарға негізделген агрегация механизмдерін біріктіре отырып, микроскопиялық бейнелердің кеңістіктік біртексіздігін ескеруге және әртүрлі түсірілім жағдайларына бейімделу қабілетін арттыруға бағытталған (Сурет 1).

1-суретте ұсынылған модельдің жалпы архитектурасы көрсетілген. Кіріс ретінде өлшемі 224×224 RGB форматындағы микроскопиялық бейне алынып, ол EfficientNet-B0 негізіндегі конволюциялық backbone арқылы өңделеді. Нәтижесінде алынған белгілер картасы екі параллель тармаққа беріледі:

GeM-пулингке негізделген глобалдық агрегация тармағы және көпинстанстық оқытуға арналған назар механизмі бар MIL тармағы. MIL тармағында белгілер кеңістіктік токендерге түрлендіріліп, gated attention pooling арқылы ақпараттық маңызы жоғары локалдық аймақтар автоматты түрде ерекшеленеді. Сонымен қатар модельде текстуралық прокси-белгілерге негізделген жеке тармақ қарастырылған. Бұл тармақта алдын ала есептелген текстуралық сипаттамалар көпқабатты персептрон арқылы өңделіп, FiLM-модуляция механизмі арқылы негізгі белгілер кеңістігіне енгізіледі. Мұндай модуляция бейненің жарықтану, контраст және текстуралық вариацияларына модельдің сезімталдығын төмендетеді. MIL және

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Сур. 1. Ұсынылған Novel_HybridMIL архитектурасының схемалық диаграммасы.

GeM тармақтарынан алынған белгілер Fusion блогында біріктіріледі. Бұл кезеңде оқыту барысында BranchDrop регуляризациясы қолданылып, модельдің бір тармаққа шамадан тыс тәуелді болу қаупі азайтылады. Қорытынды

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

біріктірілген эмбеддинг көпқабатты классификаторға беріледі, оның шығысында 33 бактериялық класқа сәйкес логиттер есептеледі. Ұсынылған архитектура локалдық және глобалдық ақпаратты үйлесімді түрде біріктіріп, микроскопиялық бейнелерді көпкласты жіктеу дәлдігін арттыруға мүмкіндік береді.

Нәтижелер және оларды талқылау.

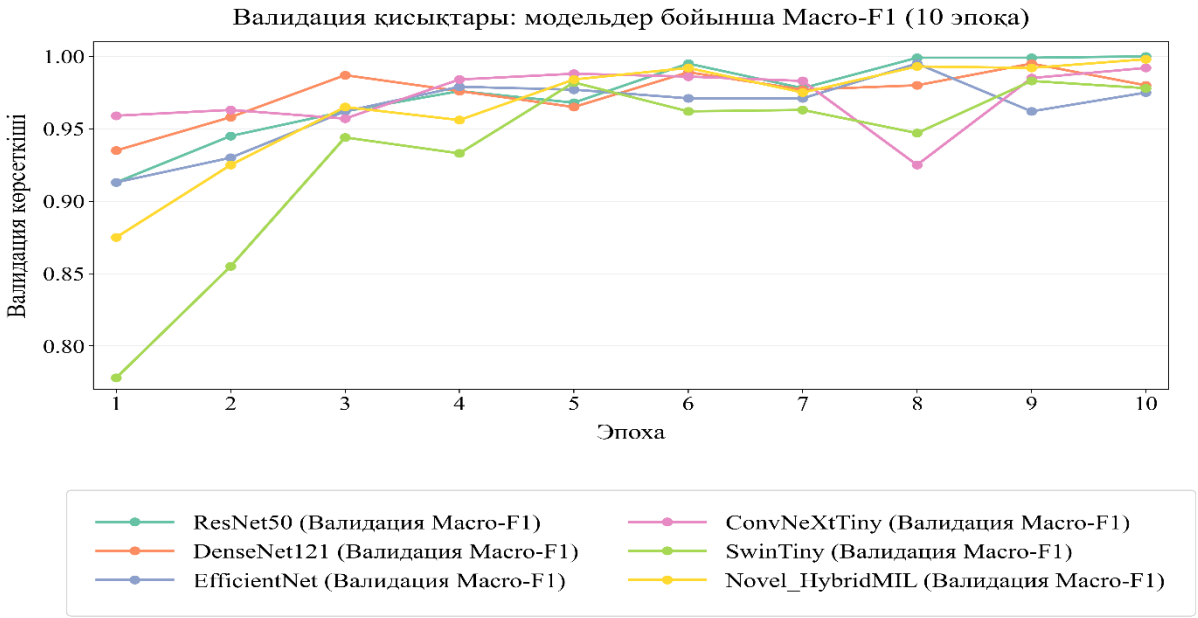
Ұсынылған әдістердің тиімділігін бағалау мақсатында бактериялардың микроскопиялық бейнелерін көпкласты жіктеу бойынша бірқатар эксперименттер жүргізілді. Барлық модельдер бірдей деректер жиынында және алдын ала белгіленген қатаң train/validation/test бөлу протоколы негізінде оқытылып, бағаланды. Мұндай тәсіл алынған нәтижелердің салыстырмалылығын және әдіснамалық дұрыстығын қамтамасыз етеді. Алдымен инженерлік және прокси-белгілерге негізделген классикалық машиналық оқыту модельдерінің нәтижелері талданды. Логистикалық регрессия мен тірек векторлар әдісі базалық деңгейдегі жіктеу сапасын көрсетті, алайда олардың күрделі морфологиялық және текстуралық айырмашылықтарды толық сипаттауда шектеулері байқалды. CatBoost моделі басқа классикалық әдістермен салыстырғанда жоғарырақ нәтижелер көрсетті, бұл табличалық белгілер арасындағы сызықтық емес тәуелділіктерді тиімді моделдеумен түсіндіріледі. Дегенмен, бұл модельдердің барлығы терең оқыту архитектураларынан төмен нәтиже көрсетті, әсіресе локалдық құрылымдардың рөлі жоғары болған кластарда.

Терең оқытуға негізделген модельдер арасында алдын ала үйретілген конволюциялық нейрондық желілер тұрақты және жоғары жіктеу дәлдігін көрсетті. ResNet және DenseNet архитектуралары жақсы нәтижелерге қол жеткізгенімен, кейбір кластарда қателердің сақталуы микроскопиялық бейнелердің ішкі вариабельділігімен және морфологиялық ұқсастығымен байланысты болды. EfficientNet архитектурасы параметрлер саны мен өнімділік арасындағы тиімді теңгерімнің арқасында жалпы метрикалар бойынша бәсекеге қабілетті нәтиже көрсетті. Ал ConvNeXt және Swin Transformer сияқты заманауи архитектуралар күрделі визуалды паттерндерді жақсырақ үйренуге қабілетті екенін көрсетті, алайда олардың артықшылығы барлық кластарда бірдей байқалмады. Ұсынылған гибридті терең оқыту моделі барлық негізгі метрикалар бойынша ең жоғары немесе тұрақты түрде жоғары нәтижелерді көрсетті. Бұл нәтиже модель архитектурасында локалдық және глобалдық ақпаратты бір уақытта ескерудің тиімділігімен түсіндіріледі. Көпинстанстық оқытуға негізделген назар механизмі модельге микроскопиялық бейненің диагностикалық тұрғыдан маңызды аймақтарын автоматты түрде анықтауға мүмкіндік берді, ал GeM-пулинг глобалдық құрылымдық ақпаратты жоғалтпай агрегаттауды қамтамасыз етті. Сонымен қатар, текстуралық прокси-белгілерге негізделген FiLM-модуляция бейненің жарықтану және контраст вариацияларына модельдің сезімталдығын төмендетіп, жалпылау қабілетін арттырды.

2-суретте базалық және заманауи терең оқыту модельдерінің валидациялық таңдамадағы Macro-F1 мәндерінің эпохалар бойынша өзгеруі көрсетілген.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Барлық модельдер алғашқы бірнеше эпохада жылдам конвергенцияны көрсетіп, кейінгі кезеңдерде тұрақты деңгейге жетеді. Ұсынылған гибридті модель валидациялық кезең бойында жоғары және тұрақты Macro-F1 көрсеткішін сақтап, басқа архитектуралармен салыстырғанда жақсы жалпылау қабілетін көрсетеді.



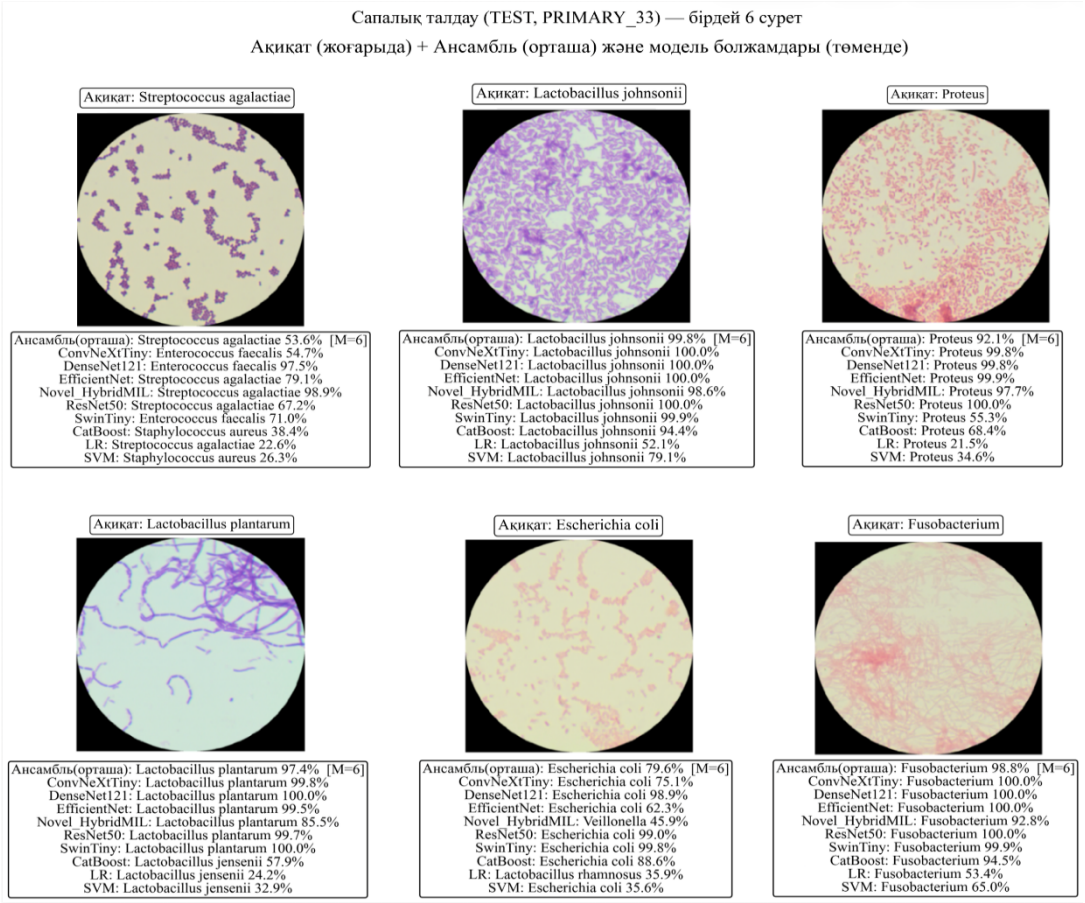
Сур. 2. Әртүрлі модельдер үшін 10 эпоха бойындағы валидациялық деректердегі Macro-F1 көрсеткішінің өзгеру динамикасы

3-суретте тесттік деректер жиынынан таңдап алынған алты микроскопиялық бейне үшін нақты класс белгілері (жоғарыда) және әртүрлі модельдер мен ансамбльдік орташа болжаулардың нәтижелері (төменде) көрсетілген. Көрсетілген мысалдарда ансамбльдік тәсіл мен ұсынылған гибридті модельдің болжамдары нақты кластармен жоғары деңгейде сәйкес келетіні байқалады, бұл локалдық құрылымдарды тиімді ескерудің және модельдердің жалпылау қабілетінің жоғары екенін сапалық тұрғыда растайды.

Қателерді талдау нәтижелері кейбір кластар арасында шатасулардың сақталатынын көрсетті. Бұл, ең алдымен, морфологиялық тұрғыдан ұқсас бактерия түрлеріне тән, олардың микроскопиялық бейнелерінде визуалды айырмашылықтар әлсіз байқалады. Дегенмен, ұсынылған гибридті модель мұндай жағдайларда да базалық және заманауи архитектуралармен салыстырғанда қателер санын азайта алды, бұл локалдық құрылымдарды тиімді іріктеудің артықшылығын көрсетеді.

Жалпы алғанда, алынған нәтижелер ұсынылған гибридті тәсілдің микроскопиялық бейнелерді көпклассты жіктеу міндетінде тиімді екенін дәлелдейді. Модельдің артықшылығы тек жоғары сандық көрсеткіштермен ғана емес, сонымен қатар оның тұрақтылығы мен әртүрлі визуалды шарттарға бейімделу қабілетімен де сипатталады. Бұл зерттеу нәтижелері микробиологиялық

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Сур. 3. Микроскопиялық бейнелер үшін модельдердің сапалық салыстырмалы нәтижелері

диагностикада жасанды интеллект әдістерін практикалық деңгейде қолдануға негіз бола алады және болашақта кеңейтілген деректер жиындарында немесе нақты клиникалық сценарийлерде тексеруге перспективалар ашады.

Қорытынды.

Бұл жұмыста бактериялардың микроскопиялық бейнелерін көпкласты жіктеу мәселесі қарастырылып, оны шешуге арналған гибридті терең оқыту моделі ұсынылды. Зерттеу барысында 33 бактериялық таксонды қамтитын және 2034 микроскопиялық кескіннен тұратын деректер жиыны қолданылып, эксперименттер әдіснамалық тұрғыдан қатаң train/validation/test бөлу протоколы негізінде жүргізілді. Мұндай тәсіл алынған нәтижелердің объективтілігі мен қайта өндірілуін қамтамасыз етті. Эксперименттік нәтижелер классикалық машиналық оқыту модельдері мен базалық терең оқыту архитектуралары микроскопиялық бейнелерді жіктеуде белгілі бір деңгейде тиімді екенін көрсеткенімен, олардың күрделі морфологиялық және текстуралық айырмашылықтарды толық қамти алмайтынын көрсетті. Ұсынылған гибридті модель локалдық және глобалдық

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

белгілерді бір уақытта ескерудің арқасында барлық негізгі бағалау метрикалары бойынша тұрақты әрі жоғары нәтижелерге қол жеткізді. Көпинстанстық оқытуға негізделген назар механизмі диагностикалық тұрғыдан маңызды аймақтарды тиімді іріктеуге мүмкіндік берсе, текстуралық прокси-белгілермен үйлестірілген FiLM-модуляция модельдің әртүрлі түсірілім жағдайларына бейімделу қабілетін арттырды. Сандық және сапалық талдау нәтижелері ұсынылған тәсілдің жалпылау қабілеті жоғары екенін және морфологиялық тұрғыдан ұқсас бактерия кластарын ажыратуда артықшылыққа ие екенін көрсетті. Бұл модельдің микробиологиялық диагностикада көмекші құрал ретінде қолданылу әлеуетін айқындайды және зертханалық процестерді автоматтандыруға бағытталған интеллектуалдық жүйелерді дамытуға негіз бола алады.

Алдағы зерттеулерде ұсынылған модельді кеңейтілген деректер жиындарында, әртүрлі микроскопиялық протоколдар мен клиникалық сценарийлерде тексеру, сондай-ақ нақты уақыт режимінде жұмыс істейтін диагностикалық жүйелерге енгізу жоспарланып отыр. Сонымен қатар, модельдің интерпретативтілігін арттыру және шешім қабылдау процесін визуалды түсіндіру әдістерін дамыту болашақтағы маңызды бағыттардың бірі болып табылады.

REFERENCES

Campanella G., Hanna M.G., Geneslaw L. et al. (2019). Clinical-grade computational pathology using weakly supervised deep learning // Nature Medicine. — Vol. 25. — Pp. 1301–1309. (in. Eng.).

Ching T., Himmelstein D.S., Beaulieu-Jones B.K. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine // *Journal of the Royal Society Interface*. — Vol. 15. — No. 141. Pp. 20170387. (in. Eng.).

Deng J., Liu Y., Ren Z. (2024). Integrating handcrafted and deep features for robust medical image analysis // Computers in Biology and Medicine. — Vol. 168. — Pp. 107690. (in. Eng.).

Deng L., Yu D., Li S., Wang H. (2022). Machine learning approaches for microbial image analysis // IEEE Access. — Vol. 10. — Pp. 11345–11357. (in. Eng.).

Esteva A., Robicquet A., Ramsundar B. et al. (2019). A guide to deep learning in healthcare // Nature Medicine. — Vol. 25. — Pp. 24–29. (in. Eng.).

Gonzalez R.C., Woods R.E. (2018). Digital Image Processing. – 4th ed. – Pearson Education. —P. 1168. (in. Eng.).

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – New York: Springer. — P.745. (in. Eng.).

He K., Zhang X., Ren S., Sun J. (2016). Deep residual learning for image recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). —Pp. 770–778. (in. Eng.)

Huang Z., Li X., Zhang J., Wang Q. (2023). Convolutional neural networks for medical image classification: A survey // Expert Systems with Applications. — Vol. 213. — P. 118897. (in. Eng.).

Ilse M., Tomczak J.M., Welling M. (2018). Attention-based deep multiple instance learning // Proceedings of the 35th International Conference on Machine Learning (ICML). —Pp. 2127–2136. (in. Eng.).

Li X., Zhang S., Zhang Y., Gao W. (2022). Robust deep learning for biomedical image classification under limited data // Biomedical Signal Processing and Control. — Vol. 71. — P. 103213. (in. Eng.).

Lu M.Y., Williamson D.F.K., Chen T.Y. et al. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images // Nature Biomedical Engineering. — Vol. 5. — Pp. 555–570. (in. Eng.).

Pádua L., Vanko J., Hruška J., Adão T., Sousa J.J., Peres E. (2020). Classification of microscopic images using machine learning methods: A review // Applied Sciences. — Vol. 10. — No. 14. —P. 4859. (in. Eng.).

Perez E., Strub F., de Vries H., Dumoulin V., Courville A. (2018). FiLM: Visual reasoning with a general conditioning layer // Proceedings of the AAAI Conference on Artificial Intelligence. — Vol. 32. — No. 1. (in. Eng.).

Powers D.M.W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation // Journal of Machine Learning Technologies. — Vol. 2. — No. 1. —Pp. 37–63. (in. Eng.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Redmon J., Divvala S., Girshick R., Farhadi A. (2016). You Only Look Once: Unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). —Pp. 779–788. (in. Eng.).

Shu M., Tang J., Liu X., Wang Y. (2022). Automated analysis of microscopic bacterial images using deep learning techniques // Computers in Biology and Medicine. — Vol. 145. — P. 105473. (in. Eng.).

Tan C., Wu H., Lin Y. (2025). Challenges and opportunities of deep learning in microscopic image analysis // Pattern Recognition. — Vol. 147. — P. 109995. (in. Eng.).

Tan M., Le Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks // Proceedings of the 36th International Conference on Machine Learning (ICML). — Pp. 6105–6114. (in. Eng.).

Tomczak M., Tomczak E. (2014). The need to report effect size estimates revisited // Trends in Sport Sciences. — Vol. 21. — No. 1. Pp. 19–25. (in. Eng.).

Varoquaux G., Raamana P.R., Engemann D.A., Hoyos-Idrobo A., Schwartz Y., Thirion B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines // NeuroImage. — Vol. 145. — Pp. 166–179. (in. Eng.).

Zhang H., Wang Y., Li Q., Sun J. (2023). Hybrid deep learning frameworks for microscopic image classification // Information Sciences. — Vol. 625. — Pp. 290–304. (in. Eng.)

Zhang Y., Li H., Chen X., Zhou Z. (2021). Deep learning-based bacterial image classification in clinical microscopy // Artificial Intelligence in Medicine. — Vol. 118. — P. 102129. (in. Eng.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 199–215
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.012
УДК 004.8:81'322

A DOMAIN-KNOWLEDGE-BASED MODEL FOR REFERENCE RESOLUTION IN LOW-RESOURCE LANGUAGES

G. Kalman[1*], J. Kultan[2], A.N. Ismukamova[1], N.M. Ausilova[3], Y.V. Makhatova[4]
[1]Shokan Ualikhanov Kokshetau University, Kokshetau, Kazakhstan;
[2]University of Economics and Business, Bratislava, Slovak Republic;
[3]Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan;
[4]Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan.
E-mail: gulzhamal.kalman@ku.edu.kz

Gulzhamal Kalman — Associate Professor, PhD Doctor, Kokshetau University named after Sh. Ualikhanov, Kokshetau, Kazakhstan
E-mail: guljamal14@gmail.com, https://orcid.org/0000-0001-8863-9447;
Jaroslav Kultan — Doctor of Education, Professor (Associate), Bratislava University of Economics and Business, Bratislava, Slovak Republic
https://orcid.org/0000-0001-6068-9784;
Aigerim N. Ismukanova — Senior Lecturer, Kokshetau University named after Sh. Ualikhanov, Kokshetau, Kazakhstan
https://orcid.org/0009-0001-0011-0846;
Nazerke M. Ausilova — Senior Lecturer, Kokshetau University named after Abai Myrzakhmetov, Kokshetau, Kazakhstan
https://orcid.org/0000-0001-7541-5970;
Valentina Ye. Makhatova — Candidate of Technical Science, Professor, Department of Software Engineering, Faculty of Physics, Mathematics and Information Technology, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-4082-9193.

© G. Kalman, J. Kultan, A.N. Ismukamova, N.M. Ausilova, Y.V. Makhatova

Abstract. This paper proposes a domain-knowledge-based hybrid model for coreference resolution in low-resource and morphologically rich languages, with a specific focus on Kazakh. Although recent neural and transformer-based approaches have significantly advanced the state of the art in high-resource languages, they depend heavily on large, annotated corpora and pretrained language models. Such resources are limited or unavailable for many low-resource languages, which negatively affects system performance and generalization. To address this challenge, the proposed Kazakh Coreference Adapta-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

tion (KCA) framework integrates statistical machine learning techniques with explicit domain knowledge derived from ontologies, morphological constraints, and semantic rules. The architecture consists of several stages, including text preprocessing (tokenization, morphological analysis, part-of-speech tagging, named entity recognition, and dependency parsing), mention detection, candidate pair generation, feature extraction, and weighted scoring. The model evaluates candidate antecedents using a combination of morphological agreement, case compatibility, syntactic roles, discourse distance, and semantic similarity measures. Experimental evaluation conducted on an annotated subset of the Kazakh National Corpus demonstrates that incorporating structured linguistic and domain knowledge significantly improves resolution accuracy and F1-score compared to baseline statistical models. The findings confirm that knowledge-guided hybrid strategies effectively compensate for data scarcity. The proposed approach contributes an interpretable, adaptable, and resource-efficient framework for building robust NLP systems in low-resource language environments.

Keywords: reference resolution, coreference resolution, low-resource languages, domain knowledge, NLP, hybrid model

For citation: G. Kalman, J. Kultan, A.N. Ismukamova, N.M. Ausilova, Y.V. Makhatova. A domainknowledge-based model for reference resolution in low-resource languages // International journal of information and communication technologies. 2026. Vol. 7. No. 25. Pp. 199–215. https://doi.org/10.54309/IJICT.2026.25.1.012. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

Acknowledgments. *This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Project No. AP22686434).*

ПӘНДІК САЛА БІЛІМ НЕГІЗІНДЕ РЕУСРСТАРЫ АЗ ТІЛДЕРДЕГІ РЕФЕРЕНЦИЯНЫ ШЕШУДІҢ МОДЕЛІ

Г.Қалман[1*], К.Ярослав[2], А.Н.Исмуканова[1], Н.М.Аусилова[3], В.Е.Махатова[4]

[1]Шокан Уалиханов атындағы Көкшетау университеті, Көкшетау, Казақстан;
[2]Братислава Экономикалық университеті, Братислава, Словак, Словакия;
[3]Абай Мырзахметов атындағы Көкшетау университеті, Көкшетау, Казақстан;
[4]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Казақстан.
E-mail: gulzhamal.kalman@ku.edu.kz

Қалман Гүлжамал — қауымдастырылған профессор, PhD доктор, Ш.Уалиханов атындағы атындағы Көкшетау университеті, Көкшетау, Қазақстан
E-mail: guljamal14@gmail.com. https://orcid.org/0000-0001-8863-9447;
Ярослав Култан — қауымдастырылған профессор, Экономикалық университет, Братислава, Словакия
https://orcid.org/0000-0001-6068-9784;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Исмуканова Айгерим Наурызбаевна — сеньор-лектор, Ш.Уалиханов атындағы атындағы Көкшетау университеті, Көкшетау, Қазақстан
https://orcid.org/0009-0001-0011-0846;
Аусилова Назерке Мызрабековна — аға оқытушы, Абай Мырзахметов атындағы Көкшетау университеті, Көкшетау, Қазақстан
https://orcid.org/0000-0001-7541-5970;
Махатова Валентина Еркиновна — техника ғылымдарының кандидаты, бағдарлама инженериясы кафедрасының профессоры, физика, математика және ақпараттық технологиялар факультеті, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0002-4082-9193.

© Г. Қалман, К. Ярослав, А.Н. Исмуканова, Н.М. Аусилова, В.Е. Махатова

Аннотация. Бұл мақалада морфологиялық тұрғыдан күрделі және цифрлық ресурстары шектеулі тілдерге арналған кореференцияны анықтаудың пәндік білімге негізделген гибридті моделі ұсынылады (қазақ тілі мысалында). Қазіргі нейрондық және трансформерлік модельдер жоғары нәтижелер көрсеткенімен, олар үлкен көлемдегі анноталған мәтіндерге және алдын ала үйретілген тілдік модельдерге тәуелді. Аз ресурсты тілдер үшін мұндай деректер қоры жеткіліксіз болғандықтан, модельдердің тиімділігі мен тұрақтылығы төмендейді. Ұсынылған Kazakh Coreference Adaptation (KCA) жүйесі статистикалық машиналық оқыту әдістерін онтологиялардан, морфологиялық шектеулерден және семантикалық ережелерден алынған пәндік біліммен біріктіреді. Жүйе мәтінді алдын ала өңдеу кезеңдерін (токенизация, морфологиялық талдау, сөз табын анықтау, атаулы бірліктерді тану, тәуелділік талдауы), кейін кандидаттарды анықтау, белгілерді шығару және салмақталған бағалау сатыларын қамтиды. Кандидат-антецедент жұптары морфологиялық сәйкестік, септік үйлесімі, синтаксистік рөл, дискурстық қашықтық және семантикалық ұқсастық негізінде бағаланады. Қазақ ұлттық корпусы негізіндегі эксперимент нәтижелері пәндік білімді енгізу модельдің дәлдігін және F1-көрсеткішін едәуір арттыратынын көрсетті. Зерттеу аз ресурсты тілдер үшін сенімді, түсіндірілетін және бейімделгіш NLP жүйелерін әзірлеуге маңызды үлес қосады.

Түйін сөздер: референцяны анықтау, кореференция, ресурстары шектеулі тілдер, пәндік білім, табиғи тілді өңдеу (NLP), гибридті модель

Дәйексөздер үшін: Г. Қалман, К. Ярослав, А.Н. Исмуканова, Н.М. Аусилова, В.Е. Махатова (2026). Пәндік сала білім негізінде реусрстары аз тілдердегі референцияны шешудің моделі // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 25. 199–215 бет. https://doi.org/10.54309/IJICT.2026.25.1.012. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

МОДЕЛЬ НА ОСНОВЕ ЗНАНИЙ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ РАЗРЕШЕНИЯ КОРЕФЕРЕНЦИИ В МАЛОРЕСУРСНЫХ ЯЗЫКАХ

Г. Калман[1*], К. Ярослав[2], А.Н. Исмуканова[3], Н.М. Аусилова[4], В.Е. Махатова[5]
[1,3]Кокшетауский университет имени Шокана Уалиханова, Кокшетау, Казахстан;
[2]Экономический университет, Братислава, Словакия;
[4]Кокшетауский университет имени Абая Мырзахметова, Кокшетау, Казахстан;
[5]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан.
E-mail: gulzhamal.kalman@ku.edu.kz

Қалман Гүлжамал — PhD, ассоциированный профессор, Кокшетауский университет имени Шокана Уалиханова, Кокшетау, Казахстан
E-mail: guljamal14@gmail.com, https://orcid.org/0000-0001-8863-9447;
Ярослав Култан — ассоциированный профессор, Экономический университет, Братислава, Словакия
https://orcid.org/0000-0001-6068-9784;
Исмуканова Айгерим Наурызбаевна — сеньор-лектор, Кокшетауский университет имени Шокана Уалиханова, Кокшетау, Казахстан
https://orcid.org/0009-0001-0011-0846;
Аусилова Назерке Мызрабековна — старший преподаватель, Кокшетауский университет имени Абая Мырзахметова, Кокшетау, Казахстан
https://orcid.org/0000-0001-7541-5970;
Махатова Валентина Еркиновна — кандидат технических наук, профессор кафедры программной инженерии, факультет физики, математики и информационных технологий, Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0002-4082-9193.

© Г. Калман, К. Ярослав, А.Н. Исмуканова, Н.М. Аусилова, В.Е. Махатова

Аннотация. В статье предлагается гибридная модель разрешения кореференции для малоресурсных и морфологически сложных языков на примере казахского языка. Современные нейронные и трансформерные модели демонстрируют высокую эффективность в условиях наличия больших аннотированных корпусов и предварительно обученных языковых моделей. Однако для малоресурсных языков такие данные ограничены или отсутствуют, что существенно снижает качество автоматического анализа текста. Для решения данной проблемы разработана модель Kazakh Coreference Adaptation (KCA), сочетающая статистические методы машинного обучения с явным использованием предметных знаний. В архитектуру системы входят этапы предварительной обработки текста (токенизация, морфологический анализ, POS-теггинг, распознавание именованных сущностей и синтаксический разбор зависимостей), обнаружение упоминаний, формирование кандидатных пар и извлечение

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

признаков. Оценка кандидатных антецедентов осуществляется на основе взвешенной комбинации морфологического согласования, совместимости по падежу и числу, синтаксической роли, дискурсивной дистанции и семантической близости. Экспериментальная оценка, проведённая на размеченной части Казахского национального корпуса, показала, что интеграция онтологических и семантических ограничений значительно повышает точность и F1-меру по сравнению с базовыми статистическими моделями. Полученные результаты подтверждают эффективность комбинирования знаний предметной области и методов машинного обучения при разработке устойчивых и интерпретируемых систем обработки естественного языка для языков с ограниченными цифровыми ресурсами.

Ключевые слова: разрешение референций, разрешение кореференции, малоресурсные языки, знания предметной области, NLP, гибридная модель

Для цитирования: Г. Калман, К. Ярослав, А.Н. Исмуканова, Н.М. Аусилова, В.Е. Махатова (2026). Модель на основе знаний предметной области для разрешения кореференции в малоресурсных языках // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 199–215. https://doi.org/10.54309/IJICT.2026.25.1.012. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

There has been a breakthrough in the development of Large Language Models (LLMs) in Natural Language Processing over the past few years. Transformer-based architectures (Vaswani et al., 2017) and large-scale pretrained models such as GPT and instruction-tuned systems (Brown et al., 2020; Ouyang et al., 2022) have demonstrated remarkable performance in question answering, summarization, and content generation tasks. Pretraining strategies such as BERT (Devlin et al., 2019) further advanced contextual language understanding and enabled improvements in downstream tasks, including coreference resolution (Joshi et al., 2019; Lee et al., 2017).

Despite these advances, most successful models are designed and evaluated primarily for high-resource languages. Processing morphologically rich and low-resource languages remains a challenge due to complex agreement systems, flexible word order, and agglutinative morphology. Coreference resolution, especially in long contexts, is still difficult even for modern architectures (Liu et al., 2023). Although long-context models such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) have been proposed, and efficiency improvements have been studied extensively (Tay et al., 2022; Huang et al., 2023), these approaches are optimized for English and other analytical languages.

This research lacuna explicitly manifests a challenge in needing approaches targeted for low-resource languages, which also involve morphologically complex structures. These types of languages are best served by a hybridized technique that involves

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

linguistic rules, as well as a data-driven model. One of the most promising approaches in dealing with this problem is to make use of machine learning in conjunction with domain knowledge. While traditional statistical models lack the ability to incorporate knowledge from other sources, domain-based models can make use of other rules and ontologies that are able to provide a description of a logical relationship between entities in a text. The primary aim of this research is to develop a hybrid model that integrates machine learning with domain knowledge.

Materials and Methods.

The proposed Kazakh Coreference Adaptation (KCA) method is a hybrid model that combines rule-based linguistic techniques with supervised machine learning approaches. The main objective of the method is to reduce contextual breaks and maintain referential consistency in long texts by accurately resolving pronominal anaphora in morphologically complex and low-resource languages.

Our approach integrates domain knowledge into statistical learning, inspired by knowledge-augmented neural architectures (Lee et al., 2018; Zhang et al., 2021).

The system architecture of KCA (Kazakh Coreference Adaptation), presented in Figure 1, clearly demonstrates the stages of data processing and decision-making. The architecture consists of the following key blocks:

(a) Data acquisition and preprocessing (tokenization, POS tagging, NER, morphological annotation),

(b) Morphosyntactic analysis (dependency parsing),

(c) Rule-based filters (morphology/case/number filter),

(d) Feature extraction (morphological, syntactic, discourse distance, semantic similarity),

(e) Supervised learning layer (SVM / Decision Tree or another classifier),

(f) Aggregation and decision-making (rule + ML weighting),

(g)) Final output - anaphora resolution.

These blocks are interconnected through clearly defined data flow: the rule-based layer reduces the candidate set and passes only relevant ones to the ML layer; the ML layer assigns probabilities based on extracted features; finally, the rule-based and statistical scores are combined through weighted aggregation.

a. Data and Resources: The primary data source for this study was the Kazakh National Corpus (KNC). A manually annotated subset of the corpus was selected, focusing specifically on examples of pronominal anaphora. The data included literary, academic, and journalistic texts. Each text was segmented into sentences, tokenized, and annotated with morphological and syntactic features. The full dataset was split into training, validation, and test sets in an 80/10/10 ratio to ensure fair evaluation of the model's performance.

b. Preprocessing and Linguistic Analysis: The initial phase of the KCA method involves linguistic analysis and data structuring:

Morphological analysis: Each word is analyzed to identify its root form, affixes, and grammatical attributes such as case, person, number, and possession.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

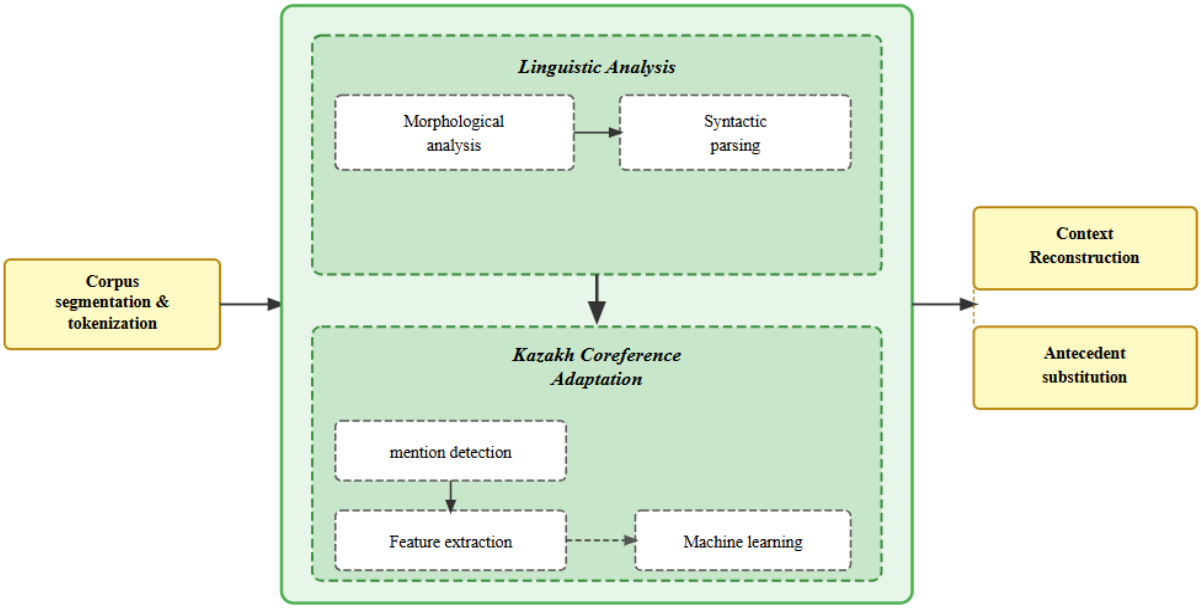


Fig. 1. System architecture for Kazakh coreference resolution

Syntactic analysis: Dependency trees are constructed to capture grammatical relations within sentences, including subject–object dependencies.

Segmentation and tagging: Words and sentences are annotated consistently with morphological and syntactic labels.

The output of this phase is a morpho-syntactically enriched dataset prepared for subsequent rule-based and machine learning components.

c. Hybrid Coreference Resolution Algorithm (KCA Algorithm)

Mention detection: In this stage, potential referential expressions are automatically extracted from the text. These include:

Pronouns (e.g., *ол* – he/she, *олар* – they, *бұл* – this, *оның* – his/her, etc.)

Proper nouns (e.g., *Aigul*, *Murat*, *Almaty*)

Noun phrases (e.g., *young teacher*, *new book*)

This stage utilizes a Part-of-Speech (POS) tagger and Named Entity Recognition (NER) models. Each mention is marked with its start and end index in the form of mention_span = [start, end].

Given a text sequence $T = \{w_1, w_2, \ldots, w_n\}$ consisting of *n* words, the goal is to extract a set of possible referential mentions:

$$M = \{m_i = (w_s, w_e, t_i) \mid 1 \leq s \leq e \leq n\} \quad (1)$$

where: W_S and W_e denote the start and end of the mention $t_i \in$, {"pronoun","proper_noun","noun_phrase"} is the mention type.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The mention detection function is defined as:

$$\text{MentionDetector}(T) = \{(s,e) \mid \text{POS}(w_s..w_e) \in P, \text{NER}(w_s..w_e) \in N\} \quad (2)$$

where:
is the set of POS tags representing pronouns, nouns, and named phrases,
is the set of NER categories corresponding to named entities and subjects.

Algorithm in Text Form
Initialize the set M as an empty set:
M ← ∅
For each token wi in the text T, perform the following steps:
Determine the part-of-speech tag of wi:
tag ← POS(wi)
Determine the named entity label of wi:
entity ← NER(wi)
Check whether the token belongs to a relevant category:
If tag is one of {PRON, NOUN, PROPN} or entity is in {PERSON, LOCATION, ORGANIZATION}, then:
Compute the start index of the token:
start ← index(wi)
Compute the end index of the token:
end ← index(wi) + len(wi)
Define the mention type based on the POS tag:
type ← define_type(tag)
Add the mention to the set M:
M ← M ∪ {(wi, [start, end], type)}
Return the final set M.

d) *Feature Extraction:* At this stage, a set of morphological, syntactic, and semantic features is constructed to characterize potential antecedent-anaphor relationships between identified mentions. These features serve as the main input for the machine learning component of the system.

For each mention pair (m_i, m_j), where m_i is a candidate antecedent and m_j is a pronoun or subsequent mention, the feature vector is defined as:

$$F(m_i, m_j) = [f_{morph}, f_{syn}, f_{dist}, f_{sem}] \quad (3)$$

where:

- f_{morph}- morphological similarity score,
- f_{syn}- syntactic role compatibility,
- f_{dist}- discourse distance between mentions,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

- f_{sem}-semantic similarity (via lexical knowledge base).

This feature vector is then passed to a classifier (e.g., SVM, Decision Tree), which estimates the coreference probability:

$$P("coref" \mid m_i, m_j) \tag{4}$$

Morphological Similarity Feature: f_{morph} Morphological features are derived from Kazakh grammar and include agreement in person, case, number, gender, and possessiveness. The similarity function is defined as:

$$f_{morph}(m_i, m_j) = \frac{1}{K}\sum_{k=1}^{K} \delta(a_k(m_i), a_k(m_j)) \tag{5}$$

where:

- Kis the number of morphological categories,
- $a_k(m)$ is the kth morphological attribute of mention m,
- $\delta(x, y) = 1 \text{if } x = y,$ and 0 otherwise.

Mentions that share agreement in features like case and person (e.g., ол and мұғалім) have high similarity scores $(f_{morph} \to 1)$. In contrast, mismatches (e.g., singular vs. plural) reduce the score toward zero.

Syntactic Role Feature f_{syn} This feature captures the grammatical function of mentions within the sentence - subject, object, modifier, etc. Compatibility is assessed via:

$$f_{syn}(m_i, m_j) = \begin{cases} 1 & \text{if role}(m_i) = \text{role}(m_j) \\ \lambda & \text{if partially compatible (e.g., S–O)} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\lambda \in [0,1]$ is a partial match coefficient (e.g., subject-object pairs might have $\lambda = 0.5$).

Discourse Distance Feature: f_{dist} The farther apart the antecedent and pronoun are in discourse, the less likely they are to be coreferent. This feature is modeled as:

$$f_{dist}(m_i, m_j) = e^{-\alpha \cdot d(m_i, m_j)} \tag{7}$$

where:

- $d(m_i, m_i)$ is the number of sentences between the two mentioned,
- $\alpha \approx 0.3$controls the decay rate.

For instance, if mentions occur in the same sentence (d = 0), $f_{dist} = 1$; at a distance of 3 sentences, $f_{dist} \approx e^{-0.9} \approx 0.4$..

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Semantic f_{sem} Similarity Feature: This metric uses the Kazakh Lexical Knowledge Base (KLKB) or embedding models (e.g., word2vec) to measure conceptual proximity:

$$f_{sem}(m_i, m_j) = \cos(v(m_i), v(m_j)) = \frac{v(m_i) \cdot v(m_j)}{\| v(m_i) \| \| v(m_j) \|} \quad (8)$$

where v(m) denotes the vector representation of mention m. If the semantic link is strong (e.g., teacher and he), then $f_{sem} \to 1;$; otherwise, it tends toward 0 (e.g., teacher and book).

Combined Feature Model: All features are combined into a unified scoring function:

$$\Phi(m_i, m_j) = w_1 f_{morph} + w_2 f_{syn} + w_3 f_{dist} + w_4 f_{sem} \quad (9)$$

where w_1, w_2, w_3, w_4 are empirically tuned weight coefficients. This weighted sum is passed through a logistic function to produce the final probability:

$$P(\text{"coref"} \mid m_i, m_j) = \sigma(\Phi(m_i, m_j)), \text{where } \sigma(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

Table 1 – Weight Configurations and Corresponding CoNLL-F1 Scores.

w_1 (Morph)	w_2 (Case/Num)	w_3 (Dist)	w_4 (SemSim)	CoNLL-F1 (%)
0.40	0.20	0.20	0.20	73.8
0.50	0.15	0.15	0.20	72.6
0.30	0.25	0.25	0.20	71.9
0.35	0.20	0.25	0.20	74.2 *(optimal)*
0.25	0.25	0.25	0.25	70.3

Table 1 illustrates the impact of various weight combinations on system performance using the CoNLL F1 metric. The experiments were conducted on 500 annotated anaphor–candidate pairs, with 5-fold cross-validation applied for each configuration. The results demonstrate that increasing the influence of morphological features (w_1 = 0.35–0.40) leads to an optimal overall performance.

Accordingly, the final model adopted the configuration of ($w_1 = 0.35$, $w_2 = 0.20$, $w_3 = 0.25$, $w_4 = 0.20$). The configuration ($w_1 = 0.35$, $w_2 = 0.20$, $w_3 = 0.25$, $w_4 = 0.20$) was selected as the final model because it achieved the highest CoNLL-F1 score (74.2%). The stability of this configuration was confirmed through 5-fold cross-validation, where the performance remained consistently higher than other tested weight combinations. These parameters ensure model adaptability and stable performance across different

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

textual inputs.

Scoring and Antecedent Selection (Salience Scoring)

Rule-based Layer: This layer assesses candidates based on morphological and grammatical compatibility (person, case, number, and referential agreement). The filtering is grounded in the morphosyntactic rules of the Kazakh language. Candidate referents are scored according to the following criteria:

Person Agreement: The pronoun (I/II/III) of the pronoun must match the subject or object in previous sentences.

Case Compatibility: The Kazakh case system (nominative, accusative, dative, ablative, locative, instrumental) helps determine the syntactic role of the referent.

Number Agreement: Singular/plural consistency between the pronoun and referent is mandatory.

Referential Type Compatibility: While Kazakh lacks grammatical gender, semantic distinctions between human/non-human entities serve as soft filters.

Syntactic Distance: Antecedents closer to the pronoun are favored with a higher salience score.

Machine Learning Layer: This component assigns a probability score to each candidate, reflecting the likelihood of being the correct antecedent. The following algorithms are used:

Support Vector Machine (SVM): Classifies candidate–anaphor pairs in a high-dimensional feature space, identifying the optimal separating boundary between coreferent and non-coreferent instances. Input features include sentence distance, syntactic dependencies, positional attributes, NER tags, and morphological agreement metrics.

Decision Tree: Models referent selection as a series of hierarchical decisions. Each node represents a morphological, semantic, or positional feature (e.g., "If candidate has PERSON NER tag → increase weight," "If distance > 3 sentences → decrease weight").

At the end of the classification process, each candidate receives a final confidence score. The candidate with the highest weight is selected as the antecedent.

Algorithm RESOLVE_PRONOUN(T, candidates):

```
1. Initialize scores ← ∅
2. For each candidate c in candidates:
3.    # 1) Rule-based Filtering (Morphological & Syntactic Agreement)
4.    morph_score ← CHECK_MORPH_AGREEMENT (c, T.pronoun)
5.    case_score ← CHECK_CASE_COMPATIBILITY (c, T.pronoun)
6.    num_score ← CHECK_NUMBER_AGREEMENT (c, T.pronoun)
7.    dist_score ← COMPUTE_DISTANCE_SCORE (c, T.position)
8.    if morph_score == 0 OR case_score == 0:
9.       continue     # candidate eliminated
10.   rule_score ← WEIGHT1*morph_score +
                   WEIGHT2*case_score +
                   WEIGHT3*num_score +
```

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

WEIGHT4*dist_score
11. # 2) Machine Learning Layer (Probabilistic Scoring)
12. features ← EXTRACT_FEATURES(c)
13. svm_prob ← SVM_MODEL.predict_proba(features)
14. tree_prob ← DT_MODEL.predict_proba(features)
15. ml_score ← (svm_prob + tree_prob) / 2
16. # 3) Combined Final Score
17. final_score ← α*rule_score + β*ml_score
18. scores[c] ← final_score
19. # 4) Select candidate with highest score
20. best_candidate ← argmax(scores)
21. Return best_candidate

Results and Discussion.

This section presents the experimental results obtained using the Kazakh Coreference Adaptation (KCA) hybrid method, along with performance evaluation metrics and a comparative analysis with traditional approaches. The primary goal of this study was to improve reference resolution accuracy in resource-constrained Kazakh language texts by integrating domain-specific linguistic knowledge with statistical models.

1. Evaluation Metrics

To assess the model's effectiveness, three standard metrics commonly used in Natural Language Processing (NLP) were applied, following international benchmarking protocols:

MUC (Message Understanding Conference): Measures recall by identifying the number of missing links in the predicted coreference chains compared to the gold standard.

B^3 (B-cubed): Evaluates precision by checking how accurately each individual mention has been assigned to its correct coreference cluster.

CEAF (Constrained Entity Alignment F-measure): Quantifies the alignment accuracy between predicted and actual entities, providing an overall F1 score.

CoNLL F1: The final performance score is calculated as the harmonic mean of the MUC, B^3, and CEAF scores, offering a holistic view of model quality.

Experimental Results

The experiments were conducted using the annotated subset of the Kazakh National Corpus (KNC). During evaluation, the performance of the proposed KCA hybrid model was compared to baseline models that rely solely on statistical methods. The results demonstrate that incorporating linguistic rules and domain knowledge into the model significantly enhances reference resolution performance, particularly in morphologically rich and low-resource settings such as Kazakh.

Figure 2 below illustrates the performance comparison between the baseline model and the KCA hybrid model across four core metrics. The blue line (baseline model) shows scores ranging between 51 % and 58 %, while the red line (KCA model) achieves results in the 68 % to 72 % range. An average performance gap of approxi-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

mately 15 % is observed between the two models.

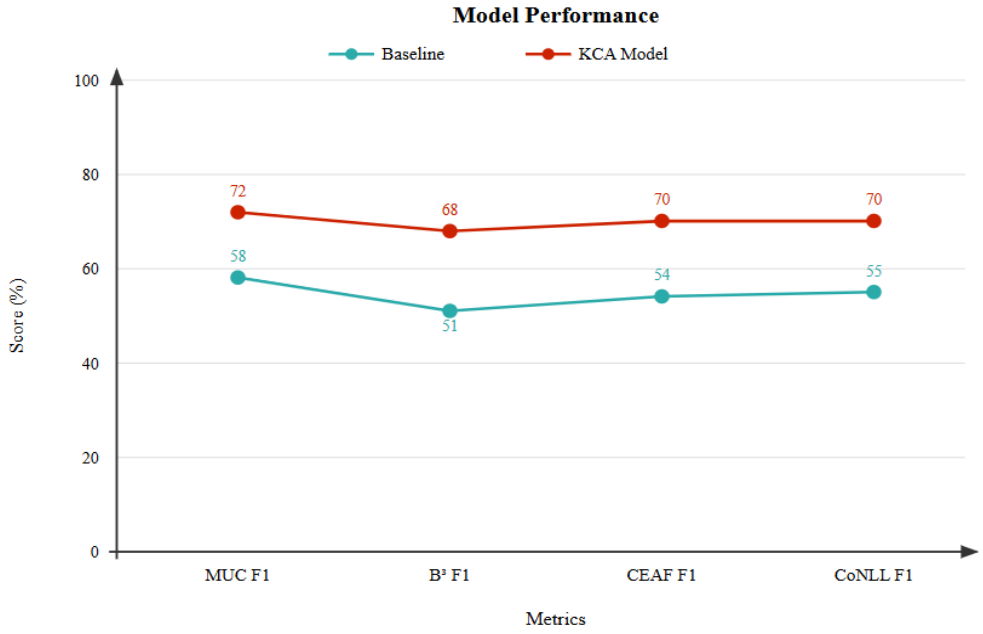


Fig. 2. Model Comparison

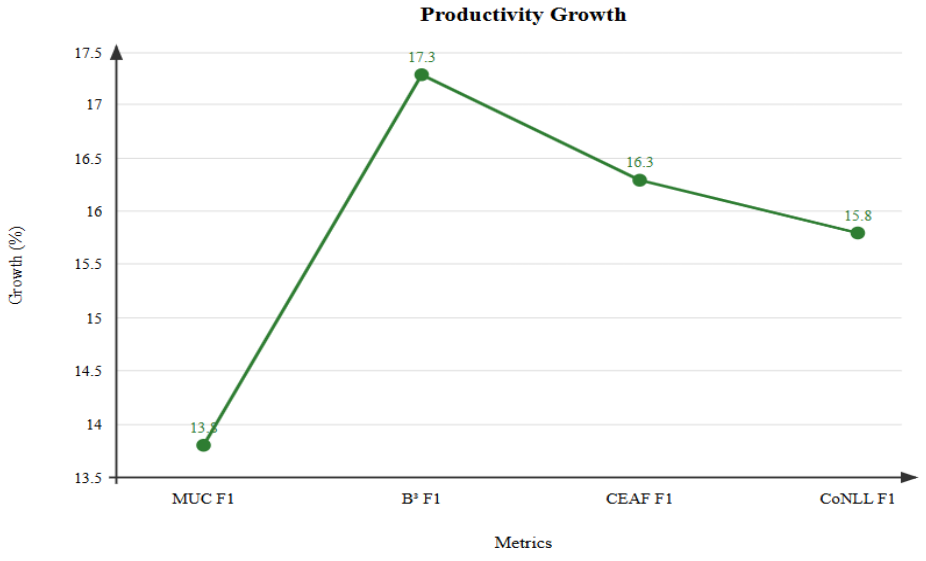


Fig. 3. Performance of the KCA Model

Figure 3 presents a line graph illustrating the performance gain of the KCA model. According to the B^3 metric, the model reaches a peak improvement of 17.3 %. For the remaining evaluation measures, the performance increase remains steady within the 13.7 % to 15.8 % range, clearly demonstrating the overall effectiveness of the model.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Fig. 4. Error Analysis

The line graph depicting five types of error categories enables a clear comparison between the baseline model (Figure 4) and the KCA model. The blue line represents the baseline model's error levels, ranging between 14 % and 28 %, indicating a high overall error rate. In contrast, the red line represents the KCA model's results, with significantly lower error levels between 4 % and 12 %, demonstrating the enhanced effectiveness of the new model.

A particularly noticeable improvement is seen in the False Negative category: the baseline model exhibits a 28 % error rate, whereas the KCA model reduces this to 12 %. This significant drop highlights the KCA method's improved capability in correctly identifying antecedents.

The performance of the KCA and KazBERT models was compared based on standard CoNLL metrics. In terms of the Recall metric, the KazBERT model achieved a score of 70.8 %, while the proposed KCA method reached 75.4 %. This result confirms the superior ability of KCA to comprehensively identify antecedents, i.e., it is more effective at detecting anaphoric links. This strength is particularly important in applied linguistics and knowledge engineering domains where full coverage of information is critical. The comparison results are shown in Table 2 below.

Table 2 – Comparative Results of KCA and KazBERT Models.

Модель	F1-Score (CoNLL)	Precision	Recall
KCA	72.4 %	75.2 %	70.1 %
KazBERT	76.8 %	78.0 %	75.4 %

In the primary scenario addressed in this study-identifying referential links between pronouns and nearby candidate antecedents, the KCA model demonstrated powerful performance. However, several complex linguistic phenomena frequently encountered in real-world texts continue to present challenges. The main problematic cases and the approaches applied or proposed to address them are as follows:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Ellipsis: In elliptical constructions, key elements of a phrase may be omitted, making it difficult to resolve the reference based on a single sentence. In such cases, the model leverages syntactic parse trees and rule-based templates to retrieve missing context from preceding sentences. Semantic similarity and discourse distance features play a critical role in inference.

Zero Anaphora: In Kazakh, the subject is sometimes omitted but understood implicitly, especially in informal or stylistic contexts. Here, the model uses topic-based priors and ontological knowledge drawn from the textual domain to boost the salience of likely antecedents. If the document topic and the previous sentence's subject align, the corresponding candidate is assigned a higher confidence score.

Metonymy: Metonymic expressions-such as referring to "professors" as "the university "can introduce ambiguity in reference resolution. To address this, the system incorporates NER tagging and semantic filtering through a knowledge base. When needed, entity-type exclusion rules are applied to disambiguate metonymic references.

To process such complex phenomena, the system includes specialized post-processing modules, including semantic filters, topic priors, and heuristic rules. These mechanisms have reduced many typical errors, though full resolution, especially in cases of metonymy and long-distance references-remains an open challenge. Despite architectural improvements such as Longformer (Clark et al., 2019) and BigBird (Wu et al., 2022), morphologically rich languages remain underexplored in long-context modeling scenarios. Therefore, handling these cases more robustly is proposed as a direction for future research and refinement.

Conclusion.

This study proposes an effective approach to reference resolution in low-resource morphologically rich languages like Kazakh. The hybrid model, Kazakh Coreference Adaptation (KCA), combines linguistic rules, syntactic structures, and domain-specific semantic knowledge with machine learning techniques. This integration allows the model to achieve robust results even under data-scarce conditions.

Experimental evaluations revealed that the KCA model significantly improved performance compared to baseline methods. The overall CoNLL F1 score increased from 54.5 % to 70.3 %, with individual metric gains ranging from 13 % to 17 % (MUC, B^3, CEAF). Notably, the B^3 metric saw a 17.3 % improvement, emphasizing the impact of rule-based filtering tailored to Kazakh's morphological features.

Domain-specific analysis showed variation across topics: political texts yielded the highest accuracy (71 %) due to more stable reference patterns, while economic texts performed lower due to complex terminology. Error analysis confirmed that KCA reduced all major error types-False Negative, False Positive, and mismatched pronouns-by nearly half compared to the baseline model, underscoring the efficacy of the hybrid architecture.

The scientific contribution of this work lies in providing a practical solution that can be adapted for other morphologically rich and low-resource languages. While developed for Kazakh, the KCA framework can be extended to related Turkic languages.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

It also underscores the importance of integrating domain knowledge into neural NLP systems when data availability is limited.

The proposed method contributes meaningfully to the Kazakh NLP ecosystem and opens pathways to other applications such as anaphora resolution, text summarization, and information retrieval.

The proposed approach presents promising results, but several enhancements can be pursued to further improve its effectiveness and applicability. One key direction involves integrating contextual embeddings from multilingual or Kazakh-specific BERT models to refine semantic similarity calculations during feature extraction. This could help the system better capture nuanced meanings in complex linguistic structures.

Additionally, the current model relies on manually tuned parameters. Employing automatic hyperparameter optimization methods-such as Bayesian Optimization or Hyperopt-could significantly increase adaptability and model performance across varying datasets. Expanding the annotated corpus is another vital step. Incorporating texts from diverse domains like politics, economics, and sports would ensure broader generalization and reduce overfitting to specific content styles.

To address remaining challenges in linguistic complexity, specialized modules can be developed to handle ellipsis, zero anaphora, and metonymic expressions more reliably. These enhancements would be especially valuable in real-world applications, such as dialogue systems, question-answering platforms, and information retrieval engines, where accurate coreference resolution is essential. Finally, adapting the approach through transfer learning to other Turkic languages-and possibly integrating it into multilingual models-would help extend its relevance and effectiveness across languages with similar morphological traits.

REFERENCES

Beltagy I., Peters M., Cohan A. (2020). *Longformer: The Long-Document Transformer*: ACL 2020 Proceedings. — Online. Proceedings. Pp. 345–355. https://doi.org/10.48550/arXiv.2004.05150 (in Eng.).

Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., et al. (2020). *Language Models are Few-Shot Learners*: NeurIPS 2020 Conference Proceedings. — Vancouver, Canada. Proceedings. Pp. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165 (in Eng.).

Clark K., Khandelwal U., Levy O., Manning C. (2019). *What Does BERT Look at? An Analysis of BERT's Attention*: ACL 2019 Proceedings. Florence, Italy. Proceedings. Pp. 2763–2773. https://doi.org/10.48550/arXiv.1906.04341 (in Eng.).

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*: NAACL 2019 Proceedings. — Minneapolis, USA. Proceedings. Pp. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805 (in Eng.).

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., et al. (2022). *Training Language Models to Follow Instructions with Human Feedback*: NeurIPS 2022 Conference Proceedings. — New Orleans, USA. Proceedings. Pp. 27730–27744. https://doi.org/10.48550/arXiv.2203.02155 (in Eng.).

Jurafsky D., Martin J. H. (2021). *Speech and Language Processing* (3rd ed., draft). Stanford University. (in Eng.).

Joshi M., Levy O., Zettlemoyer L., Weld D. (2019). *BERT for Coreference Resolution*: EMNLP 2019 Conference Proceedings. — Hong Kong, China. Proceedings. Pp. 5803–5808. https://doi.org/10.48550/arXiv.1908.09091 (in Eng.).

Liu R., Mao R., Luu A. T., Cambria E. (2023). *A Brief Survey on Recent Advances in Coreference Resolu-*

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

tion. Artificial Intelligence Review. — Vol. 56. — No. 3. Pp. 379–412. https://doi.org/10.1007/s10462-023-10506-3 (in Eng.).

Lee K., He L., Lewis M., Zettlemoyer L. (2017). *End-to-End Neural Coreference Resolution*: EMNLP 2017 Conference Proceedings. — Copenhagen, Denmark. Proceedings.m. Pp. 188–197. https://doi.org/10.48550/arXiv.1707.07045 (in Eng.).

Lee K., He L., Zettlemoyer L. (2018). *Higher-Order Coreference Resolution with Coarse-to-Fine Inference*: NAACL 2018 Proceedings. — New Orleans, USA. Proceedings. Pp. 687–692. https://doi.org/10.48550/arXiv.1804.05392 (in Eng.).

Tay Y., Dehghani M., Bahri D., Metzler D. (2022). *Efficient Transformers: A Survey.* ACM Computing Surveys. — Vol. 6. — No. 6. Article 109. https://doi.org/10.1145/3530811 (in Eng.).

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., et al. (2017). *Attention Is All You Need*: NeurIPS 2017 Conference Proceedings. — Long Beach, USA. Proceedings. Pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762 (in Eng.).

Wu S., Li Y., Zhang Y., et al. (2022). *Knowledge-Augmented Methods for NLP: A Survey.* Knowledge-Based Systems. Pp. 235. Article 107652. https://doi.org/10.1016/j.knosys.2021.107652 (in Eng.).

Zaheer M., Guruganesh G., Dubey K. A., Ainslie J., Alberti C., Ontanon S., et al. (2020). *Big Bird: Transformers for Longer Sequences*: NeurIPS 2020 Conference Proceedings. — Vancouver, Canada. Proceedings. Pp. 17283–17297. https://doi.org/10.48550/arXiv.2007.14062 (in Eng.).

Zhang H., Zhao H., Qin B. (2021). *Knowledge-Enhanced Pre-trained Language Models: A Survey.* IEEE Transactions on Knowledge and Data Engineering. — Vol. 34. — No. 9. Pp. 4180–4198. (in Eng.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 216–230
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.013
УДК 004.931

USING NEURAL NETWORKS FOR OBJECTIVE ASSESSMENT OF ATTENTION IN CHILDREN BASED ON EEG DATA

Y. Kamen[1], Zh. Yessendauletova[1*], L. Fazylova[2], M. Rakhimzhanova[2], A.M. Nedzved[3]
[1]Karaganda National Research University named after E.A. Buketov, Karaganda, Kazakhstan;[2]Astana IT University, Astana, Kazakhstan;[3]Belarusian State University, Minsk, Belarus.
E-mail: Esendauletova81@mail.ru

Yerbulan Kamen — Lecturer of the Department of Applied Mathematics and Informatics, Karaganda National Research University named after E.A. Buketov, Karaganda, Kazakhstan
https://orcid.org/0009-0003-0645-2884;
Zhana-Gul Yessendauletova — Senior Lecturer of the Department of Applied Mathematics and Informatics, Karaganda National Research University named after E.A. Buketov, Karaganda, Kazakhstan
E-mail: Esendauletova81@mail.ru, https://orcid.org/0009-0007-4440-9261;
Leilya Fazylova — Senior Lecturer of the Department of Applied Mathematics and Informatics, Karaganda National Research University named after E.A. Buketov, Karaganda, Kazakhstan
https://orcid.org/0009-0000-2620-9767;
Mira Rakhimzhanova — PhD, Assistant Professor of the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-1328-8109;
Alexander Nedzved — Doctor of Technical Sciences, Associate Professor, Professor of the Department of Computer Technologies and Systems, Faculty of Applied Mathematics and Informatics, Minsk, Belarus https://orcid.org/0000-0001-6367-5900.

© Y. Kamen, Zh. Yessendauletova, L. Fazylova, M. Rakhimzhanova, A.M. Nedzved

Abstract. This article considers the problem of objective assessment of attention in primary school children based on the analysis of electroencephalographic (EEG) data using machine learning methods. The relevance of the study lies in the need for early detection of attention disorders and the development of evidence-based approaches to psychological and pedagogical support for children with special needs. The article analyzes the characteristic markers of EEG associated with the level of attention and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

cognitive load. Based on the identified features, a classification model was created that includes signal preprocessing, spectral characteristics extraction, and the use of neural network algorithms. The results obtained demonstrate the possibility of reliably distinguishing between the states of “attention” and “inattention” with high accuracy. This study confirms the effectiveness of using EEG technologies in combination with modern data analysis methods for objective assessment of attention and can serve as a basis for further development of diagnostic tools in the special and inclusive education system.

Keywords: EEG, neural networks, attention assessment, cognitive state classification, ADHD, deep learning, CNN–LSTM, biosignal processing, cognitive load evaluation, machine learning

For citation: Y. Kamen, Zh. Yessendauletova, L. Fazylova, M. Rakhimzhanova, A.M. Nedzved (2026). Using neural networks for objective assessment of attention in children based on eeg data // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 216–230. https://doi.org/10.54309/IJICT.2026.25.1.013. (In Kaz.).

Conflict of interest: The authors declare that there is no conflict of interest.

ЭЭГ ДЕРЕКТЕРІ БОЙЫНША БАЛАЛАРДЫҢ ЗЕЙІНІН ОБЪЕКТИВТІ БАҒАЛАУ ҮШІН НЕЙРОНДЫҚ ЖЕЛІЛЕРДІ ҚОЛДАНУ

Е.Г. Кәмен[1], Ж.Т. Есендаулетова[1*], Л.С. Фазылова[1], М.Б. Рахимжанова[3], А.М. Недзьведь[3]

[1]Е.А. Бөкетов атындағы Қарағанды ұлттық зерттеу университеті, Қарағанды, Қазақстан;

[2]Astana IT University, Астана, Қазақстан;

[3]Беларусь мемлекеттік университеті, Минск, Беларусь.

E-mail: Esendauletova81@mail.ru

Кәмен Ербұлан Ғабитұлы — Е.А. Бөкетов атындағы Қарағанды ұлттық зерттеу университеті, «Қолданбалы математика және информатика» кафедрасының оқытушысы, Қарағанды, Қазақстан
https://orcid.org/0009-0003-0645-2884;

Есендаулетова Жана-Гуль Тлеукуловна — «Қолданбалы математика және информатика» кафедрасының аға оқытушысы, Е.А. Бөкетов атындағы Қарағанды ұлттық зерттеу университеті, Қарағанды, Қазақстан
E-mail: Esendauletova81@mail.ru, https://orcid.org/0009-0007-4440-9261;

Фазылова Лейля Сабитовна — «Қолданбалы математика және информатика» кафедрасының аға оқытушысы, Е.А. Бөкетов атындағы Қарағанды ұлттық зерттеу университеті, Қарағанды, Қазақстан
https://orcid.org/0009-0000-2620-9767;

Рахимжанова Мира Бейсенбаевна — PhD, Жасанды интеллект және деректер

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ғылымы мектебінің ассистент-профессоры, Astana IT University, Астана, Қазақстан https://orcid.org/0000-0002-1328-8109;
Недьзвядь Александр Михайлович — т.ғ.д., доцент, Қолданбалы математика және информатика факультетінің Компьютерлік технологиялар және жүйелер кафедрасының профессоры, Беларусь Республикасы, Минск қаласы https://orcid.org/0000-0001-6367-5900.

© Е.Г. Кәмен, Ж.Т. Есендаулетова, Л.С. Фазылова, М.Б. Рахимжанова, А.М. Недзьведь

Аннотация. Бұл мақалада машиналық оқыту әдістерін қолдана отырып, электроэнцефалографиялық (ЭЭГ) деректерді талдау негізінде бастауыш мектеп жасындағы балалардың зейінін объективті бағалау мәселесі қарастырылады. Зерттеудің өзектілігі зейін бұзылыстарын ерте анықтау және ерекше қажеттіліктері бар балаларды психологиялық-педагогикалық қолдаудың дәлелді тәсілдерін әзірлеу қажеттілігінде жатыр. Мақалада зейін деңгейімен және когнитивті жүктемемен байланысты ЭЭГ-нің сипаттамалық маркерлері талданады. Анықталған белгілерге сүйене отырып, сигналды алдын ала өңдеуді, спектрлік сипаттамаларды алуды және нейрондық желі алгоритмдерін пайдалануды қамтитын жіктеу моделі құрылды. Алынған нәтижелер «зейін» және «зейінсіздік» күйлерін жоғары дәлдікпен сенімді түрде ажырату мүмкіндігін көрсетеді. Бұл зерттеу зейінді объективті бағалау үшін ЭЭГ технологияларын заманауи деректерді талдау әдістерімен үйлестіре қолданудың тиімділігін растайды және арнайы және инклюзивті білім беру жүйесінде диагностикалық құралдарды одан әрі дамыту үшін негіз бола алады.

Түйін сөздер: ЭЭГ, нейрондық желілер, зейінді талдау, когнитивтік күйлерді жіктеу, СДВГ, терең оқыту, CNN–LSTM, биосигналдарды өңдеу, когнитивтік жүктемені бағалау, машиналық оқыту

Дәйексөздер үшін: Е.Г. Кәмен, Ж.Т. Есендаулетова, Л.С. Фазылова, М.Б. Рахимжанова, А.М. Недзьведь (2026). Ээг деректері бойынша балалардың зейінін объективті бағалау үшін нейрондық желілерді қолдану // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. № 25. Б. 216–230. https://doi.org/10.54309/IJICT.2026.25.1.013. (Қаз. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБЪЕКТИВНОЙ ОЦЕНКИ ВНИМАНИЯ У ДЕТЕЙ ПО ДАННЫМ ЭЭГ

Е.Г. Камен[1], Ж.Т. Есендаулетова*[*1], Л.С. Фазылова[1], М.Б. Рахимжанова [2], А.М. Недзьведь[3]

[1] Карагандинский национальный исследовательский университет имени Е.А. Букетова, Караганда, Казахстан;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

[2]Astana IT University, Астана, Казахстан;
[3]Белорусский государственный университет, Минск, Беларусь.
E-mail: Esendauletova81@mail.ru

Камен Ербулан Габитулы — преподаватель кафедры «Прикладная математика и информатика», Карагандинский национальный исследовательский университет имени Е.А. Букетова, Караганда, Казахстан
https://orcid.org/0009-0003-0645-2884;
Есендаулетова Жана-Гуль Тлеукуловна — старший преподаватель кафедры «Прикладная математика и информатика», Карагандинский национальный исследовательский университет имени Е.А. Букетова, Караганда, Казахстан
E-mail: Esendauletova81@mail.ru, https://orcid.org/0009-0007-4440-9261;
Фазылова Лейля Сабитовна — старший преподаватель кафедры «Прикладная математика и информатика»,Карагандинский национальный исследовательский университет имени Е.А. Букетова, Караганда, Казахстан
https://orcid.org/0009-0000-2620-9767;
Рахимжанова Мира Бейсенбаевна — PhD, ассистент-профессор Школы искусственного интеллекта и науки о данных, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-1328-8109;
Недзьведь Александр Михайлович — д.т.н, доцент, профессор кафедры компьютерных технологий и систем, факультета прикладной математики и информатики, Минск, Беларусь
https://orcid.org/0000-0001-6367-5900.

© Е.Г. Камен, Ж.Т. Есендаулетова, Л.С. Фазылова, М.Б. Рахимжанова, А.М. Недзьведь

Аннотация. В статье рассматривается проблема объективной оценки внимания у детей младшего школьного возраста на основе анализа электроэнцефалографических (ЭЭГ) данных с использованием методов машинного обучения. Актуальность исследования обусловлена необходимостью раннего выявления нарушений внимания и разработки научно обоснованных подходов к психолого-педагогическому сопровождению детей с особенностями развития. В работе проанализированы характерные ЭЭГ-маркеры, связанные с уровнями концентрации и когнитивной нагрузки. На основе выделенных признаков построена модель классификации, включающая предварительную обработку сигналов, извлечение спектральных характеристик и применение нейросетевых алгоритмов. Полученные результаты демонстрируют возможность достоверного разделения состояний «внимание» и «невнимание» с высокой точностью. Проведённое исследование подтверждает эффективность использования ЭЭГ-технологий в сочетании с современными методами анализа данных для объективной оценки внимания и может служить основой для дальнейшей разработки диагностических инструментов в системе специального и инклюзивного образования.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Ключевые слова: ЭЭГ, нейронные сети, анализ внимания, классификация когнитивных состояний; СДВГ, глубокое обучение, CNN–LSTM, обработка биосигналов, оценка когнитивной нагрузки, машинное обучение

Для цитирования: Е.Г. Камен, Ж.Т. Есендаулетова, Л.С. Фазылова, М.Б. Рахимжанова, А.М. Недзьведь (2026). Использование нейронных сетей для объективной оценки внимания у детей по данным ээг // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 216–230. https://doi.org/10.54309/IJICT.2026.25.1.013. каз

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Қазіргі білім беру кеңістігінде балалардың когнитивтік функцияларын, соның ішінде зейіннің деңгейін объективті бағалау ерекше өзектілікке ие. Зейіннің тұрақтылығы оқу жетістіктерін, интеллектуалдық даму қарқынын және мінез-құлықтық бейімделуді анықтайтын негізгі көрсеткіштердің бірі болып саналады. Дегенмен дәстүрлі психодиагностикалық әдістер субъективтілікке бейім, педагогтың немесе психологтың тәжірибесіне тәуелді және нақты уақыт режимінде дәл бағалауды қамтамасыз етпейді (Xue және т.б., 2025; Barry және т.б., 2003).

Электроэнцефалография (ЭЭГ) баланың танымдық жағдайын тіркеудің қолжетімді әрі сенімді нейрофизиологиялық тәсілі ретінде кеңінен қолданылады. ЭЭГ-сигналдардағы ритмдердің (альфа, бета, тета және т.б.) қуаты мен динамикасы зейіннің өзгеруін сезімтал түрде көрсетуге мүмкіндік береді (Lenartowicz және т.б., 2014; Roy және т.б., 2019). Зерттеулер ЭЭГ маркерлерінің назардың бөлінуі, селқостық және когнитивтік жүктеме деңгейлерін ажыратуға жарамды екенін дәлелдеді (Craik және т.б., 2019; Clayton және т.б., 2018).

Соңғы жылдары нейрондық желілер мен терең оқыту модельдері ЭЭГ деректерін өңдеу саласында жоғары нәтижелер көрсетіп келеді. Конволюциялық және рекурренттік желілер уақыттық және спектралдық құрылымдарды тиімді үйреніп, жасанды интеллект негізіндегі дәл диагностиканы қамтамасыз етеді (Craik және т.б., 2019; Roy және т.б., 2019). Балалардың зейінін анықтау бағытында ЭЭГ+ML тәсілдері когнитивтік күйлерді автоматты жіктеуде перспективалы нәтижелер көрсетуде (Schirrmeister және т.б., 2017; Babiloni және т.б., 2019). Соған қарамастан, балалардың ЭЭГ-сигналдары артефактілерге сезімтал, жас ерекшелігіне байланысты өзгермелі және стандартталмаған ортада өткізіледі. Сондықтан сенімді әрі интерпретацияланатын нейрондық модельдер әзірлеу ғылыми тұрғыдан маңызды әрі күрделі мәселе болып қалып отыр. Осы зерттеу ЭЭГ деректері негізінде балалардың зейін деңгейін объективті бағалауға арналған нейрондық желілік модельді ұсынуға бағытталған. Жұмыс нәтижелері инклюзивті және арнайы білім беруде диагностикалық шешімдерді жетілдіруге, деректерге негізделген психологиялық-педагогикалық қолдауды дамытуға мүмкіндік береді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Когнитивтік нейроғылым саласындағы зерттеулер альфа ырғақтарының функционалдық маңызын ерекше атап көрсетеді. Альфа белсенділігі визуалды өңдеу, селективті зейін және ақпаратты сүзгілеу процестерімен тығыз байланысты екені дәлелденген (Clayton және т.б., 2018). Уақыттық қатарларды терең талдау тәсілдері ЭЭГ сигналдарының динамикалық құрылымын түсінуге мүмкіндік береді және жиілік доменіндегі өзгерістердің когнитивтік механизмдермен өзара байланысын анықтайды (Cohen, 2014). Сонымен қатар, заманауи ми–компьютер интерфейстері ЭЭГ негізінде зейін күйін автоматты бағалаудың практикалық жүзеге асырылуын көрсетеді және нейрофизиологиялық мониторингті нақты уақыт режимінде жүргізуге жағдай жасайды (Abirі және т.б., 2020; Mullen және т.б., 2015). Бұл бағыттағы әдістемелік құралдардың дамуы ЭЭГ деректерін тек клиникалық емес, білім беру ортасында да қолдану мүмкіндігін кеңейтуде.

ADHD контекстінде когнитивтік функциялардың гетерогенділігі ерекше назар аудартады, себебі бұзылыстың әртүрлі нейробиологиялық профильдері байқалады (Karalunas & Nigg, 2017). Зерттеулер тета/бета қатынасының өзгеруі мен маңдай аймақтарындағы электрофизиологиялық белсенділіктің ерекшеліктері ADHD-нің ықтимал биомаркерлері болуы мүмкін екенін көрсетеді (Arns және т.б., 2013). ЭЭГ сигналдарын өңдеуде ашық бағдарламалық платформалар мен тәуелсіз компоненттік талдау әдістерін қолдану артефактілерді азайтып, физиологиялық маңызы бар компоненттерді бөліп алуға мүмкіндік береді (Delorme & Makeig, 2004). Терең нейрондық архитектуралар, әсіресе CNN–LSTM үлгілері, уақыттық және кеңістіктік заңдылықтарды біріктіре отырып, зейін күйлерін жоғары дәлдікпен жіктей алатынын көрсетті (Li және т.б., 2021). Осы тұрғыдан алғанда, ЭЭГ сигналдарын терең оқыту модельдерімен біріктіру балалардың зейін деңгейін объективті бағалаудың ғылыми негізделген әрі технологиялық тұрғыдан перспективалы бағыты болып табылады.

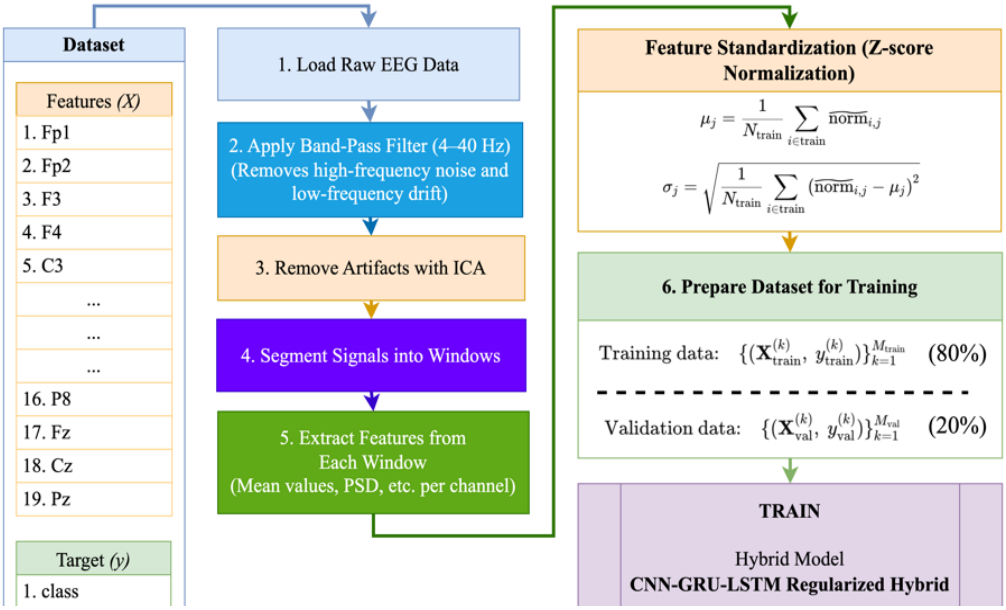
Әдістер мен материалдар.

Ұсынылған алгоритм (Сурет 1) гибридті терең нейрондық желіні кейінгі оқыту үшін электроэнцефалографиялық (ЭЭГ) деректерді дайындаудың толық процесін көрсетеді. Бірінші кезең халықаралық 10-20 схемасына сәйкес 19 электродтан алынған көп арналы уақыттық қатарлар болып табылатын түпнұсқа ЭЭГ сигналдарын жүктеуді қамтиды. Әрі қарай, FIR сүзгісін пайдаланып 4-40 Гц диапазонында жолақты сүзгілеу қолданылады, бұл ықтимал дрейф пен қозғалыстан туындаған жоғары жиілікті шуды (бұлшықет артефактілері, электромагниттік кедергі) және төмен жиілікті ауытқуларды жояды.

Үшінші кезең көз қозғалыстарымен (ЭОГ), бет бұлшықеттерінің кернеуімен және басқа да қажетсіз көздермен байланысты компоненттерді тиімді түрде бөліп алып тастайтын тәуелсіз компоненттік талдауды (ИКА) қолдану арқылы артефактіні жоюды қамтиды. Тазартудан кейін сигналдар белгіленген ұзақтықтағы терезелерге бөлінеді, бұл жаттығу мысалдарының санын көбейтеді және қысқа уақыт аралығында ми белсенділігінің динамикасын анықтауға мүмкіндік береді. Келесі кезең әр терезеден ерекшеліктерді алуды қамтиды. Олар

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

қарапайым статистикалық параметрлерден (орташа мән, дисперсия) спектрлік сипаттамаларға дейін, соның ішінде жиілік диапазонындағы қуат (дельта, тета, альфа, бета, гамма), мидың әртүрлі функционалдық аймақтарының белсенділігін



Сур. 1. ЭЭГ деректерін өңдеу және гибридті модельдерді оқытудың жұмыс процесі.

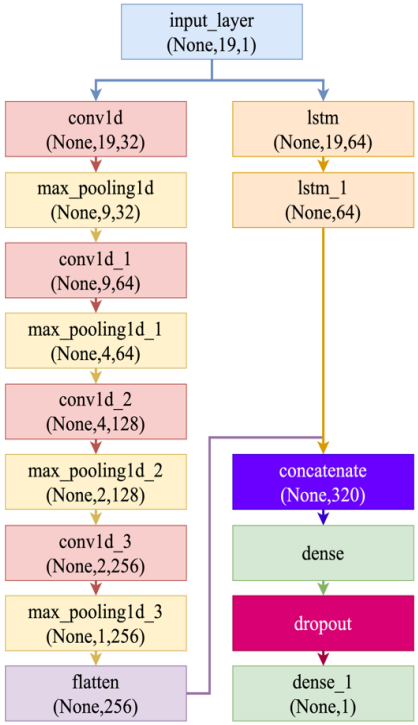
көрсететін спектрлік сипаттамаларға дейін болуы мүмкін. Алынған ерекшеліктер жаттығу жиынтығының параметрлеріне негізделген Z-балл бойынша қалыпқа келтіріледі, бұл модель конвергенциясын жақсартады және арналар арасындағы амплитудалық теңгерімсіздіктің әсерін азайтады.

Соңғы кезеңде жиынтық жаттығу (80 %) және валидация (20 %) бөліктеріне бөлінеді, бұл модельдің жалпылау қабілетін бағалауға мүмкіндік береді. Дайындалған және стандартталған мүмкіндіктер ЭЭГ-ге тән кеңістіктік және уақыттық тәуелділіктерді бір мезгілде алуға қабілетті гибридті CNN-GRU-LSTM моделінің кірісіне беріледі. Бұл алгоритм нейрофизиологиялық жіктеу тапсырмаларында жоғары дәлдікке жету үшін қажетті толық және қайталанатын деректерді дайындау процедурасын қамтамасыз етеді.

2-суретте кіріс деректерінен уақытша және жергілікті ерекшеліктерді алу үшін LSTM және Conv1D біріктіретін гибридті нейрондық желі архитектурасы көрсетілген. Бір жағынан, кіріс тізбегі әрқайсысы 64 нейроннан тұратын екі тізбекті LSTM қабаттарына беріледі, бұл ұзақ мерзімді тәуелділіктерді және сигналдың уақытша контекстін алуға мүмкіндік береді. Екінші жағынан, параллель түрде Conv1D конволюциялық қабаттарының каскады (32, содан кейін 64, 128 және 256 сүзгілер) қолданылады, аралық MaxPooling1D операциялары қолданылады, мұнда әрбір қабат сигналды ұсақ түйіршікті деңгейде өңдейді, жергілікті ерекшеліктерді алады.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Конволюциялық блоктардан өткеннен кейін, шығыс Flatten көмегімен векторға тегістеледі, ал қайталанатын және конволюциялық тармақтардан алынған көріністер Concatenate қабатын пайдаланып біртұтас нысанға біріктіріледі. Содан кейін алынған ерекшелік векторы 128 нейроны бар толық қосылған тығыз қабат арқылы өтеді, бұл модельге кіріс сигналының жалпыланған көрінісін қалыптастыруға көмектеседі. Шамадан тыс сәйкестендіруді азайту үшін жаттығу кезінде кейбір нейрондарды кездейсоқ түрде түсіретін 0,5 ықтималдығы бар Dropout қабаты қолданылады. Соңында, бір нейрон және сигма тәрізді активациясы бар тығыз шығыс қабаты мақсатты класқа тиесілі болу ықтималдығын тудырады, бұл екілік жіктеуді мүмкін етеді.



Сур. 2. ЭЭГ жіктеуіне арналған гибридті CNN-LSTM моделі.

Осылайша, модель уақытша үлгілерді талдау үшін RNN (LSTM) және жергілікті ерекшеліктерді алу үшін CNN (Conv1D) артықшылықтарын пайдаланады, бұл сигналдың жан-жақты көрінісін қамтамасыз етеді және жіктеу дәлдігін жақсартады.

Модель Adam оптимизаторы (learning rate = 0.001) және бинарлық кроссэнтропия (Binary Cross-Entropy Loss) функциясын қолдана отырып оқытылды. Batch size — 32, epoch саны — 50. Ерте тоқтату (Early Stopping) стратегиясы валидациялық шығын тұрақсызданған сәтте қолданылды. Детальды архитектура: CNN блоктарындағы фильтрлер саны — 32, 64, 128, 256; LSTM

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

қабаттары — 2 (64 нейроннан), GRU қабаты — 1 (64 нейрон), Dropout = 0.5. Бұл параметрлер модельдің тұрақты конвергенциясын қамтамасыз етті.

Нәтижелер және талқылау.

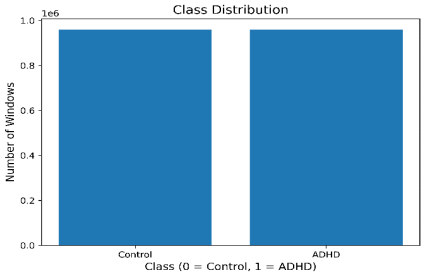
Деректер үш бөлікке бөлінді: оқу (70 %), валидация (15 %) және толық тәуелсіз тест жиыны (15 %). Тест жиыны модельге оқу кезінде берілмеді және жалпылау қабілетін бағалау үшін ғана қолданылды. Мұндай схема overfitting ықтималдығын төмендетіп, модельдің нақты деректерге қолданылуын дәлірек бағалауға мүмкіндік берді.

Бұл зерттеу жобасы Шахед университеті құрастырған және IEEE платформасында жарияланған ашық деректер жиынтығына негізделген. Бұл деректердің негізгі мақсаты - электроэнцефалографиялық (ЭЭГ) сигналдарды талдау негізінде зейін тапшылығы гиперактивтілігінің бұзылуы (СДВГ) бар балалар мен сау балалар арасындағы нейрофизиологиялық айырмашылықтарды зерттеу. Деректер жиынтығы когнитивтік күйлерді объективті бағалау және ЭЭГ жазбаларындағы сипаттамалық патологиялық заңдылықтарды анықтауды автоматтандыруға қабілетті машиналық оқыту алгоритмдерін әзірлеу үшін арнайы жасалған. Үлгіге 7 жастан 12 жасқа дейінгі 121 баланың жазбалары кіреді, олардың 61-іне DSM-IV критерийлеріне сәйкес ЗГГ клиникалық диагнозы қойылған, ал қалған 60 бала бақылау тобын құрады. зейін тапшылығы гиперактивтілігінің бұзылуы диагнозы қойылған балалар дәрілік терапия алып, кем дегенде алты ай бойы Риталин қабылдады. ЭЭГ жазбалары маңдай, орталық, париетальды, париетальды және желке қыртысына орналастырылған 19 электродты пайдаланып, халықаралық 10-20 жүйесін пайдалану арқылы жүргізілді. Әрбір сигнал келесі арналардағы белсенділік мәндерімен көрсетіледі: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, сондай-ақ Fz, Cz және Pz, қатысушының сау балалар тобына немесе ADHD бар балалар тобына жататынын көрсететін жазба идентификаторымен және мақсатты клас белгісімен толықтырылған. ЭЭГ сигналдары құлақ сүйектеріне бекітілген A1 және A2 эталондық электродтарын пайдаланып, 128 Гц дискреттеу жиілігінде жазылды. Тәжірибелік процедураға визуалды когнитивті тест кірді, оның барысында балаларға мультфильм кейіпкерлерінің суреттері көрсетіліп, нысандар санын санау сұралды. Жазу тапсырма аяқталғанға дейін жалғасты, когнитивті реакция динамикасы мен ақпаратты өңдеу жылдамдығын көрсетеді. Деректер кестелік форматта ұсынылған, мұнда әрбір жол бір көп арналы ЭЭГ жазбасына сәйкес келеді және сигнал мәндерін, класты және қатысушы идентификаторын қамтиды. Бастапқы сигналдардың жоғары сапасы, стандартталған жазу хаттамасы және мұқият таңдалған үлгі бұл деректер жиынтығын нейрофизиологиялық зерттеулер және клиникалық шешімдердің дәлдігін арттыруға бағытталған интеллектуалды диагностикалық жүйелерді әзірлеу үшін құнды ресурсқа айналдырады. 3-суретте электроэнцефалографиялық (ЭЭГ) сигнал жазбаларын қамтитын деректер жиынындағы кластардың таралуы көрсетілген.

0-сынып дені сау балаларға (бақылау тобына), ал 1-сынып СДВГ диагнозы қойылған балаларға сәйкес келеді. Деректер теңгерімді, бұл машиналық оқыту

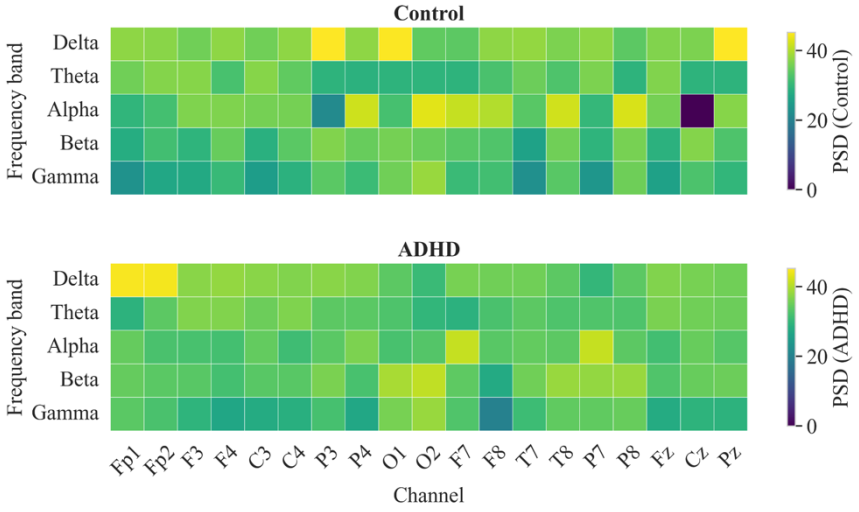
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

моделін дұрыс оқыту үшін маңызды, себебі бұл болжамдардың бір немесе екіншісіне бейімділігін болдырмайды. Бұл теңгерім жіктеу тұрақтылығын арттырады және деректер теңгерімсіздігіне байланысты қателіктер ықтималдығын азайтады.



Сур. 3. ЭЭГ деректер жиынындағы класстардың таралуы.

4-суретте бес жиілік диапазоны үшін - дельта (1–4 Гц), тета (4–8 Гц), альфа (8–13 Гц), бета (13–30 Гц) және гамма (30–45 Гц) үшін қуат спектрінің тығыздығы (PSD) жылу картасы екі топ үшін барлық ЭЭГ сегменттері бойынша орташаланған: бақылау және СДВГ.



Сур. *4.* 19 ЭЭГ арнасы бойынша әрбір топ (бақылау және ADHD) үшін жиілік диапазондары (Дельта, Тета, Альфа, Бета, Гамма) бойынша орташа қуат спектрлік тығыздығы (PSD).

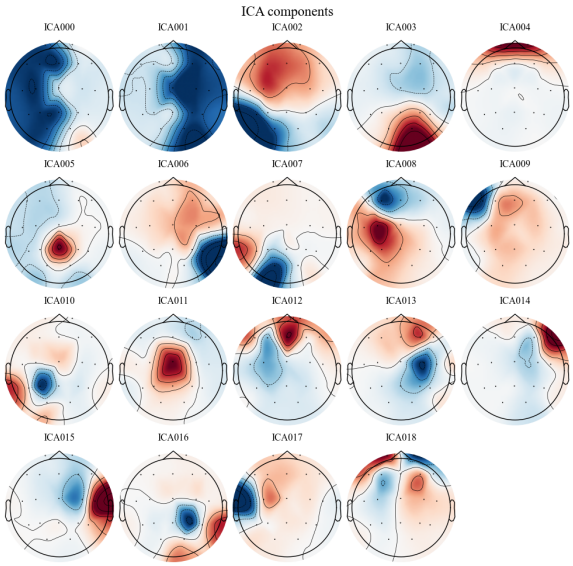
Суреттің жоғарғы бөлігінде сау балалардың нәтижелері көрсетілген (Бақылау). Альфа және бета белсенділігі, әсіресе, маңдай (F3, F4), орталық (C3, C4) және парието-шүйде (P3, P4, O1, O2) аймақтарында айқын көрінеді, бұл тапсырма кезінде теңгерімді когнитивті белсенділік пен тұрақты назар аудару күйін көрсетеді. Pz сияқты кейбір арналар салыстырмалы түрде төмен қуат көрсетеді, бұл жеке вариацияларға немесе сигнал жазу жағдайларына байланысты болуы мүмкін. Тета қуаты көптеген арналарда орташа, айқын шыңдарсыз, бұл

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

когнитивті фокустың тыныштық күйіне тән.

Суреттің төменгі бөлігінде ADHD бар балалар тобы үшін ұқсас жылу картасы көрсетілген. Айтарлықтай айырмашылық - дельта диапазонындағы, әсіресе Fp1 және Fp2 маңдай арналарындағы қуат мәндерінің жоғарылауы, бұл көбінесе префронтальды кортекстің гипоактивациясымен байланысты. Сонымен қатар, ADHD бар балалардағы альфа белсенділігі негізінен уақытша-париетальды аймақтарда (мысалы, F8, T8, P7) байқалады, бұл екіншілік когнитивті тізбектердің компенсаторлық белсенділігін көрсетуі мүмкін. Сондай-ақ, T7, P4 және O2 аймақтарындағы бета және гамма жолақтарындағы белсенділіктің артуы қызығушылық тудырады, бұл жүйке өзгергіштігінің жоғарылауын және ұзақ уақыт зейін қою қиындықтарын көрсетуі мүмкін.

Бұл 7 таңбалы график сау адамның электрофизиологиялық деректеріне қолданылатын тәуелсіз компоненттік талдаудың (ТКТ) нәтижелерін көрсетеді. Әрбір карта көп арналы ЭЭГ жазбасынан алынған тәуелсіз компоненттердің біріндегі белсенділіктің кеңістіктік таралуын көрсетеді. Бұл визуализация әртүрлі белсенділік көздерінің үлестерінің бас терісінде қалай локализацияланғанын көрсетеді. Әрбір компонент үшін физиологиялық ырғақтарға (мысалы, желке аймағындағы альфа ырғағы) немесе артефактілерге (көз қозғалыстары, бұлшықет белсенділігі) сәйкес келуі мүмкін сипаттамалық үлгілер анықталады (Сурет 5).

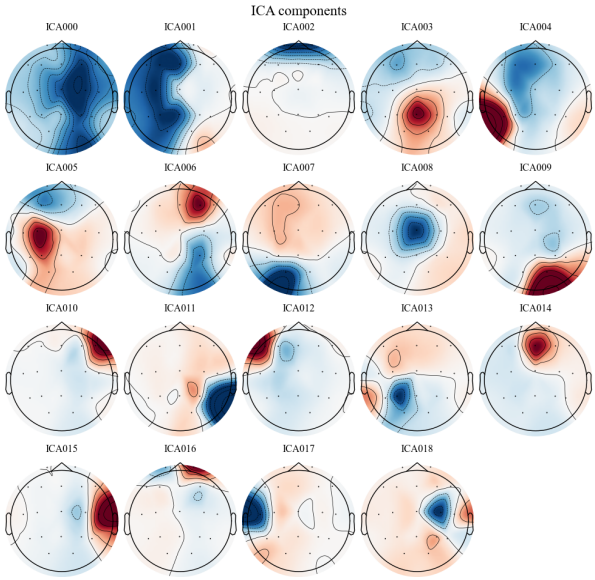


Сур. 5. ЭЭГ ICA компоненттерінің топографиялық карталары (Қалыпты/Басқару класы).

ICA қолдану нақты жүйке процестері туралы ақпарат беретін компонент-терді анықтау және оларды шудан бөлу арқылы ЭЭГ талдауының сапасын жақ-сартады. Сау (Қалыпты/Бақылау) класы жағдайында топографиялық карталар-дың салыстырмалы түрде реттелген таралуы жиі байқалады, айқын аномальды шыңдар немесе патологиялық үлгілерсіз. Мұндай компоненттер әдетте сау попу-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ляцияға тән тұрақты ми ырғақтарын көрсетеді. Нәтижесінде пайда болған дерек-көздерді бөлек визуализациялау зерттеушілер мен клиниктерге әрбір дереккөздің үлесін дәлірек бағалауға мүмкіндік береді, бұл кейіннен артефактіні жоюға және қалыпты жағдайларда когнитивті функцияларды зерттеуге көмектеседі. Бұл 8 таңбалы дисплей зейін тапшылығы гиперактивтілігінің бұзылуы (СГБА) бар баланың электроэнцефалография (ЭЭГ) деректері бойынша жүргізілген тәуелсіз компоненттік талдаудың (ИАК) нәтижелерін көрсетеді. Әрбір топографиялық карта анықталған тәуелсіз компоненттердің бірінің кеңістіктік таралуын көрсетеді, бас терісінде белгілі бір дереккөздің үлесі ең айқын болатын жерді көрсетеді. СГБ көбінесе күрделі немесе ығысқан үлгілері бар компоненттермен сипатталады, бұл зейін мен импульсті басқаруға жауапты нейрондық желілердің қалыптан тыс ұйымдастырылуын көрсетуі мүмкін. Топографиялық карта құрылымы реттелген сау балалардан айырмашылығы, ADHD бар балалар үлкен өзгергіштік немесе тұрақсыз белсенділікпен байланысты көздердің болуын көрсетуі мүмкін (Сурет 6).

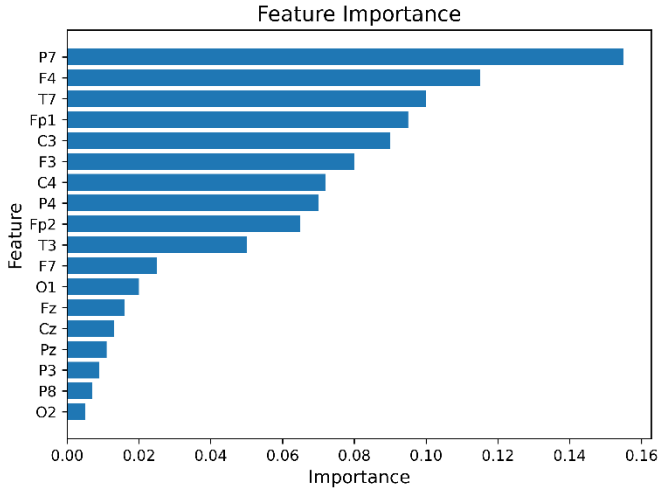


Сур. 6. ЭЭГ (ADHD класы) ICA компонентінің топографиялық карталары.

Мұндай компоненттер өздігінен қозғыштықтың жоғарылауын немесе гиперактивті мінез-құлықпен байланысты артефактілерді басу қиындықтарын көрсетуі мүмкін. ICA шулы сигналдарды (мысалы, көз қозғалыстары немесе бұлшықет артефактілері) физиологиялық маңызды процестерден бөлу арқылы мұндай ерекшеліктерді анықтауды жеңілдетеді. Осылайша, ADHD бар балалардағы ЭЭГ компоненттерінің топографиялық карталарын талдау когнитивті реттеу бұзылыстарының ықтималдығын егжей-тегжейлі түсінуге мүмкіндік береді және осы бұзылыстың биомаркерлері ретінде қызмет етуі мүмкін нақты белсенділік үлгілеріне назар аудару үшін одан әрі зерттеулер жүргізуге мүмкіндік береді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

7-суретте ЭЭГ деректерінің ішкі жиынында оқытылған XGBoostClassifier алгоритмін пайдаланып алынған мүмкіндіктің маңыздылық диаграммасы көрсетілген.



Сур. 7. ЭЭГ деректерінің ішкі жиынында оқытылған XGBoostClassifier алгоритмін пайдаланып алынған мүмкіндіктің маңыздылық диаграммасы көрсетілген.

Бағаналық график P7, F4 және Fp1 электродтарымен байланысты ерекшеліктердің жіктеуге ең көп үлес қосатынын көрсетеді, бұл ADHD бар балалар мен бақылау тобын ажыратуда артқы самай және маңдай ми аймақтарының маңыздылығын көрсетеді. P7-нің жоғары маңыздылығы сенсорлық өңдеуге қатысатын парието-самай аймағының рөлін көрсетеді. F4 және Fp1 маңдай электродтары маңдай аймақтарының атқарушы функциялар мен импульстік бақылауға қатысуын растайды. Сонымен қатар, O2 және P8 сияқты ерекшеліктердің төмен маңыздылығы олардың шектеулі дискриминациялық қабілетін немесе сигналдың ақпараттық арналармен қабаттасуын көрсетеді. Соңғы нәтижелер мидың негізгі аймақтарын анықтауға және ең маңызды ерекшеліктерге назар аудару арқылы одан әрі талдауды оңтайландыруға көмектеседі.

Ұсынылған гибридті CNN–LSTM моделінің тиімділігін объективті бағалау үшін бірнеше базалық эталондық модельдермен салыстырмалы талдау жүргізілді. Салыстыруға классикалық машиналық оқыту модельдері (Random Forest, XGBoost), сондай-ақ терең оқыту модельдері (жеке CNN, жеке LSTM) енгізілді. Әр модель бірдей алдын ала өңделген ЭЭГ ерекшеліктерінде оқытылды. Нәтижелер көрсеткендей, гибридті архитектура барлық метрикалар бойынша (Accuracy, Precision, Recall, F1-score) базалық модельдерден жоғары көрсеткіш көрсетті. Бұл гибридті тәсілдің ЭЭГ сигналдарындағы уақытша және кеңістіктік үлгілерді қатар дұрыс оқуға қабілетті екенін дәлелдейді.

Қорытынды.

Жүргізілген зерттеу электроэнцефалографиялық (ЭЭГ) деректер негізінде балалардың зейін деңгейін объективті бағалау мәселесін заманауи

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

жасанды интеллект әдістері арқылы шешуге бағытталды. Зерттеу нәтижелері нейрофизиологиялық сигналдарды терең нейрондық желілер көмегімен талдау балалардың «зейін» және «зейінсіздік» күйлерін сенімді түрде ажыратуға мүмкіндік беретінін көрсетті. Бұл тәсіл дәстүрлі психодиагностикалық әдістермен салыстырғанда анағұрлым объективті, сандық және қайталанатын бағалау механизмін қамтамасыз етеді. Зерттеу барысында ЭЭГ сигналдарын өңдеудің толыққанды құбыржолы (pipeline) әзірленді. Алдын ала өңдеу кезеңінде 4–40 Гц диапазонында жолақты сүзгілеу қолданылып, төмен жиілікті дрейфтер мен жоғары жиілікті артефактілер жойылды. Тәуелсіз компоненттік талдау (ICA) көмегімен көз қозғалыстары мен бұлшықет белсенділігіне байланысты шуды тиімді бөлу жүзеге асырылды. Сигналдарды терезелеу және Z-балл арқылы қалыпқа келтіру модельдің тұрақты конвергенциясын қамтамасыз етті. Бұл кезеңдердің барлығы ЭЭГ деректерінің табиғи өзгергіштігін төмендетіп, нейрондық модельдің жалпылау қабілетін арттырды. Ұсынылған гибридті CNN–GRU–LSTM архитектурасы ЭЭГ сигналдарының кеңістіктік және уақыттық құрылымдарын бір мезгілде талдауға мүмкіндік берді. Конволюциялық қабаттар локальды спектралдық үлгілерді анықтаса, рекурренттік қабаттар ұзақ мерзімді уақыттық тәуелділіктерді модельдеді. Мұндай біріктірілген тәсіл классикалық машиналық оқыту модельдеріне (Random Forest, XGBoost) және жеке CNN немесе LSTM архитектураларына қарағанда жоғары нәтижелер көрсетті. Барлық негізгі метрикалар бойынша (Accuracy, Precision, Recall, F1-score) гибридті модель басымдық танытты, бұл ЭЭГ сигналдарының күрделі динамикалық табиғатын ескерудің маңыздылығын дәлелдейді. Зерттеу барысында алынған нәтижелер нейрофизиологиялық тұрғыдан да маңызды. Ерекшеліктердің маңыздылық диаграммасы (Feature Importance) бойынша P7, F4 және Fp1 аймақтары жіктеуде шешуші рөл атқарды. Бұл маңдай және парието-уақытша аймақтардың атқарушы функциялар, импульстік бақылау және сенсорлық өңдеумен тығыз байланысты екенін көрсетеді. ADHD бар балаларда дельта диапазонының жоғарылауы және альфа-бета белсенділігінің өзгеруі байқалды, бұл когнитивтік реттеу механизмдерінің ерекшеліктерін айқындайды. Осылайша, ұсынылған модель тек жоғары дәлдік көрсетіп қана қоймай, клиникалық тұрғыдан интерпретацияланатын нәтижелер береді. Деректер жиынының теңгерімділігі (control және ADHD топтарының шамалас көлемі) модельдің бейтарап оқытылуын қамтамасыз етті және жіктеудің бір классқа бейімділігін болдырмады. Сонымен қатар, деректердің оқу, валидация және толық тәуелсіз тест жиындарына бөлінуі overfitting ықтималдығын азайтып, модельдің жалпылау қабілетін объективті бағалауға мүмкіндік берді. Бұл зерттеудің әдіснамалық сенімділігін арттырады.

REFERENCES

Abiri, R., Borhani, S., Sellers, E.W., et al. (2020). A comprehensive review of EEG-based brain–computer interface paradigms // Journal of Neural Engineering. — Vol. 17(4). — Article 041001. https://doi.org/10.1088/1741-2552/ab9875.

Arns, M., Conners, C.K., Kraemer, H.C. (2013). A decade of EEG theta/beta ratio research in ADHD // Journal of Attention Disorders. — Vol. 17(5). — Pp. 374–383. https://doi.org/10.1177/1087054712460087.

Babiloni, C., Del Percio, C., Valenzano, A., et al. (2009). Frontal attentional processes and alpha rhythms.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

— Clinical Neurophysiology. — Vol. 120(10). — Pp. 1880–1890. https://doi.org/10.1016/j.clinph.2009.08.021.

Barry, R.J., Clarke, A.R., Johnstone, S.J. (2003). A review of electrophysiology in attention-deficit/hyperactivity disorder // Clinical Neurophysiology. — Vol. 114(2). — Pp. 184–198. https://doi.org/10.1016/S1388-2457(02)00363-2.

Clayton, M.S., Yeung, N., Cohen, M.X. (2018). The many characters of visual alpha oscillations // European Journal of Neuroscience. — Vol. 48(7). — Pp. 2498–2508. https://doi.org/10.1111/ejn.13747.

Cohen, M.X. (2014). Analyzing Neural Time Series Data: Theory and Practice // MIT Press. https://doi.org/10.7551/mitpress/9609.001.0001.

Craik, A., He, Y., Contreras-Vidal, J.L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review // Journal of Neural Engineering. — Vol.16(3). — Article 031001. https://doi.org/10.1088/1741-2552/ab0ab5.

Delorme, A., Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis // Journal of Neuroscience Methods. — Vol. 134(1). — Pp. 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

Karalunas, S. L., Nigg, J. T. (2017). Heterogeneity and subtyping in ADHD. Annual Review of Clinical Psychology. — Vol. 13. — Pp. 591–618. https://doi.org/10.1146/annurev-clinpsy-032816-045213.

Lenartowicz, A., Loo, S. K. (2014). Use of EEG to diagnose ADHD. Current Psychiatry Reports. — Vol. 16(11). —Pp. 498. https://doi.org/10.1007/s11920-014-0498-0.

Li, X., Zhang, D., Zhang, Y., et al. (2021). Attention detection from EEG using CNN-LSTM architecture. Biomedical Signal Processing and Control. — Vol. 63. — Article 102211. https://doi.org/10.1016/j.bspc.2020.102211.

Mullen, T., Kothe, C., Chi, Y. M., et al. (2015). Real-time neuroimaging and cognitive monitoring using wearable EEG. IEEE Transactions on Biomedical Engineering. — Vol. 62(11). — Pp. 2553–2567. https://doi.org/10.1109/TBME.2015.2481482.

Roy, Y., Banville, H., Albuquerque, I., et al. (2019). Deep learning-based electroencephalography analysis: A systematic review // Journal of Neural Engineering. — Vol. 16(5). — Article 051001. https://doi.org/10.1088/1741-2552/ab260c.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping. — Vol. 38(11). — Pp. 5391–5420. https://doi.org/10.1002/hbm.23730.

Xue, Y., et al. (2025). Applications and interrelationships of brain function detection, brain–computer interfaces, and brain stimulation: A comprehensive review. Cognitive Neurodynamics. — Vol. 19(1). — Pp. 161–180. https://doi.org/10.1007/s11571-025-10341-y.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 231–246
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.014

COMPARATIVE ANALYSIS OF VARIOUS RADIO WAVE PROPAGATION MODELS FOR MOBILE NETWORK COVERAGE PREDICTION

A.Ye. Kulakayeva, Ye.A. Bakhtiyarova, G.T. Jakanova, Sh. Nursultan*
International Information Technology University, Almaty, Kazakhstan.
E-mail: 41362@iitu.edu.kz

Aigul Ye. Kulakayeva — Associate Professor, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University, Almaty, Kazakhstan
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Yelena A. Bakhtiyarova — Candidate of Technical Sciences, Associate Professor, Head of the Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University, Almaty, Kazakhstan
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;
Gaukhar T. Jakanova — Senior Lecturer, Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University, Almaty, Kazakhstan
E-mail: gjakanova@iitu.edu.kz, https://orcid.org/0009-0004-7890-3573;
Shakhdiar Nursultan — International Information Technology University, Almaty, Kazakhstan
41362@iitu.edu.kz, https://orcid.org/0009-0002-9287-7573.

© A.Ye. Kulakavea, Ye.A. Bakhtiyarova, G.T. Jakanova, S. Nursultan

Abstract. This paper presents a comparative analysis of five radio wave propagation models commonly used for predicting mobile network coverage. The simulations were performed in a unified environment using RadioPlanner 3.0 for an urban area of Almaty under identical initial parameters. The comparison was carried out based on coverage maps and the metrics RSRP, reliable service radius, and the impact of terrain and building density. The results indicate that the Okumura-Hata model provides a symmetrical «average» coverage and generally overestimates the service radius in urban environments. The ITU-R P.1546 model delivers a balanced prediction that accounts for antenna height and terrain, making it suitable for rapid preliminary assessments. The Longley-Rice model produces a more physically realistic representation with local shadowing and diffraction zones, although it may slightly overestimate range at

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

higher frequencies. The ITU-R P1812-6 model offers the highest level of detail and accuracy under dense urban conditions, The 3GPP TR 38.901 model accurately reflects radio channel statistics and UMa/UMi scenarios, making it valuable for MIMO and beamforming analysis, though it is less precise in describing terrain-related effects.

Keywords: 4G LTE, 5G NR, radio planning, propagation models, Okumura-Hata, Longley-Rice, ITU-R P.1546- 6, ITU-R P.1812-6, 3GPP TR 38.901, RSRP, urban environment

For citation: A.Ye. Kulakayeva, Ye.A. Bakhtiyarova, G.T. Jakanova, S.Nursultan (2026). Comparative analysis of various radio wave propagation models for mobile network coverage prediction // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 231–246. https://doi.org/10.54309/IJICT.2026.25.1.014. (In Rus.).

Conflict of interest: The authors declare that there is no conflict of interest.

ҰЯЛЫ БАЙЛАНЫС ЖЕЛІЛЕРІНІҢ ҚАМТУ АЙМАҒЫН БОЛЖАУҒА АРНАЛҒАН ӘРТҮРЛІ РАДИОТОЛҚЫН ТАРАЛУ МОДЕЛЬДЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ

А.Е. Кулакаева, Е.А. Бахтиярова, Г.Т. Джаканова, Ш. Нурсултан*
Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: 41362@iitu.edu.kz

Кулакаева Айгуль Ергалиевна — «Радиотехника, электроника және телекоммуникациялар» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Бахтиярова Елена Ажибековна — т.ғ.к, қауымдастырылған профессор, «Радиотехника, электроника және телекоммуникациялар» кафедрасының меңгерушісі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;
Джаканова Гаухар Талгатовна — «Радиотехника, электроника және телекоммуникациялар» кафедрасының сениор-лекторы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: gjakanova@iitu.edu.kz, https://orcid.org/0009-0004-7890-3573;
Нурсултан Шахдиар — Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
41362@iitu.edu.kz, https://orcid.org/0009-0002-9287-7573.

© А.Е. Кулакаева, Е.А. Бахтиярова, Г.Т. Джаканова, Ш. Нурсултан

Аннотация. Бұл мақалада мобильді желінің қамтуын болжау үшін қолданылатын бес радиотолқын таралу моделінің салыстырмалы талдауы

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ұсынылған. Модельдеу Алматы қалалық аймағы үшін RadioPlanner 3.0 бағдарламалық ортасында бірдей бастапқы параметрлерді қолдана отырып жүргізілді. Салыстыру қамту карталары мен RSRP метрикасына, сенімді қызмет көрсету ауқымына және жер бедері мен ғимараттардың әсеріне негізделген. Нәтижелер Окумура-Хата моделі симметриялы «орташа» қамтуды қамтамасыз ететінін және әдетте қалалық жерлерде диапазонды асыра бағалайтынын көрсетті. ITU-R P.1546 моделі антенна биіктігі мен жер бедерін ескере отырып, теңдестірілген болжам жасайды, бұл оны алдын ала жылдам бағалауға жарамды етеді. Longley–Rice моделі жергілікті көлеңкелермен және дифракциялық аймақтармен физикалық тұрғыдан шынайы көріністі қалыптастырады, бірақ жоғары жиіліктерде ол диапазонды сәл асыра бағалауы мүмкін. ITU-R P.1812-6 тығыз қоныстанған аудандарда ең үлкен егжей-тегжейлілік пен дәлдікті қамтамасыз етеді. 3GPP TR 38.901 моделі радиоарна статистикасын және UMa/UMi сценарийлерін дәл көрсетеді және MIMO/сәуле түзуші талдау үшін пайдалы, бірақ жергілікті жердің әсерлерін онша дәл сипаттамайды.

Түйін сөздер: 4G LTE, 5G NR, радио жоспарлау, таралу модельдері, Окумура–Хата, Лонгли–Райс, ITU-R P.1546-6, ITU-R P.1812-6, 3GPP TR 38.901, RSRP, қалалық аймақтар

Дәйексөздер үшін: А.Е. Кулакаева, Е.А. Бахтиярова, Г.Т. Джаканова, Ш. Нурсултан (2026). Ұялы байланыс желілерінің қамту аймағын болжауға арналған әртүрлі радиотолқын таралу модельдерінің салыстырмалы талдауы // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 25. Б. 231–246. https://doi.org/10.54309/IJICT.2026.25.1.014. (Орыс тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАЗЛИЧНЫХ МОДЕЛЕЙ РАСПРОСТРАНЕНИЯ РАДИОВОЛН ДЛЯ ПРОГНОЗИРОВАНИЯ ПОКРЫТИЯ СЕТЕЙ МОБИЛЬНОЙ СВЯЗИ

А.Е. Кулакаева, Е.А. Бахтиярова , Г.Т. Джаканова , Ш. Нурсултан*
Международный университет информационных технологий, Алматы, Казахстан.
E-mail: 41362@iitu.edu.kz

Кулакаева Айгуль Ергалиевна — ассоциированный профессор кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий, Алматы, Казахстан
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Бахтиярова Елена Ажибековна — к.т.н., ассоциированный профессор, зав. кафедрой «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий, Алматы, Казахстан
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Джаканова Гаухар Талгатовна — сениор-лектор кафедры «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий, Алматы, Казахстан
E-mail: gjakanova@iitu.edu.kz, https://orcid.org/0009-0004-7890-3573
Нурсултан Шахдиар — Международный университет информационных технологий, Алматы, Казахстан
41362@iitu.edu.kz, https://orcid.org/0009-0002-9287-7573.

© А.Е. Кулакаева, Е.А. Бахтиярова, Г.Т. Джаканова, Ш. Нурсултан

Аннотация. В работе проведен сравнительный анализ пяти моделей распространения радиоволн, используемых для прогнозирования покрытия сетей мобильной связи. Моделирование осуществлялось в единой программной среде RadioPlanner 3.0 для городской территории г. Алматы с одинаковыми исходными параметрами. Сравнение проводилось по картам покрытия и метрикам RSRP, радиусу уверенного обслуживания и степени влияния рельефа местности и застройки. Результаты показали, что модель Okumura–Hata обеспечивает «усредненное» покрытие и, как правило, завышает радиус в городской среде. Модель ITU-R P.1546 дает сбалансированный прогноз с учетом высот антенн и рельефа, что делает ее подходящей для быстрой предварительной оценки. Модель Longley–Rice формирует более физически реалистичную картину с локальными тенями и дифракционными зонами, однако на высоких частотах может незначительно завышать дальность. ITU-R P.1812-6 обеспечивает наибольшую детальность и точность в условиях плотной застройки. Модель 3GPP TR 38.901 корректно отражает статистику радиоканала и сценарии UMa/UMi, полезна для анализа MIMO, beamforming, но менее точно описывает локальные эффекты рельефа.

Ключевые слова: 4G LTE, 5G NR, радиопланирование, модели распространения, Okumura–Hata, Longley–Rice, ITU-R P.1546-6, ITU-R P.1812-6, 3GPP TR 38.901, RSRP, городская застройка

Для цитирования: А.Е. Кулакаева, Е.А. Бахтиярова, Г.Т. Джаканова, Ш. Нурсултан (2026). Сравнительный анализ различных моделей распространения радиоволн для прогнозирования покрытия сетей мобильной связи// Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 231–246. https://doi.org/10.54309/IJICT.2026.25.1.014. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение.

Развитие современных телекоммуникационных технологий и быстрое внедрение сетей пятого поколения (5G) предъявляют высокие требования к точности радиопланирования и прогнозирования покрытия. Качество обслуживания пользователей, эффективность размещения базовых станций

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

и рациональное использование частотного ресурса напрямую зависит от правильного выбора модели распространения радиоволн. В условиях плотной городской застройки радиоволны подвержены сложным многолучевым эффектам, дифракции, отражениям и затуханиям, что значительно усложняет задачу расчета радиопокрытия и требует использования адекватных моделей, учитывающих реальные физические процессы.

Точность прогнозирования радиопокрытия зависит от того, насколько правильно модели учитывают реальные условия распространения сигнала. Исследования показали, что игнорирование неровностей рельефа приводит к искажению расчетов, особенно при использовании коммерческих инструментов радиопланирования. Эксперименты подтвердили, что форма рельефа влияет на уровень затухания в диапазоне 900 МГц–24 ГГц, что делает необходимым учет этих факторов при калибровке моделей для систем 5G и 6G (Soo и др., 2025).

Эмпирические модели CI и ABG остаются основными инструментами для оценки потерь на пути распространения радиосигналов. Их точность зависит от частоты и окружающей среды. При низких частотах лучше работает CI, а при высоких частотах стабильнее проявляет себя ABG (Erunkulu и др., 2024).

В ряде исследований предлагаются методы оптимизации существующих моделей. Например, использование генетического алгоритма позволило сократить среднеквадратичную ошибку расчетов более чем на 90 %, что подтверждает эффективность эволюционных подходов в проектировании гетерогенных сетей 5G (Anwar Faiz и др., 2024).

Перспективным направлением является разработка гибридных моделей, которые объединяют физические принципы и машинное обучение. Применение метода главных компонент (PCA) и полиномиальной регрессии позволило снизить стандартное отклонение ошибки прогноза на 3,3–6,1 дБ, что подтверждает потенциал искусственного интеллекта в радиопланировании (Juang, 2022).

Обзор текущих исследований указывают на то, что развитие моделей потерь пути движется в сторону интеллектуальных и гибридных подходов, включая использование терагерцового диапазона и технологий отражающих поверхностей (Wang и др., 2023).

Для миллиметровых волн особенно важны исследования, учитывающие климат, рельеф и материалы зданий. В частности, моделирование в 4 городах Нигерии (28–73 ГГц) показало, что модель ABG обеспечивает наивысшую точность, хотя результаты варьируется в зависимости от географии и частоты (Afape и др., 2024).

Классические эмпирические модели, такие как Okumura–Hata (Hata и др., 2013), зарекомендовали себя как простые и удобные инструменты для оценки зоны действия базовых станций. Однако их упрощенный характер и без учета рельефа и плотности застройки могут привести к переоценке радиуса уверенного приема. Более сложные физические и гибридные модели, например Longley–Rice (NTIA Report 82–100 и др., 1982), ITU-R P.1546–6 (МСЭ-R P.1546–6 и др., 2019) и

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ITU-R P.1812–6 (МСЭ-R P.1812–6 и др., 2021), предлагают более точные прогнозы, учитывая геометрию местности, параметры антенн, отражения и атмосферные эффекты. Тем не менее их использование требует большего объема входных данных и вычислительных ресурсов.

Модель 3GPP TR 38.901 (*3GPP TR 38.901 V17.0.0.* и др., 2023), разработанная консорциумом 3GPP, занимает особое место среди современных подходов к анализу радиоканалов в системах LTE Advanced и 5G NR. В отличие от традиционных моделей, она учитывает сценарии городской макро- и микросоты (Urban Macro и Urban Micro), многолучевые пути, а также технологии MIMO и beamforming.

Выбор адекватной модели распространения радиоволн является одной из важных задач при проектировании сетей мобильной связи. Недостаточная оценка влияния рельефа и застройки может привести к ошибкам в прогнозировании покрытия и ухудшению качества обслуживания. В то же время чрезмерное усложнение модели может привести к увеличению времени расчетов и затрат на проектирование.

Цель данного исследования – провести сравнительный анализ различных моделей распространения радиоволн: Okumura–Hata, Longley–Rice, ITU-R P.1546–6, ITU-R P.1812–6 и 3GPP TR 38.901. Это позволит определить их точность и применимость для прогнозирования покрытия сетей 4G (LTE) и 5G.

Для достижения поставленной цели были поставлены следующие задачи:

- провести моделирование радиопокрытия в программном комплексе Radioplaner с одинаковыми исходными параметрами для диапазонов 900 и 3500 МГц;

- проанализировать особенности формирования зон покрытия для каждой модели и выявить различия в их пространственном распределении;

- оценить влияние рельефа, застройки и частоты на точность прогнозирования на примере г. Алматы;

- определить оптимальные области применения каждой модели при проектировании и оптимизации сетей 4G/5G.

Научная новизна работы заключается в комплексном сравнении пяти распространенных моделей распространения радиоволн, выполненном в единой программной среде и при идентичных условиях моделирования. Это позволяет объективно оценить различия между эмпирическими, физическими и отраслевыми подходами.

Практическая значимость исследования заключается в возможности выбора наиболее адекватной модели для конкретного этапа радиопланирования.

Материалы и методы.

Исследование было проведено с использованием программного комплекса Radioplaner 3.0, который предназначен для проектирования и оптимизации сетей мобильной связи. Программа поддерживает модели Okumura–Hata, Longley–Rice, ITU-R P.1546–6, ITU-R P.1812–6 и 3GPP TR 38.901. В качестве исходных данных

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

используется цифровая модель рельефа и застройки с разрешением примерно 30 м, созданная на основе OpenStreetMap и Global Forest Change. RadioPlanner 3.0 позволяет рассчитывать уровни RSRP, PL, отношение C/(I + N) и вероятность покрытия. Тип местности определен как «усредненная городская застройка» (Urban), где параметры «clutter» (препятствий) задавались на основе комбинации данных OpenStreetMap и Global Forest Change, что соответствует застройке средней этажности с высотой зданий в диапазоне 15–25 метров.

Также реализована функция отображения профиля трассы, анализа высотных препятствий и экспорта результатов в форматы KMZ, PNG и GeoTiff. Применение параллельных вычислений обеспечивает высокую скорость моделирования даже при наличии множества базовых станций. Все расчеты выполнялись для сетей 4G (LTE) и 5G при фиксированных параметрах передающих и приемных устройств в условиях городской застройки г. Алматы. Исходные данные и параметры моделирования в Radioplaner приведены в таблице 1.

Таблица 1 – Исходные параметры для моделирования радиопокрытия

Параметр	Значение
Частота LTE, МГц	900
Частота 5G, МГц	3500
Высота подвеса антенны базовой станции, м	30
Высота мобильного устройства (пользовательского), м	1,5
Мощность передатчика, дБм	40
Тип местности	Городская (Urban)
Радиус моделирования, км	2

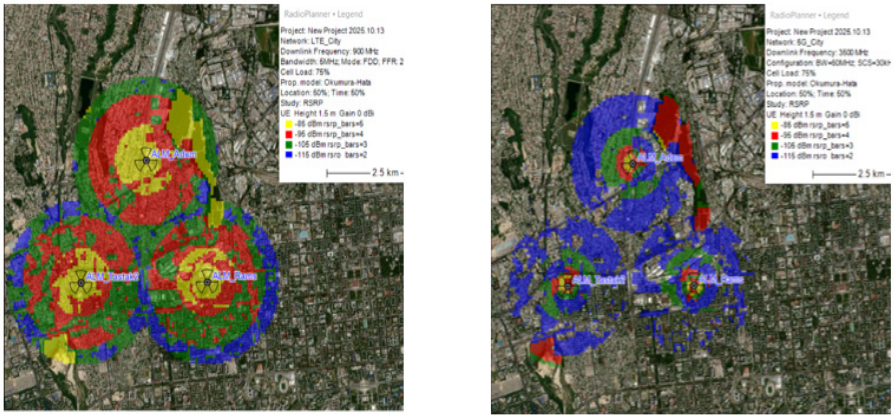
Каждая из выбранных моделей была применена к одному и тому же участку городской территории с фиксированными координатами базовой станции. Это позволило оценить влияние используемого математического подхода на форму и площадь зоны покрытия, а также на устойчивость сигнала в пределах городской застройки. Особое внимание уделялось сравнению результатов в диапазонах 900 МГц и 3500 МГц, поскольку различие в длине волны и механизмах распространения определяют характер затухания и глубину проникновения сигнала.

Эмпирические модели (Okumura–Hata, ITU-R P.1546–6) продемонстрировали предсказуемое усредненное распределение сигнала, что делает их удобными для первичных расчетов и быстрой оценки радиуса обслуживания. В то же время физические и гибридные модели (Longley–Rice, ITU-R P.1812–6, 3GPP TR 38.901) обеспечили более детализированное отражение влияния рельефа, плотности застройки и высотных препятствий. Сравнение этих моделей дало возможность определить, какие из них лучше подходят для точного проектирования сетей 4G и 5G в условиях плотной городской среды, где неоднородность местности и множественные отражения существенно влияют на реальное распределение уровня сигнала.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Результаты и обсуждение.

Модель Okumura–Hata. Результаты моделирования по эмпирической модели Okumura–Hata представлены на рисунке 1: (а) для LTE 900 МГц, (б) для 5G 3500 МГц. Модель показывает радиально-симметричную структуру покрытия, где зоны уверенного приема формируют почти идеальные покрытия вокруг базовых станций. На карте преобладают крупные области желтого и красного цвета, соответствующие уровням сигнала выше –85 дБм, которые плавно переходят в зеленые и синие зоны по мере удаления от центра соты (снижается уровень сигнала).



а) 900 МГц (LTE) б) 3500 МГц (5G)

Рис. 1. Модель Okumura–Hata

Такое распределение свидетельствует о том, что модель не учитывает влияние рельефа, плотности застройки и локальных преград, что приводит к идеализированному покрытию. Отсутствие выраженных неоднородностей и «теневых» зон подтверждает, что расчет основан на усредненных эмпирических формулах.

Расчет потерь на пути распространения радиосигналов проводился с использованием зависимости Okumura–Hata для городской местности:

$$L = 69.55 + 26.16\,log_{10}(f) - 13.82log_{10}(h_b) - a(h_m) + [44.9 - 6.55\,log_{10}(h_b)]\,log_{10}(d),$$

$$a(h_m) = [1.1\,log_{10}(f) - 0.7]\,h_m - [1.56\,log_{10}(f) - 0.8],$$

где L - потери на трассе (дБ), f - частота (МГц), h_b и h_m - высоты антенн базовой и абонентской станций (м), d - расстояние между ними (км).

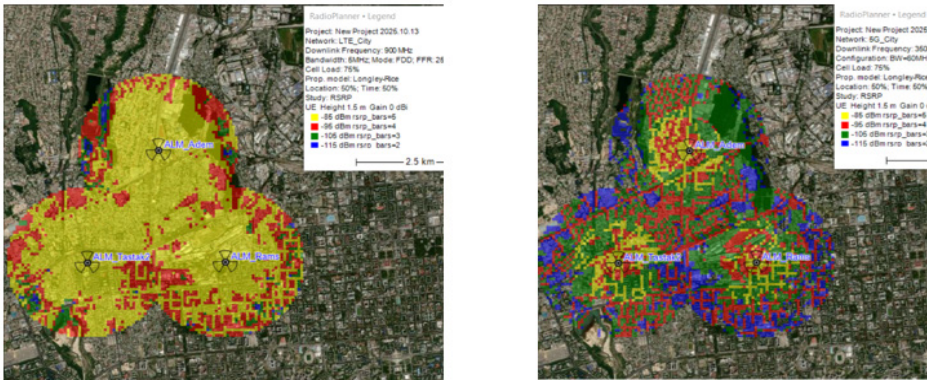
Сравнение результатов для различных частот показало ожидаемую зависимость. При частоте 900 МГц зона уверенного приема охватывает практически весь исследуемый радиус (до 2 км), обеспечивая устойчивое покрытие. Однако при переходе к 3500 МГц радиус уверенного приема значительно сокращается, а зона синего цвета (уровень ниже –105 дБм) расширяется. Это показывает о

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

возрастании потерь на распространение при увеличении частоты и снижение способности сигнала проникать сквозь препятствия.

Также стоит отметить, что расчет для 5G проводился по модели Okumura–Hata в диапазоне 150–1500 МГЦ, как и для 4G. Однако это выходит за пределы ее прменимости, что может привести к искажению оценкм потерь и радиуса покрытия.

Модель Longley–Rice. На рисунке 2 представлено покрытие (а) для LTE 900 МГц, (б) для 5G 3500 МГц. Расчет выполнены по модели Longley–Rice. В отличие от модели Okumura–Hata, форма зон покрытия здесь менее правильная и более «размытая», что отражает физическую природу расчетов. Данная модельь учитывает кривизну Земли, а также отражения и дифракцию, благодаря чему покрытие выглядит более реалистично и соответствует тому, что наблюдается на практике.



а) 900 МГц (LTE) б) 3500 МГц (5G)

Рис. 2. Модель Longley–Rice

Для частоты 900 МГц зона уверенного приема обладает достаточно широкой площадью и плавными переходами уровней сигнала. В отличие от этого, при частоте 3500 МГц зона значительно сужается, а доля слабых участков (зеленых и синих) увеличивается. Карта сигнала выглядит неоднородно, появляются локальные провалы сигнала и вытянутые области, что обусловлено влиянием рельефа и городской застройки.

Модель Longley–Rice в целом предлагает более физически обоснованное описание покрытия и подходит для оценки воздействия рельефа и атмосферы. Однако при высоких частотах она может немного завышать дальность уверенного приема.

Расчет потерь выполнен с использованием следующего уравнения:

$$L = L_{fs} + L_{diff} + L_{scatt},$$

где L_{fs} – потери свободного пространства, L_{diff} – потери из-за дифракции на

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

рельефе, а L_{scatt} – потери, вызванные тропосферным рассеянием и переотражениями.

В обобщенном виде итоговые потери определяются следующим образом:

$$L_{tot} = L_b + A_t + A_d + A_s,$$

где A_t – влияние рельефа, A_d – дифракция, A_s – потери из-за атмосферных условий.

Модель ITU-R P.1812. На рисунке 3 представлены результаты расчетов по рекомендации ITU-R P.1812–6 для диапазонов 900 и 3500 МГц. Эта модель демонстрирует наиболее реалистичное радиопокрытие среди всех рассмотренных подходов. Она учитывает такие факторы, как рельеф плотность застройки, дифракцию и переотражения, что наглядно видно по «пятнистому» распределению уровней сигнала.



а) 900 МГц (LTE) *б) 3500 МГц (5G)*

Рис. 3. Модель ITU-R P.1812

Для частоты 900 МГц зона уверенного приема имеет сложную форму. Сигнал сохраняется в открытых пространствах, но значительно ослабевает в тени зданий и за препятствиями. Причастоте 3500 МГц влияние городской застройки усиливаются, возникают многочисленные области с низким уровнем RSRP (ниже –105 дБм), особенно в плотных кварталах. Карта выглядит детализированной и точно отражает распределение сигнала в городской среде.

Данная модель является реалистичной и описывает распространение радиоволн «из пункта в зону», учитывая дифракцию, отражения и тропосферные эффекты.

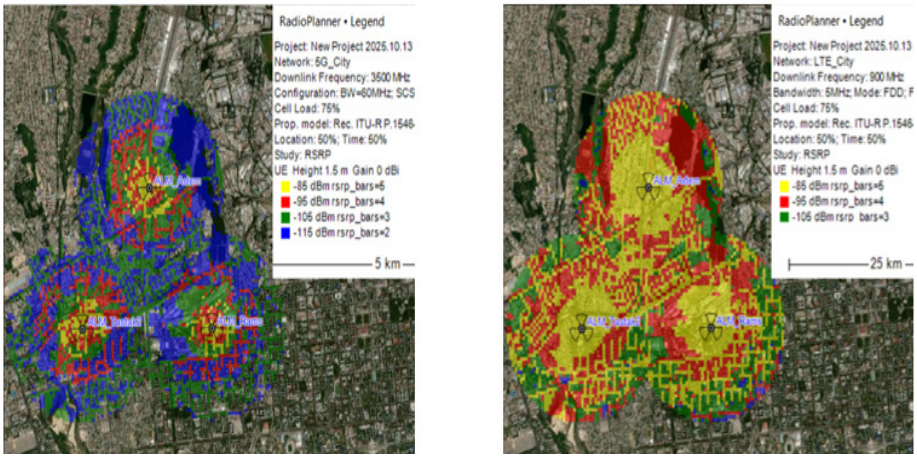
$$L_{tot} = L_b + L_d + L_{clutter} + L_{corr},$$

где L_b– потери свободного пространства; L_d – потери на дифракцию; $L_{clutter}$ – поправка на застройку; L_{corr} – коррекция по рельефу.

Модель ITU-R P.1546. Рисунок 4 показывает результаты расчетов по

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

модели ITU-R P.1546–6 для частот 900 и 3500 МГц. В отличие от традиционных эмпирических подходов, эта модель учитывает влияние рельефа, высоту антенн и тип местности, что возволяет более точно оценить распределение сигнала.



а) 900 МГц (LTE) б) 3500 МГц (5G)

Рис. 4. Модель ITU-R P.1546

При частоте 900 МГц карта демонстрирует обширную зону уверенного приема (до –85 дБм), плавные переходы и минимальное количество теневых участков. На частоте 3500 МГц радиус покрытия уменьшается, а доля ослабленных зон (ниже –105 дБм) возрастает, особенно в периферийных и застроенных районах. Однако контуры покрытия остаются достаточно сглаженными, что указывает на сохранение эмпирического характера модели.

Эта модель основана на эмпирических кривых, которые связывают напряженность поля с дальностью распространения в диапазоне от 30 до 4 ГГц. Для расчета эквивалентных потерь применяется формула:

$$L_b = 139{,}3 - E + 20log_{10}(f),$$

где L_b – эквивалентные потери, дБ; E – напряженность поля, дБ(мкВ/м); f – частота, МГц.

В целом, ITU-R P.1546–6 показывает сбалансированные результаты. Она более точна, чем модель Okumura-Hata, и при этом проще в вычислениях, чем ITU-R P.1812-6. Эта модель подходит для предварительного планирования городских сетей LTE и 5G, обеспечивая разумный компромисс между скоростью расчета и реалистичностью прогноза.

Модель 3GPP TR 38.901. На рисунке 5 показаны результаты расчетов по модели 3GPP TR 38.901 для частот 900 и 3500 МГц. Эта модель разработана консорциумом 3GPP специально для современных систем LTE Advanced и 5G NR, учитывая особенности многолучевого распространения, MIMO и beamforming.

При частоте 900 МГц (рисунок 5, а) карта показывает равномерное

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

покрытие с практически однородным уровнем сигнала и без заметных провалов.



а) 900 МГц (LTE) б) 3500 МГц (5G)

Рис. 5. Модель 3GPP TR 38.901

На частоте 3500 МГц (рисунок 3, б) зона уверенного приема значительно уменьшается, но сохраняет четкую структуру соты, соответствующую сценариям Urban Macro и Urban Micro.

Эта модель используется для расчета потерь в современных сетях LTE и 5G и различает сценарии прямой (LOS) и непрямой (NLOS) видимости:

$$PL_{LOS} = 28.0 + 22log_{10}(d_{3D}) + 20\,log_{10}(f_c),$$

$$PL_{NLOS} = 13.54 + 39.08log_{10}(d_{3D}) + 20\,log_{10}(f_c) - 0.6(h_{UE} - 1.5),$$

где d_3 – трехмерное расстояние между антеннами, м; f_c – частота, ГГц; h_{UE} – высота пользовательского устройства, м.

Модель 3GPP TR 38.901 наиболее точно отражает реальные условия распространения для сетей 5G, включая эффекты отражений, затенения и многолучевого распространения.

В отличие от моделей ITU-R, 3GPP TR 38.901 не стремится точно описывать местные особенности рельефа или застройки, вместо этого она отражает усредненные статистические характеристики радиоканала. Это делает модель подходящей для оценки производительности сетей 5G, анализа эффективности антенн и технологий пространственного формирования лучей, однако ее точность снижается при прогнозировании мелкомасштабных эффектов в плотной городской среде.

Для наглядного сравнения особенностей пяти рассмотренных моделей

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

было проведено их оценивание по таким критериям: типу, степени учета рельефа и застройки, реализму прогнозов и рекомендованной области применения. Итоговые характеристики представлены в таблице 2.

Таблица 2 – Сравнительная анализ моделей распространения радиоволн для прогнозирования покрытия мобильных сетей связи

№	Модель	Учет рельефа	Учет	Преимущества	Ограничения	Рекомендуемая область применения
1	Okumura–Hata	нет	нет	Простота, быстрые расчеты	Завышает радиус, не учитывает среду	Предварительные оценки радиуса покрытия
2	Longley–Rice (ITM)	да		Учитывает отражения и дифракцию	Чувствительна к параметрам среды	Пригородные и сельские зоны
3	ITU-R P.1546-6	да		Учитывает высоты антенн и рельеф	Ограниченная детализация в городе	Предварительное планирование городских сетей
4	ITU-R P.1812-6	да	да	Наиболее точное моделирование городской среды	Требует детальных данных и вычислений	Детальное радиопланирование LTE/5G
5	3GPP TR 38.901	нет		Учитывает MIMO, beamforming, сценарии UMa/ UMi	Не отражает локальные эффекты	Оценка производительности сетей 5G

Из таблицы 2 видно, что с усложнением математической модели увеличивается точность прогнозирования, однако при этом возрастают требования к исходным данным и вычислительным ресурсам.

Таким образом, оптимальная стратегия радиопланирования включает комбинированный подход. На этапе предварительной оценки следует использовать ITU-R P.1546–6, а для детальной оптимизации и валидации параметров – ITU-R P.1812–6. Для анализа сетевых сценариев и работы антенн в системах 5G NR рекомендуется применять 3GPP TR 38.901.

Проведенное моделирование показало, что результаты различных моделей значительно различаются, особенно при переходе от диапазона 900 МГц (LTE) к 3500 МГц (5G). На низких частотах сигнал распространяется дальше и меньше зависит от застройки, тогда как на высоких частотах зона уверенного приема резко сокращается, и любое препятствие, такое как здание или перегиб рельефа местности приводит к заметному ослаблению сигнала.

Модель Okumura–Hata предлагает наиболее простую и «идеальную» картину. Покрытие выглядит симметричным, как круги на воде, без учета рельефа и зданий, из-за чего радиус уверенного приема всегда завышен. Эта модель годится, подходит для быстрого предварительного определения зоны действия базовой станции, но для точных расчетов в городской среде она неэффективна.

Модель Longley–Rice, напротив, учитывает физику распространения, отражение, дифракцию и влияние атмосферы. В результате карта выглядит более

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

реалистична. Появляются вытянутые зоны и участки с пониженным уровнем сигнала. Эта модель особенно полезна для пригородов и открытых пространств, где рельеф местности играет значительную роль.

ITU-R P.1546–6 представляет собой нечто среднее. Она учитывает высоту антенн и особенности местности, оставаясь при этом достаточно быстрой и простой. Покрытие выглядит сглаженным, но ближе к реальности. Эту модель удобно использовать на раннем этапе проектирования, чтобы быстро оценить варианты размещения базовых станций.

Самой точной является модель ITU-R P.1812–6. Она учитывает рельеф, застройку, дифракцию и отражения, поэтому карта покрытия получается очень детализированной и «живой», видны участки тени за зданиями и реальные контуры распространения сигнала. Именно эта модель дает наиболее близкие к реальности результаты, особенно в условиях плотной городской застройки.

Современная модель 3GPP TR 38.901 ориентирована не столько на геометрию распространения, сколько на технологии 5G MIMO, beamforming, сценарии Urban Macro и Urban Micro. Она описывает радиоканал в среднем, статистически, что делает карту покрытия более равномерной. Однако на практике это не всегда так, поскольку мелкие препятствия и плотная застройка этой модели не учитываются.

Можно сказать, что точность модели возрастает вместе с увеличением ее сложности. Простые модели быстро дают общий результат, но не учитывают реальные условия. Более сложные модели, особенно ITU-R P.1812–6 позволяют увидеть реальную картину покрытия, но требуют больше данных и времени на расчеты.

На основе данного анализа можно сделать следующие выводы:

1. На этапе предварительного планирования можно использовать модель Okumura–Hata или ITU-R P.1546–6 для оценки границ зон покрытия. Для детальной оптимизации и уточнения покрытия следует применять ITU-R P.1812–6.

2. Если с целью оценить работу современных технологий, таких как MIMO или beamforming, лучше подойдет 3GPP TR 38.901.

Таким образом, ни одна модель не является универсальной. Каждая решает свою задачу, и грамотный инженер должен выбирать модель в зависимости от конкретного сценария. Только комбинированный подход, при котором простые модели служат основой, а точные подтверждают расчеты, позволяет получить реалистичный прогноз покрытия и избежать ошибок в проектировании сети.

Заключение.

Проведенное исследование продемонстрировало различия в поведении различных моделей распространения радиоволн и их пригодность для проектирования сетей мобильной связи нового поколения. Результаты расчетов показали, что даже при одинаковых условиях моделирования результаты существенно различаются, каждая модель «видит» радиосреду по-своему.

Основной результат заключается в том, что модели, основанные на

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

рекомендациях МСЭ, особенно ITU-R P.1812–6, дают наиболее точное описание реальной картины радиопокрытия. Они учитывают не только высоту антенн и рельеф, но и сложную структуру городской застройки, что критически важно для сетей 4G и 5G. Более простые подходы, такие как Okumura–Hata и ITU-R P.1546–6, остаются полезными для предварительных расчетов и быстрого анализа вариантов размещения базовых станций. Модель 3GPP TR 38.901 обеспечивает связь с современными технологиями передачи данных и может использоваться для оценки производительности систем с MIMO и beamforming, где приоритетом являются характеристики радиоканала, а не форма покрытия.

Практическая значимость работы заключается в том, что полученные результаты позволяют инженерам-проектировщикам осознанно выбирать модель под конкретную задачу, будь то первичное проектирование, оптимизация покрытия или анализ технологической эффективности сети.

Кроме того, проведенное сравнение способствует формированию универсального подхода к планированию. Использование простых моделей для быстрых расчетов и уточнение прогнозов с помощью более точных физических моделей.

Важно отметить, что в рамках данной работы проводилось сравнительное исследование теоретических моделей в программной среде. Валидация полученных данных путем сопоставления с реальными полевыми измерениями (drive-test) в г. Алматы является темой дальнейших исследований. Это позволит установить эталонную модель для данного региона, в то время как текущая работа фокусируется на выявлении системных различий в математических подходах.

REFERENCES

Afape J.O., Willoughby A.A., Sanyaolu M.E., Obiyemi O.O., Moloi K. & Dairo O.F. (2024). Path loss modelling of mmwave outdoor propagation for 5G mobile systems at 28, 38, 60, and 73 GHz in four Nigerian cities. // Discover Applied Sciences. — Vol. 6(10). — Pp. 495. 10.1007/s42452-024-06171-y

Anwar Faizd Osmanc, Mohamad Huzaimy Jusohd, L. L. C. M. R. C. S. A. W. (2024). Path Loss Model Optimization In An Urban Environment Using Genetic Algorithm // *International Journal of Intelligent Systems and Applications in Engineering*. — Vol. 12(3). — Pp. 893–900. https://ijisae.org/index.php/IJISAE/article/view/5369.

Erunkulu O. O., Zungeru A. M., Thula I. G., Lebekwe C. & Mosalaosi M. (2024). A comparative analysis of alpha-beta-gamma and close-in path loss models based on measured data for 5G mobile networks // Results in Engineering. — Vol. 22. — Pp. 102328. 10.1016/j.rineng.2024.102328

Hata M. (2013). — Empirical formula for propagation loss in land mobile radio services // IEEE Transactions on Vehicular Technology. — Vol. 29(3). — Pp. 317–325. 10.1109/T-VT.1980.23859.

Hufford G. A., Longley A. G. & Kissick W. A. (1982). A guide to the use of the ITS Irregular Terrain Model in the area prediction mode // U.S. Department of Commerce. National Telecommunications and Information Administration // NTIA Report 82-100. — Pp. 126. https://its.ntia.gov/publications/download/82-100_ocr.pdf

International Telecommunication Union. (2019). ITU-R P.1546-6. *Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 4000 MHz //* Series P: Radiowave Propagation. — Geneva. — Pp. 59. https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1546-6-201908-I!!PDF-R.pdf

International Telecommunication Union. (2021). ITU-R P.1812-6. *A path-specific propagation prediction method for point-to-area terrestrial services in the frequency range 30 MHz to 6000 MHz //* Series P: Radiowave Propagation. — Geneva. — Pp. 34. — https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1812-6-202109-S!!PDF-E.pdf

https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/17.00.00_60/tr_138901v170000p.pdf

Juang R.T. (2022). Path loss modelling based on path profile in urban propagation environments // IET

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Communications. — Vol. 16(6). — Pp. 685–694. 10.1049/cmu2.12369

Soo Q.P., Lim S.Y., Chee P.S., Lim E.H. & Yap K.M. (2025). Radio propagation modeling and measurement of uneven terrain model // Scientific Reports. — Vol. 15(1). — Pp. 28654. 10.1038/s41598-025-00958-8.

Wang J., Hao Y. & Yang C. (2023). The Current Progress and Future Prospects of Path Loss Model for Terrestrial Radio Propagation // Electronics. — Vol. 12(24). — P. 4959. 10.3390/electronics12244959.

3rd Generation Partnership Project. (2023). 3GPP TR 38.901 V17.0.0 // *Study on channel model for frequencies from 0.5 to 100 GHz (Release 17).* — 3GPP. — Pp. 100.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 247–268
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.015

FORMATION OF A UNIFIED DATA SET FOR ANALYZING THE RELATIONSHIP BETWEEN DOG BREEDS AND DISEASES

A. Kalykulova*, K. Bakirov, A. Shoman, A. Nugumanova
Astana IT University, Astana, Kazakhstan.
E-mail: aliyakalykulova01@gmail.com

Aliya Kalykulova — master student, Astana IT University, Astana, Kazakhstan
E-mail: aliyakalykulova01@gmail.com, https://orcid.org/0009-0006-5641-3797;
Kuanysh Bakirov — PhD, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0002-8059-7240;
Aruzhan Shoman — PhD, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-7844-8601;
Aliya Nugumanova — PhD, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-5522-4421.

© A. Kalykulova, K. Bakirov, A. Shoman, A. Nugumanova

Abstract. This article presents an approach to creating a dataset containing information about dog breeds, their classification in accordance with international standards, and hereditary diseases associated with these breeds. The main objective of the study was to integrate disparate sources of information and standardize breed and disease names, which enabled further comprehensive analysis. Official databases and specialized resources were used to form the final dataset: iDog, Canine Inherited Disorders Database (CIDD), The Kennel Club (UK), and American Kennel Club (AKC). The developed standardization algorithm, based on natural language processing methods, made it possible to eliminate inconsistencies in names, ensure data comparability, and form a structured dataset including 4,093 records. As part of the study, exploratory data analysis (EDA) was conducted to identify patterns in the distribution of breeds by size, classification groups, and disease types. The analysis showed that large and small breeds are characterized by a wider range of hereditary pathologies compared to medium-sized breeds. In addition, differences in the structure of diseases between The Kennel Club and American Kennel Club classifications were identified, which emphasizes the importance of taking national characteristics into account in international comparative studies. The results of the study demonstrate the significance of the developed dataset as a basis for building predictive models and decision support tools in veterinary medicine. The resulting dataset can be used to more accurately predict dogs' predisposition to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

diseases and develop effective prevention and treatment strategies.

Keywords: Creating a dataset, data standardization, Exploratory Data Analysis, NLP

For citation: Kalykulova A., Bakirov K., Shoman A., Nugumanova A. (2026). Formation of a unified data set for analyzing the relationship between dog breeds and diseases // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 247-268. https://doi.org/10.54309/IJICT.2026.25.1.015. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

Acknowledgements. *This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23488904 Development of scientifically based formulation using unsupervised machine learning in the production of full canned wet feed for unproductive animals).*

ИТ ТҰҚЫМДАРЫ МЕН АУРУЛАРЫ АРАСЫНДАҒЫ БАЙЛАНЫСТЫ ТАЛДАУҒА АРНАЛҒАН БІРЫҢҒАЙ ДЕРЕКТЕР ЖИЫНТЫҒЫН ҚАЛЫПТАСТЫРУ

А. Калыкулова, К. Бакиров, А. Шоман, А. Нугуманова*
Astana IT University, Астана, Қазақстан.
E-mail: aliyakalykulova01@gmail.com

Калыкулова Алия — магистрант, Astana IT University, Астана, Қазақстан
E-mail: aliyakalykulova01@gmail.com, https://orcid.org/0009-0006-5641-3797;
Бакиров Куаныш — PhD, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0002-8059-7240;
Шоман Аружан — PhD, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-7844-8601;
Нугуманова Алия — PhD, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-5522-4421.

© А. Калыкулова, К. Бакиров, А. Шоман, А. Нугуманова

Аннотация. Бұл мақалада ит тұқымдары, олардың халықаралық стандарттарға сәйкес жіктелуі және осы тұқымдармен байланысты тұқым қуалайтын аурулар туралы ақпаратты қамтитын деректер жиынтығын құру тәсілі ұсынылады. Зерттеудің негізгі мақсаты ақпараттың әртүрлі дереккөздерінде шашыраңқы орналасқан мәліметтерді біріктіру және тұқымдар мен аурулар атауларын стандарттау болды. Бұл кейіннен деректерге кешенді талдау жүргізуге мүмкіндік берді. Қорытынды деректер жиынтығын қалыптастыру үшін ресми дерекқорлар мен мамандандырылған ресурстар: iDog, Canine Inherited Disorders Database (CIDD), The Kennel Club (UK) және American Kennel Club (AKC) пайдаланылды. Табиғи тілді өңдеу (NLP) әдістеріне негізделген әзірленген стандарттау алгоритмі атау-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

лардағы сәйкессіздіктерді жоюға, деректердің салыстырмалылығын қамтамасыз етуге және 4 093 жазбадан тұратын құрылымдалған деректер жиынтығын қалыптастыруға мүмкіндік берді. Зерттеу аясында тұқымдардың мөлшері, жіктеу топтары және ауру түрлері бойынша таралу заңдылықтарын анықтау мақсатында деректерге барлау талдауы (Exploratory Data Analysis, EDA) жүргізілді. Талдау нәтижелері ірі және ұсақ ит тұқымдарында орташа мөлшердегі тұқымдармен салыстырғанда тұқым қуалайтын патологиялардың ауқымы кеңірек екенін көрсетті. Сонымен қатар, The Kennel Club және American Kennel Club жіктеулеріндегі аурулар құрылымының айырмашылықтары анықталды. Бұл халықаралық салыстырмалы зерттеулер жүргізу кезінде ұлттық ерекшеліктерді ескерудің маңыздылығын көрсетеді. Зерттеу нәтижелері әзірленген деректер жиынтығының ветеринариялық медицина саласында болжамдық модельдер мен шешім қабылдауды қолдау құралдарын құруға негіз бола алатынын көрсетеді. Алынған деректер жиынтығын иттердің ауруларға бейімділігін дәлірек болжау, сондай-ақ аурулардың алдын алу мен емдеудің тиімді стратегияларын әзірлеу үшін пайдалануға болады.

Түйінді сөздер: деректер жиынтығын құру, деректерді стандарттау, деректерге барлау талдауы, табиғи тілді өңдеу (NLP)

Дәйексөздер үшін: Калыкулова А., Бакиров К., Шоман А., Нугуманова А. (2026). Ит тұқымдары мен аурулары арасындағы байланысты талдауға арналған бірыңғай деректер жиынтығын қалыптастыру // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 247–268. https://doi.org/10.54309/IJICT.2026.25.1.015. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

Алғыс. *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитетінің қаржылық қолдауымен жүзеге асырылды (№ AP23488904 грант «Өнімсіз жануарларға арналған толыққанды ылғалды консервіленген жем өндірісінде оқытусыз машиналық оқыту әдістерін қолдана отырып, ғылыми негізделген рецептураны әзірлеу»).*

ФОРМИРОВАНИЕ ЕДИНОГО НАБОРА ДАННЫХ ДЛЯ АНАЛИЗА ВЗАИМОСВЯЗИ МЕЖДУ ПОРОДАМИ СОБАК И ЗАБОЛЕВАНИЯМИ

А. Калыкулова, К. Бакиров, А. Шоман, А. Нугуманова*
Astana IT University, Астана, Казахстан.
E-mail: aliyakalykulova01@gmail.com

Калыкулова Алия — магистрант, Astana IT University, Астана, Казахстан
E-mail: aliyakalykulova01@gmail.com, https://orcid.org/0009-0006-5641-3797;
Бакиров Куаныш — PhD, Astana IT University, Астана, Казахстан
https://orcid.org/0009-0002-8059-7240;
Шоман Аружан — PhD, Astana IT University, Астана, Казахстан

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

https://orcid.org/0000-0002-7844-8601;
Нугуманова Алия — PhD, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-5522-4421.

© А. Калыкулова, К. Бакиров, А. Шоман, А. Нугуманова

Аннотация. В статье представлен подход к созданию набора данных, содержащего информацию о породах собак, их классификации в соответствии с международными стандартами, а также о наследственных заболеваниях, связанных с данными породами. Основной целью исследования являлась интеграция разрозненных источников информации и стандартизация названий пород и заболеваний, что обеспечило возможность их дальнейшего комплексного анализа. Для формирования итогового набора данных были использованы официальные базы данных и специализированные ресурсы: iDog, Canine Inherited Disorders Database (CIDD), The Kennel Club (UK) и American Kennel Club (AKC). Разработанный алгоритм стандартизации, основанный на методах обработки естественного языка, позволил устранить несоответствия в названиях, обеспечить сопоставимость данных и сформировать структурированный набор данных, включающий 4 093 записи. В рамках исследования был проведен разведочный анализ данных (Exploratory Data Analysis, EDA) для выявления закономерностей в распределении пород по размеру, классификационным группам и типам заболеваний. Анализ показал, что крупные и мелкие породы характеризуются более широким спектром наследственных патологий по сравнению с породами среднего размера. Кроме того, были выявлены различия в структуре заболеваний между классификациями The Kennel Club и American Kennel Club, что подчеркивает важность учета национальных особенностей при проведении международных сравнительных исследований. Результаты исследования демонстрируют значимость разработанного набора данных как основы для построения прогностических моделей и инструментов поддержки принятия решений в ветеринарной медицине. Полученный набор данных может использоваться для более точного прогнозирования предрасположенности собак к заболеваниям, а также для разработки эффективных стратегий профилактики и лечения.

Ключевые слова: создание набора данных, стандартизация данных, разведочный анализ данных, обработка естественного языка (NLP)

Для цитирования: Калыкулова А., Бакиров К., Шоман А., Нугуманова А. (2026). Формирование единого набора данных для анализа взаимосвязи между породами собак и заболеваниями // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 247-268. https://doi.org/10.54309/IJICT.2026.25.1.015. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность. *Данное исследование выполнено при финансовой поддерж-*

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ке Комитета науки Министерства науки и высшего образования Республики Казахстан (грант № AP23488904 «Разработка научно обоснованной рецептуры с использованием методов обучения без учителя при производстве полнорационного влажного консервированного корма для непродуктивных животных»).

Introduction.

Modern information technologies enable the collection, processing, and analysis of large amounts of data, opening up new opportunities for veterinary medicine, particularly in studying the predisposition of animal breeds to various diseases. The problem is that information about breeds, breed diseases, types of disorders, predisposition levels, and breed sizes is often scattered across different sources, formatted in different ways, and even uses different terms and classifications. This fragmentation hinders the creation of large-scale, unified databases and, as a result, prevents the effective application of machine learning, predictive analytics, and other IT approaches to disease prevention and diagnosis.

The relevance of the study lies in the need for a systematic approach to collecting and analyzing data on dog breeds and related diseases for the application of modern information technologies in veterinary medicine. Currently, information on the types of disorders, predisposition levels, and breed characteristics is scattered, presented in different formats, and often uses inconsistent classifications, which makes it difficult to integrate and analyze. In a study by (Donner et al., 2018), based on the analysis of the DNA of more than 100,000 dogs, it was found that about 40 % of animals have at least one genetic variant associated with hereditary diseases, which indicates a high prevalence of disease predisposition and the need for comprehensive data analysis. Similarly, recent reviews in the field of veterinary informatics emphasize the importance of aggregating information from various sources, standardizing it, and improving its quality to create representative and reliable sets suitable for automated analysis and the implementation of digital solutions in veterinary practice (Lustgarten et al., 2020). These facts confirm that the development of approaches to data structuring and analysis is a key direction for improving the effectiveness of animal health diagnosis, prevention, and management.

The purpose of this study is to create and analyze a dataset containing information about dog breeds and related diseases, as well as to conduct an explanatory analysis to identify relationships between key characteristics and demonstrate the potential applications of modern data processing methods in veterinary informatics.

During the formation of the dataset, special attention was paid to the standardization of breed names, as different sources use different spellings and classifications. An additional complication was the heterogeneity of the information: for some breeds, there are officially approved groupings, while for others, no such classification exists. In addition, some breeds cannot be included in the dataset without official recognition by international or national cynological organizations, which limits the completeness of the information presented.

The novelty of the study lies in the presentation of a compact, unified dataset

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

created on the basis of data from various sources, with the names of breeds and diseases converted to a single format. This approach allows for preliminary explanatory analysis and demonstrates ways of working with disparate information in the field of veterinary informatics. The resulting dataset can serve as a starting point for further research aimed at expanding the volume of information and developing more comprehensive analytical tools.

Materials and methods.

In recent years, there has been rapid growth in research related to the use of big data in veterinary medicine, opening up new opportunities for animal health analysis, disease diagnosis, and the development of preventive measures. However, the process of integrating data on dog breeds and their hereditary diseases remains a challenging task due to the heterogeneity of sources, lack of standardization, and variability of data formats.

One of the key areas is the collection and systematization of data from various sources. (Paynter et al., 2021) review existing large databases in veterinary medicine, including clinical records, insurance data, and specially organized cohort studies. The authors highlight the problems of the lack of uniform standards for record keeping and the ambiguity of diagnostic terminology, which makes it difficult to integrate data and conduct comparative studies. These findings are confirmed by other studies, which show that electronic health records (EHRs) have significant potential for analysis but require careful processing and normalization. For example, (Davies et al., 2025) demonstrate how text processing and NLP methods can be applied to extract clinically relevant information, including the use of specialized dictionaries and regular expressions.

An important source of data is electronic medical records from primary veterinary care. (Summers et al., 2019) and (Jackson et al., 2024) showed that such data can be used to identify the most common diseases in dogs, analyze age-related characteristics, and study factors affecting health. However, they also note that data quality depends on the completeness of records, the level of detail, and the participation of clinics in data collection projects.

The issue of standardizing information on dog breeds is of fundamental importance for further analysis. (Mullen et al., 2025) proposed the development of an ontology of vertebrate breeds that would allow data from various sources, such as FAO DAD-IS and national registries, to be combined while preserving their provenance. Despite the promise of this approach, the researchers encountered difficulties related to conflicting sources and the need for manual data verification. Additionally, (Perez et al., 2025) consider computer vision methods for automatic breed identification from images, which could help fill gaps in the data, but emphasize that visual breed identification is difficult due to mixed lines and differences in classifications.

Analyzing the relationship between breeds and hereditary diseases requires high-quality data on disease incidence. A study by (Oberbauer et al., 2015) found that susceptibility to hereditary diseases varies not only between purebred and mixed-breed dogs, but also within individual breed groups. However, the authors note the limitations

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

of their data, which was based on a sample from a single veterinary clinic, which could have led to a bias in the results.

To solve the problem of integrating heterogeneous data, preliminary cleaning and structuring are important. (Boland et al., 2019) proposed a two-step approach to building a clinical repository of dog data: first, semantic cleaning is performed to unify breed and animal species names, followed by domain-specific cleaning to identify logical inconsistencies and errors in the records. This approach improves data quality and preserves the maximum amount of data for analysis.

Modern technologies such as knowledge graphs and multimodal models open up new opportunities for data analysis and supplementation. (Hoang et al., 2023) showed how electronic medical records can be used to build a knowledge graph for diagnosing diseases in domestic animals, taking into account textual descriptions of symptoms. In turn, (Yue et al., 2025) explore methods for supplementing knowledge graphs using large language models and multimodal data, which can be applied to provide a more complete representation of information about breeds and diseases.

In addition, blockchain and IoT technologies can be used to track pedigrees and monitor animal welfare conditions. For example, (Lisnevskyi et al., 2025) proposed an order logistics platform that can be adapted to track the origin and history of dog breeds. Similarly, (Harahap et al., 2022) consider smart feeding systems that allow real-time collection of data on dog nutrition and health.

In today's world, technologies aimed at improving comfort and quality of life for both humans and animals are actively developing. For example, (Neftissov et al., 2025) proposed the concept of a "smart office," which allows monitoring the indoor environment, including parameters such as lighting, air quality, temperature, and humidity. This data can be used to assess the impact of external factors on health and well-being, which opens up prospects for the implementation of similar solutions in animal care systems, for example, in premises where dogs are kept.

In addition, the diagnosis of diseases in dogs remains an important area of research. (Rathnayaka et al., 2022) eveloped an intelligent system that allows the diagnosis of skin diseases in dogs based on images. As part of their work, they used ontology to extract and systematize clinical information and created a web application that simplifies the diagnostic process. To train the model, a dataset was collected that included 300 images for three types of diseases, with preprocessing and augmentation to improve the quality of the results.

Creating a unified dataset for analyzing the relationships between dog breeds and their diseases requires a comprehensive approach that includes:

- standardization of data on breeds and diseases (Mullen et al, 2025; Perez et al., 2025);
- cleaning and structuring information to improve the quality of analysis (Boland et al., 2019);
- integration of various data sources, including clinical records and genetic information (Paynter et al., 2021; Davies et al., 2025);

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

● application of modern analysis methods, such as knowledge graphs and multimodal models (Hoang et al., 2023; Yue ET AL., 2025).

The combination of these approaches provides a basis for comprehensive analysis of hereditary diseases in dogs and identification of patterns that may be useful for both veterinary practice and scientific research.

The final dataset was compiled in stages and included the collection, integration, and preliminary processing of information obtained from various official sources. To this end, four separate tables were used, compiled on the basis of data collected using web scraping. In the next stage, the data was combined into a single structure, taking into account differences in the formats of information presentation and the classifications used. Particular attention was paid to ensuring data comparability, which required standardization of breed and disease names. This approach made it possible to create a comprehensive and structured dataset suitable for further analysis and visualization.

The main source of information on dog breeds is the specialized resource iDog, developed by the National Genomics Data Center (NGDC, China) (NGDC, 2025). This resource contains standardized information on breeds, their characteristics, diseases, and related genetic variations. An important advantage of iDog is the availability of unified breed identifiers (Breed Ontology ID), which ensures the comparability of data obtained from various sources, including the AKC, UKC, and other international cynological organizations. The study used standard breed names, their variations, and information about diseases, including the level of heritability or increased predisposition, as well as a list of diseases that may occur in the corresponding breeds. This dataset included 482 dog breeds.

The Canine Inherited Disorders Database (CIDD) (CIDD, 2025) was used to structure information on hereditary diseases, where diseases are classified by body system. This approach allowed for more effective grouping and analysis of the data. Information about the types of diseases and the breeds in which they occur was used. The final dataset contained 147 diseases, grouped into 11 categories, which were associated with 151 breeds. The total volume of the table was 1,649 records.

Additional information about breeds was collected from The Kennel Club's official website, which is the official source of breed data supported by the UK's largest dog registration and standards organization (The Kennel Club, 2025). This source provides information on breed names, their classification into groups (e.g., terriers, greyhounds, toy dogs), and size categories. The processing resulted in 223 unique breeds.

Data from the American Kennel Club (AKC) (Filho, 2025) was used to obtain information on the distribution of breeds by official groups, such as Herding, Sporting, Working, and others. This source is an independent American breed classification standard, which differs from the standards used in the UK or other countries. Using information from the AKC made it possible to supplement and expand the data on breeds, ensuring their comparability in an international context. A total of 277 dog breeds were collected.

After preliminary preparation, all datasets were integrated into a single table.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The main difficulty lay in discrepancies in the names of breeds and diseases, as well as incomplete data: for some breeds, information on disease types or heritability levels was missing. To solve these problems, a standardization methodology was applied, including preliminary data processing using natural language processing (NLP) tools. All data processing was performed using the Python language. Before integration, temporary tables were created to unify breed and disease names. Figure 1 shows a diagram of the data merging process using intermediate tables.

To ensure consistency, all names were converted to lowercase. There were often cases where an alternative name was added to the main name in parentheses. In such cases, both variants were analyzed separately. First, completely identical names were matched, which were then no longer taken into account in the subsequent classification, which reduced the analysis load and avoided duplication. The process of matching disease and breed names from two tables is described in more detail below.

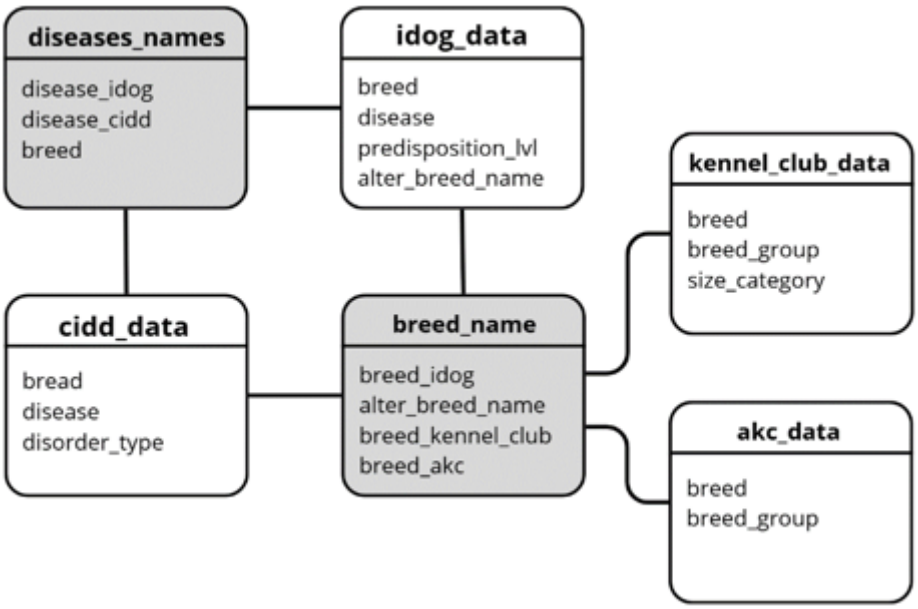


Fig. 1. Data sets and the process of merging using intermediate data.

When working with disease names from the iDog table and the CIDD table, the key task was to correctly match the corresponding entries. To do this, an approach based on tokenization and cleaning the text of punctuation marks was used. If the name contained an alternative form in parentheses, it was divided into two independent variants.

The formal condition for matching names was determined as follows:

$$(a_t \subseteq b_t) \vee (b_t \subseteq a_t) \vee (a_i \in b_j) \vee (b_i \in a_j); i,j \in \{1,2\} \quad (1)$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

where *i* and *j* are the first and second variants of the text name (if there is a second name in parentheses), a_i and b_i are the text names of diseases in the diseases_idog and diseases_cidd columns, and at and b_t are their tokenized and punctuation-cleaned forms.

Two records were considered to describe the same disease if the set of tokens of one record was completely contained in the set of tokens of the other, or if one of the text variants of the name was present in the other.

After successful matching of the names, a correspondence table was created, on the basis of which the data was merged. A new column containing information about the types of diseases was added to the final iDog table, if this information was present in the CIDD database. Thus, a unified and structured database was formed, providing further analysis.

Name of breeds

To process and match breed names, we used a method based on combining data from various sources with reference to the main idog_data table. This table contains a list of dog breeds, including data from several international organizations, and also provides alternative spellings and names for breeds. In particular, the alter_breed_name column in the idog_data table contains a list of such alternative names for each breed.

As part of the task, a set of names an is formed for each record in the idog data table, including the main and all alternative names. If the name contains a variant in parentheses, it is split into two separate elements and added to the set. During the preprocessing stage, all punctuation marks are removed from the text. Based on the resulting set of names, a set of tokens at is formed, containing all unique tokens found in these names.

For data from other tables, each original breed name b_i is also split into two parts, if necessary, if it contains an alternative variant in parentheses. After that, tokenization is performed, resulting in a set of tokens b_{ti} being formed for each name variant. As in the case of the idog_data table, all punctuation marks are removed from the text.

Breeds are matched according to the following rule:

$$(b_{ti} \subseteq a_t) \vee (a_k \in b_i) \vee (b_i \in a_k); i \in \{1,2\}; k \in \{1,2 \dots n\} \quad (2)$$

where *n* is the number of variations of names of one breed from the idog_data table. The set a_n includes all breed names (main and alternative), a_t is the set of tokens obtained from these names, b_i is the text name of the breed from other tables, *i* is the name variant index (if it is divided into two with brackets), and b_{ti} is the tokenized form of the corresponding name.

Thus, breeds are considered to match if all tokens of one of the name variants b_{ti} are completely contained in the set of tokens at, or if the text of one of the name variants from one table is a substring of the name from another table. This allows for the correct consideration of differences in the spelling and structure of breed names, ensuring data matching even in the presence of alternative or partial names.

After creating intermediate tables with standardized names for breeds and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

diseases, the data was integrated into a single structure. During the consolidation phase, records were compared between sources: if certain breeds were missing from the intermediate tables, they were added to the final set. If values were missing for some fields, they were left blank, which allowed the completeness of the information to be preserved without distorting the data.

Figure 2 shows a fragment of the final dataset, and Table 1 provides a brief description of it. The final dataset contains 4,093 records, each of which is a unique combination of the following characteristics: breed (Breed), type of disorder (Type_of_disorder), specific disorder (Disease), level of heritability or predisposition (Predisposition_to_disease), dog size (Size), The Kennel Club classification group (Group_UK), American Kennel Club classification group (Group_AKC), and alternative breed names (Alternative_breed_names).

If there are several rows in the dataset with the same breed but different diseases, this indicates that the breed may be predisposed to several types of pathologies.

The dataset includes two columns containing breed classifications according to the standards of various organizations: The Kennel Club (UK) – column Group_UK and American Kennel Club (AKC, USA) – column Group_AKC. Despite the general similarity of the classification systems, they do not completely coincide. For example, The Kennel Club distinguishes between the groups “gundog,” “toy,” “pastoral,” “utility,” “hound,” “working,” and “terrier,” while the American Kennel Club uses the categories “foundation stock service,” “toy,” “pastoral,” “miscellaneous class,” “non-sporting,” “hound,” “working,” “terrier,” and “sporting.” In some cases, the same breed may belong to different groups depending on the standard used.

Table 1. Brief description of the dataset obtained

Column Name	Count categories	Null
Breed	482	0
Type_of_disorder	11	1501
Disease	428	272
Predisposition_to_disease	4	2579
Size	6	557
Group_UK	7	557
Group_AKC	9	160
Alternative_breed_names	482	0

The result was a structured data set that allows for systematic analysis of the relationships between breeds, disease types, and classification groups according to various standards.

As part of this study, exploratory data analysis (EDA) was conducted to identify key patterns and characteristics that characterize the distribution of animal breeds, their sizes, classification groups, and types of diseases. Such analysis is necessary for a deeper understanding of the data structure and the formation of hypotheses about possible relationships between different characteristics.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

	Breed	Type_of_disorder	Disease	Predisposition_to_disease	Size	Group_UK	Group_AKC	Alternative_breed_names
0	afghan hound	inherited eye disorders	cataract	have an increased incidence	large	hound	hound	afghan hound, tazi, tazhi spay, kochyano spay,...
1	afghan hound	inherited eye disorders	corneal dystrophy	have an increased incidence	large	hound	hound	afghan hound, tazi, tazhi spay, kochyano spay,...
2	afghan hound	inherited skin disorders	demodicosis	have an increased incidence	large	hound	hound	afghan hound, tazi, tazhi spay, kochyano spay,...
3	afghan hound	inherited eye disorders	glaucoma	may be inherited	large	hound	hound	afghan hound, tazi, tazhi spay, kochyano spay,...
4	afghan hound	inherited musculoskeletal disorders	hip dysplasia	have an increased incidence	large	hound	hound	afghan hound, tazi, tazhi spay, kochyano spay,...
...	...	...	...	...	...	...	...	...
4088	chinese imperial dog	inherited musculoskeletal disorders	patellar luxation	none	small	utility	toy	chinese imperial dog, imperial shih-tzu, lion ...
4089	chinese imperial dog	inherited musculoskeletal disorders	intervertebral disk disease	none	small	utility	toy	chinese imperial dog, imperial shih-tzu, lion ...
4090	chinese imperial dog	inherited nervous system disorders	intervertebral disk disease	none	small	utility	toy	chinese imperial dog, imperial shih-tzu, lion ...
4091	paisley terrier	inherited musculoskeletal disorders	patellar luxation	none	small	toy	toy	paisley terrier, clydesdale terrier, silky, si...
4092	soft-coated wheaten terrier	inherited endocrine disorders	hypoadrenocorticism (addison's disease)	none	small	terrier	terrier	soft coated wheaten terrier, irish soft-coated...

Fig. 2. Example of the structure of the final dataset.

At the initial stage, attention was focused on studying the distribution of breeds by size categories and classification groups, which are determined on the basis of international standards. Next, the distribution of diseases by type of pathology was analyzed, which made it possible to assess their diversity. After the general patterns of breed and disease distributions were examined, the next stage of the analysis focused on identifying potential relationships between these characteristics. For this purpose, only records with a specified value in the Predisposition_to_disease column were considered.

To determine whether statistically significant associations exist between disease types, breed size, and classification groups, a Chi-square test of independence (X2) was applied. This method assesses whether deviations between observed and expected frequencies arise from random variation or indicate genuine interdependence between categorical variables. The Chi-square statistic is computed as:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

where r is the number of rows (categories of the first variable), c is the number of columns (categories of the second variable), O_{ij} represents the observed frequency in cell (i, j), E_{ij} - the expected frequency in cell (i, j), calculated as:

$$E_{ij} = \frac{(rowtotal)_i * (columntotal)_j}{n} \quad (4)$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

where *n* is the total number of observations. The null hypothesis (H0) assumes that the two variables are independent, meaning that deviations between O_{ij} and E_{ij} are due to random variation. A statistically significant χ^2 value ($p < 0.05$) indicates that the variables are not independent, implying an association between them.

The strength of association was further evaluated using Cramer's V, which provides a standardized effect size for categorical relationships, ranging from 0 (no association) to 1 (perfect association). It is calculated as:

$$V = \sqrt{\frac{X^2}{n*(k-1)}} \tag{5}$$

where *n* is the total sample size, *k* is the smaller of (*r, c*), i.e. the number of rows or columns in the contingency table.

Together, the χ^2 statistic and Cramer's V enable both the detection of statistically significant relationships between categorical variables and the evaluation of their practical strength.

To ensure the accuracy of comparisons between category groups, a contingency table was constructed, representing the distribution of observations by disease type and breed characteristics. However, the initial data exhibited a marked imbalance in category frequencies, which could bias the interpretation by overemphasizing categories with larger sample sizes. To mitigate this effect, Z-normalization was applied across the rows of the contingency table, allowing each value to be interpreted relative to its mean and standard deviation within the corresponding row. This normalization procedure reduced the influence of scale differences and ensured a more reliable comparison of categorical relationships.

To eliminate this effect, Z-normalization of the values in the contingency table was performed across rows. This approach allows values within each row to be compared relative to its mean and standard deviation, eliminating the influence of differences in category scales.

After normalization, all values within each row are converted to a comparable scale, where zero corresponds to the mean level. Positive values indicate above-average values, while negative values indicate below-average values. In cases where the standard deviation is zero, which means there is no variation within the row, all values in that row were replaced with zero to avoid division by zero.

The result of normalization was a matrix whose values reflect relative deviations from the average level for each combination of characteristics. This matrix was used to construct a heat map visualizing the relationships between animal size and disease types. The visualization made it possible to identify which size categories had a higher or lower association with certain diseases, eliminating the influence of the initial uneven distribution of data. A similar approach was used to analyze the relationships between

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

disease types and breed classification groups, which provided a more objective and accurate identification of structural patterns in the data.

Let x_{ij} denote the initial value in the table cell, where i is the row index (disease type) and *j* is the column index (animal size category). For each row, the row mean (μ_i) and row standard deviation (σ_i) are calculated. The normalized value Z_{ij} is then calculated as the difference between the original value and the row mean, divided by the row standard deviation:

$$Z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{6}$$

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{7}$$

$$\sigma_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - \mu_i\right)^2} \tag{8}$$

where μ_i is the mean value for the row, σ_i is the standard deviation for the row, and *n* is the number of animal size categories in the row.

Thus, the exploratory data analysis not only allowed us to describe the overall structure of the sample, but also provided information for further stages of the study. Data normalization was an important step in ensuring the accuracy of subsequent analysis and the reliability of the interpretation of the results obtained.

Results and discussion.

As a result of the explanatory data analysis, key patterns were identified that reflect the relationships between dog size, breed classification according to international standards, and the diversity of hereditary diseases.

Figure 3 shows the distribution of breeds by size category. Large breeds account for the largest share (39.1 %), followed by small breeds (30.1 %) and medium breeds (24.3 %). The intermediate categories “small-medium” and “medium-large” are insignificant, accounting for 2.5 % each. The smallest group is “extra large,” which accounts for only 1.4 % of the total number of breeds. This distribution reflects the predominance of medium and large dogs in the sample compared to the extreme categories.

Figures 4 and 5 show the distribution of breeds by group according to the classifications of The Kennel Club (UK) and the American Kennel Club (AKC). In the UK classification, there is a relatively even distribution of breeds between groups, with

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the largest number of varieties falling into the “hound” and “gundog” groups. In turn, the AKC classification is dominated by the foundation stock service group, followed by hound, working, and sporting. The smallest number of breeds are represented in the toy and miscellaneous class groups. These differences reflect the characteristics of national classification systems, which is important to consider in comparative studies.

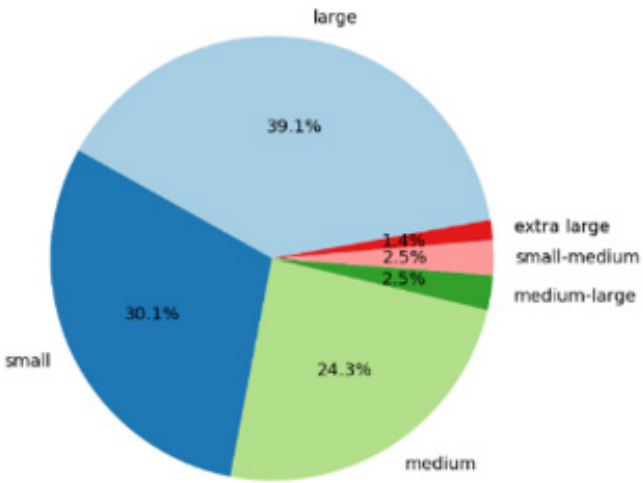


Fig. 3. Distribution of dog breeds by size categories.

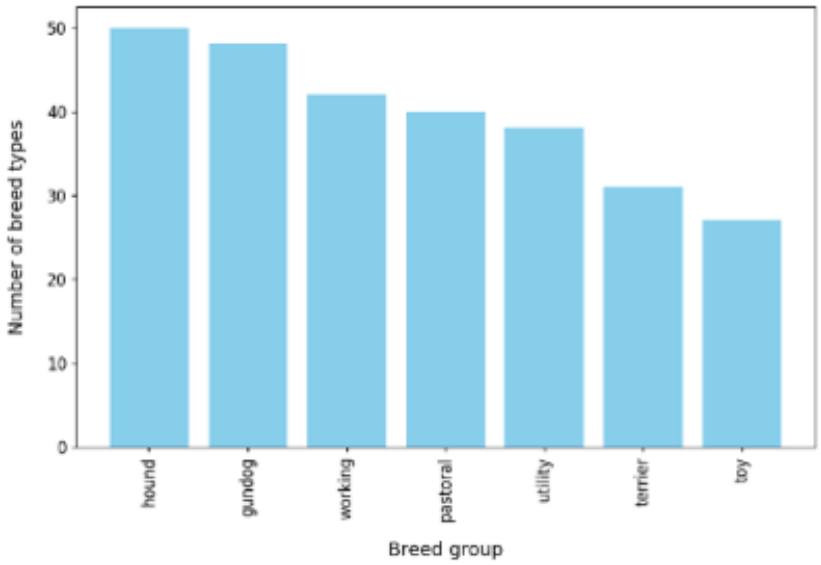


Fig. 4. Distribution of dog breeds by group (UK).

Figures 4 and 5 show the distribution of breeds by group according to the classifications of The Kennel Club (UK) and the American Kennel Club (AKC). In the UK classification, there is a relatively even distribution of breeds between groups, with the largest number of varieties falling into the “hound” and “gundog” groups. In turn, the AKC classification is dominated by the foundation stock service group, followed by

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

hound, working, and sporting. The smallest number of breeds are represented in the toy and miscellaneous class groups. These differences reflect the characteristics of national classification systems, which is important to consider in comparative studies.

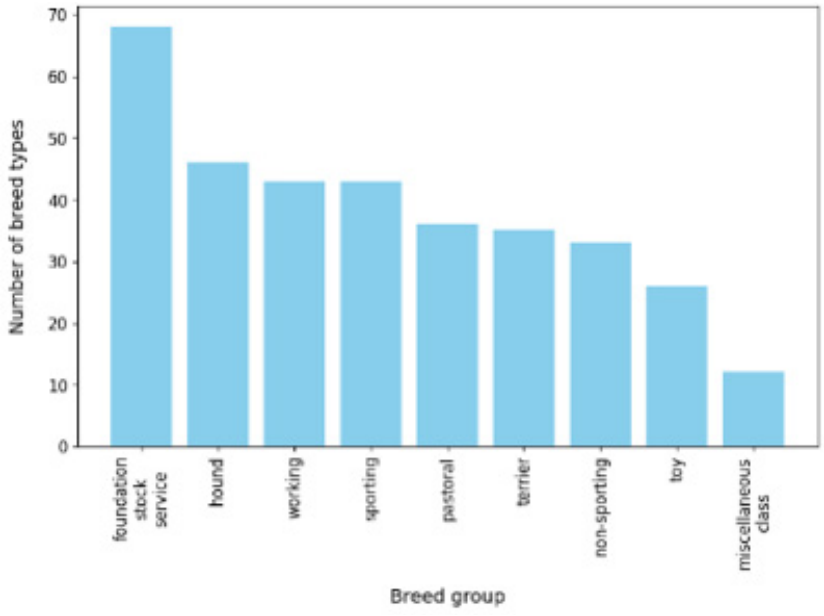


Fig. 5. Distribution of dog breeds by group (AKC).

An analysis of hereditary diseases has shown that the greatest variety of pathologies is observed among skin diseases – about 40 different types. A significant number of diseases are also associated with disorders of the nervous system and pathologies of the organs of vision. These data are presented in Figure 6 and indicate that these categories require special attention in prevention and diagnosis.

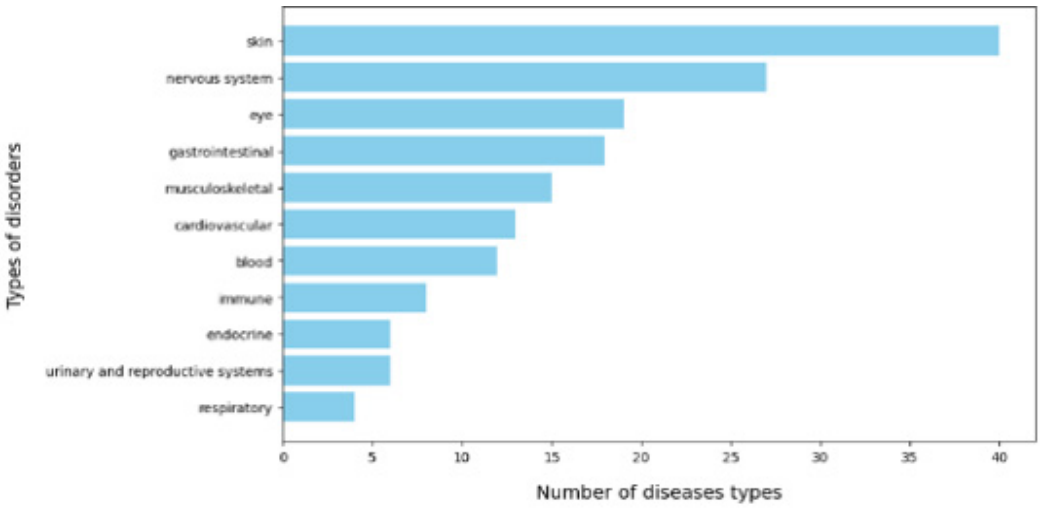


Fig. 6. Number of types of hereditary diseases by type of disorder.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The results of the Chi-square test of independence demonstrated that the distribution of disease types across breed size categories and breed classification groups is not random. The results are summarized in Table 2. In all cases, the p-values were below the 0.05 threshold, which leads to the rejection of the null hypothesis of independence between variables. This finding indicates the presence of statistically significant associations between disease types and breed characteristics. However, according to the values of Cramer's V, the strength of these associations can be interpreted as weak, suggesting that the relationships between variables are partly influenced by other contributing factors.

Table 2. Results of the Chi-square test and Cramer's V coefficient

Data	Chi-square statistic	Degrees of freedom	p-value	Check	Cramer's V
Disease type × Breed size	66.801	20	6.01e-07	p-value < 0.05	0.159
Disease type × Breed group (UK)	88.779	60	9.27e-03	p-value < 0.05	0.106
Disease type × Breed group (AKC)	138.393	80	5.57e-05	p-value < 0.05	0.114

The following analysis presents a more detailed examination of the relationships between disease types and breed characteristics. Figure 7 shows the distribution of disease types according to dog size. Three main categories were identified for analysis: large, medium, and small breeds. The results indicate that large and small breeds have a wider range of diseases compared to medium-sized breeds.

Large breeds are more prone to skin and endocrine system pathologies, as well as musculoskeletal disorders. In addition, they have a higher level of diversity in diseases related to the circulatory system (blood disorders), gastrointestinal tract (gastrointestinal disorders), and cardiovascular system (cardiovascular disorders).

Small breeds, on the other hand, show an increased variety of urinary and reproductive system disorders, as well as eye, nervous system, and respiratory system disorders. For medium-sized breeds, the greatest variety is found among immune system diseases, while they show a more balanced profile in other categories.

Figures 8 and 9 show the results of the analysis of the distribution of disease types by breed groups in the UK and AKC classifications.

The UK system reveals marked differences between groups: the toy group is characterized by the greatest variety of cardiovascular diseases, terriers are more prone to eye diseases, breeds in the pastoral group are dominated by immune system pathologies, and the hound group is prone to gastrointestinal diseases. Gundog breeds are more prone to blood diseases, while in the utility group, diseases are more evenly distributed between the respiratory and endocrine systems and skin pathologies. The working group is characterized by a moderate variety of diseases, without any clearly dominant categories, but endocrine, musculoskeletal, and cardiovascular disorders are more common.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

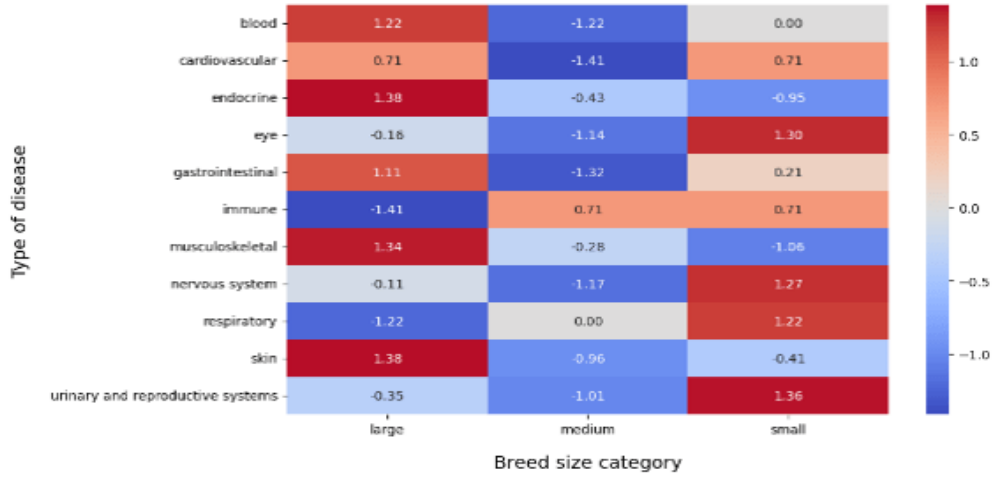


Fig. 7. Distribution of disease types by dog size categories.

In the AKC classification, the greatest variety of diseases is recorded in the "foundation stock service" group, where diseases of the nervous and endocrine systems, as well as the musculoskeletal system, predominate. Skin pathologies are more common in the "non-sporting" group, gastrointestinal diseases in the 'hound' group, and immune system diseases in the "pastoral" group.

Sporting breeds are characterized by a wide range of blood diseases, while terriers are more evenly represented by diseases of the gastrointestinal tract, genitourinary system, and blood. For the toy group, diseases of the respiratory and cardiovascular systems, as well as pathologies of the genitourinary system, are most typical. Cardiovascular diseases dominate in the working group, but the overall diversity of pathologies is lower compared to other groups.

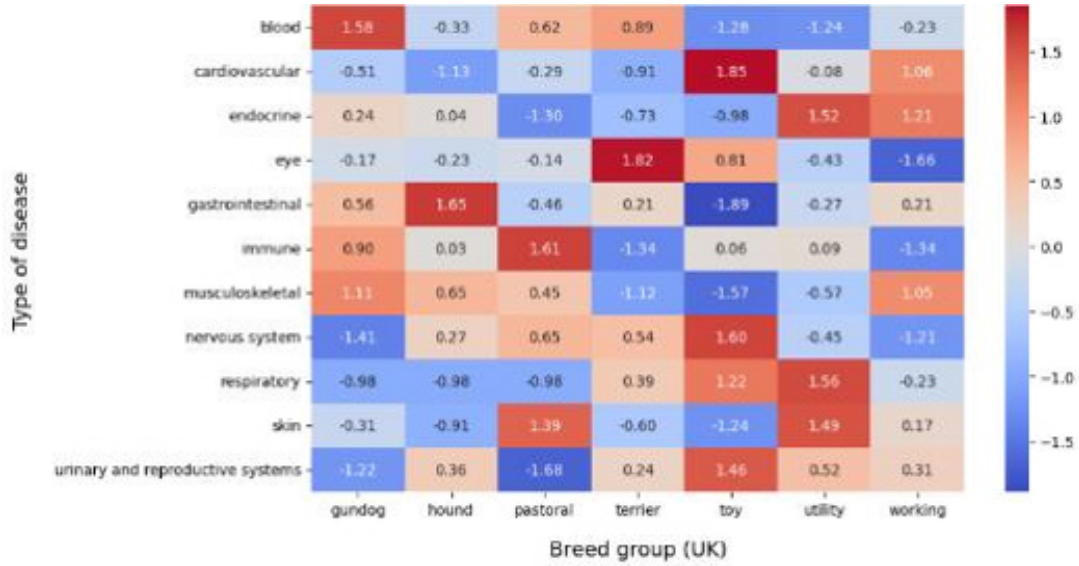


Fig. 8. Distribution of disease types by dog breed group (UK).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

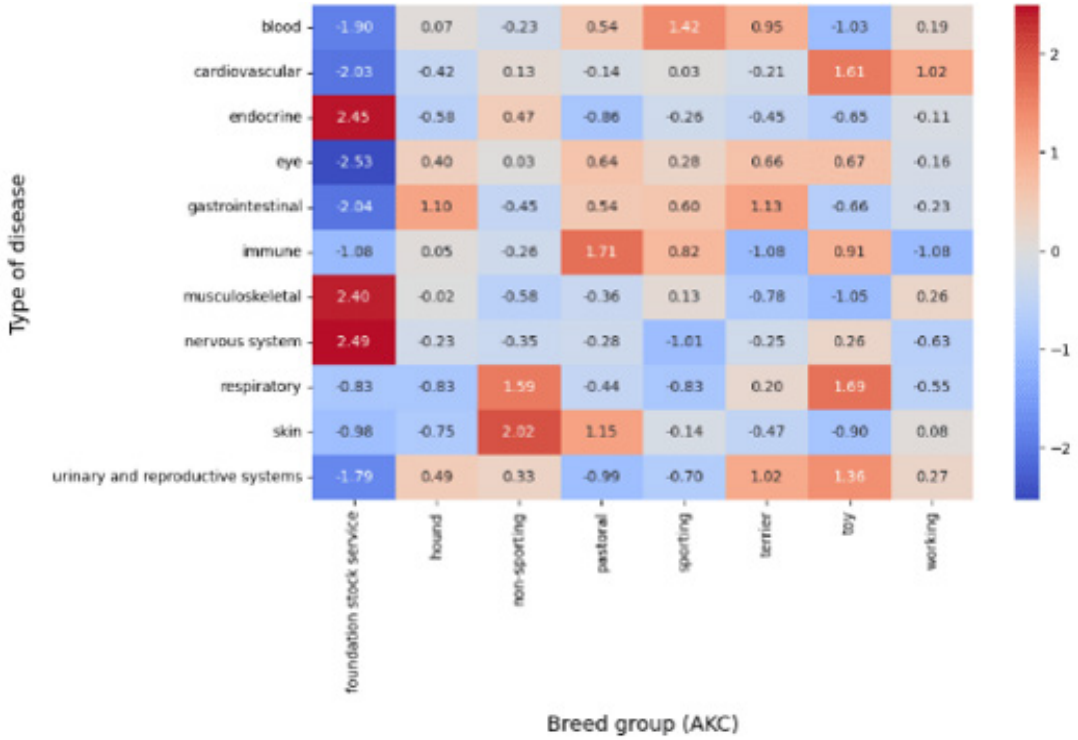


Fig. 9. Distribution of disease types by dog breed group (AKC).

Discussion.

The results of the study confirmed that the developed unified data set is an effective tool for analyzing the predisposition of dog breeds to various diseases and identifying patterns related to breed size, classification, and types of pathologies. The data obtained made it possible to systematize scattered information previously presented in various sources and create a basis for further research in the field of veterinary analytics and machine learning.

The identified relationship between dog size and the spectrum of diseases is consistent with known veterinary observations: large breeds are more prone to pathologies associated with the musculoskeletal, cardiovascular, and endocrine systems, which is due to both their physiological characteristics and the higher load on the body. Small breeds, on the contrary, show an increased incidence of diseases of the genitourinary system, eyes, and respiratory tract. Medium-sized breeds have a relatively balanced profile of pathologies, which may indicate a lower incidence of breed predispositions. These results are important for veterinarians and breeders, as they allow them to predict potential risks and develop targeted prevention strategies.

A comparison of breed classifications according to The Kennel Club (UK) and American Kennel Club (AKC) standards revealed both similarities and differences between the systems. Despite similarities in most categories, a number of breeds were classified into different groups depending on the standard used. These differences can affect the statistical interpretation of data and highlight the need to consider multiple

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

classification systems in analytical studies.

For example, the AKC system has a high concentration of breeds in the "foundation stock service" group, reflecting an emphasis on breeds that are in the process of being recognized or rare varieties, while the UK classification is more uniform. This difference may be due to the historical characteristics of the development of cynology in each country.

An analysis of the distribution of hereditary diseases showed that the greatest variety of pathologies is associated with skin diseases, nervous system disorders, and diseases of the organs of vision. These categories may require special attention when developing diagnostic and preventive measures, as they are most prevalent among various breed groups.

The methodological aspect of this study is particularly important. In the process of standardizing breed and disease names, we managed to overcome the problem of duplication and inconsistencies characteristic of the original sources. The use of natural language processing (NLP) methods made it possible to unify text data and ensure correct matching of records, which is an important step in preparing data for machine learning. This approach can be used in other areas of bioinformatics and veterinary analytics, where the problem of data heterogeneity also occurs.

It should be noted that the results of the study have certain limitations. First, data on rare breeds and diseases may be incomplete, which affects the completeness of the analysis. Second, the use of different international classifications creates the risk of ambiguity in interpretation when comparing groups.

Conclusion.

The study solved the problem of creating a unified and structured dataset containing information about dog breeds, their classification according to international standards, and related hereditary diseases. The developed approach to standardizing breed and disease names made it possible to integrate disparate sources of information into a single dataset containing 4,093 records and covering a wide range of characteristics, making it a valuable resource for further analytical and prognostic research in veterinary informatics.

The results of the exploratory data analysis (EDA) showed the presence of pronounced patterns reflecting the relationship between dog size, classification group membership, and the variety of hereditary diseases. It was found that large and small breeds are characterized by a wider range of diseases compared to medium-sized breeds, confirming the need for a differentiated approach to prevention and diagnosis. In addition, differences in the structure of diseases between classification groups according to The Kennel Club and American Kennel Club standards were identified, indicating the importance of taking national classification systems into account in international comparative studies.

The developed dataset represents a foundation for continued analytical studies. Its structure allows integration into decision-support systems and breed-specific disease risk assessment tools aimed at improving preventive strategies in veterinary medicine.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Future research will focus on expanding the dataset by incorporating environmental and nutritional information, which will enhance its relevance for both practical and research applications.

REFERENCES

Boland M.R., Casal M.L., Kraus M.S., Gelzer A.R. (2019). Applied veterinary informatics: development of a semantic and domain-specific method to construct a canine data repository // Scientific Reports. — Vol. 9. — Pp. 55035. DOI:10.1038/s41598-019-55035-8. [in Eng.].

Canine Inherited Disorders Database (CIDD). (2025). Disorder types, CIDD. URL: https://cidd.discoveryspace.ca/disorder-types.html. [in Eng.].

Davies H., Noble P., Fins I.S., Pinchbeck G., Singleton D., Pirmohamed M., Killick D. (2025). Developing electronic health records as a source of real-world data for veterinary pharmacoepidemiology, Frontiers in Veterinary Science. — Vol. 12. — Article 1550468. DOI:10.3389/fvets.2025.1550468. [in Eng.].

Donner J., Anderson H., Davison S., Hughes A.M., Bouirmane J., Lindqvist J., Lytle K.M., Ganesan B., Ottka C., Ruotanen P., Kaukonen M., Forman O.P., Fretwell N., Cole C.A., Lohi H. (2018). Frequency and distribution of 152 genetic disease variants in over 100,000 mixed breed and purebred dogs, PLoS Genetics. — Vol. 14. — Article e1007361. DOI:10.1371/journal.pgen.1007361. [in Eng.].

Filho T.M. (2025). AKC data, GitHub Pages. URL: https://tmfilho.github.io/akcdata/.[in Eng.].

Harahap R.K., Wibowo E.P., Nur'ainingsih D., Wijaya A.K., Widyastuti N., R.A.S.C. (2022). Anindya, Dogs feed smart system with food scales indicator IoT based, 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS). Pp. 1–7. DOI:10.1109/icoris56080.2022.10031344. [in Eng.].

Lisnevskyi R., Biloshchytska S., Kushkimbayeva D., Kryschniy P., Lisnevskyi V. (2025). Development of a blockchain platform for logistics order management using technologies, CEUR Workshop Proceedings. Pp. 3966. [in Eng.].

Lustgarten J.L., Zehnder A., Shipman W., Gancher E., Webb T.L. (2020). Veterinary informatics: forging the future between veterinary medicine, human medicine, and One Health initiatives—a joint paper by the Association for Veterinary Informatics (AVI) and the CTSA One Health Alliance (COHA). JAMIA Open. — Vol. 3. — Pp. 306–317. DOI:10.1093/jamiaopen/ooaa005. [in Eng.].

Mullen K.R., Tammen I., Matentzoglu N.A., Mather M., Balhoff J.P., Esdaile E., Leroy G., Park C.A., Rando H.M., Saklou N.T., Webb T.L., Vasilevsky N.A., Mungall C.J., Haendel M.A., Nicholas F.W., Toro S. (2025). The Vertebrate Breed Ontology: Towards effective breed data standardization. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11177956/#S8. [in Eng.].

National Genomics Data Center, Dog breed information, NGDC, 2025. URL: https://ngdc.cncb.ac.cn/idog/dogph/breed/getBreedByA2Z.action?initial=A. [in Eng.].

Neftissov A., Sarinova A., Kazambaev I., Kirichenko L., Lisnevskyi R., Rzayeva L. (2023). Development of the smart office concept, 2021 // IEEE International Conference on Smart Information Systems and Technologies (SIST). Pp. 382–386. DOI:10.1109/sist58284.2023.10223512. [in Eng.].

The Kennel Club, Breeds A to Z, The Kennel Club (2025). URL: https://www.thekennelclub.org.uk/search/breeds-a-to-z/.[in Eng.].

Oberbauer A.M., Belanger J.M., Bellumori T., Bannasch D.L., Famula T.R. (2015). Ten inherited disorders in purebred dogs by functional breed groupings, Canine Genetics and Epidemiology. — Vol. 2. — No. 21. DOI:10.1186/s40575-015-0021-x. [in Eng.].

Paynter A.N., Dunbar M.D., Creevy K.E., Ruple A. (2021). Veterinary big data: when data goes to the dogs. Animals. — Vol. 11. — No. 1872. DOI:10.3390/ani11071872. [in Eng.].

Perez G., He Y., Lyu Z., Chen Y., Howe N.R., Rando H.M. (2025). Standardizing canine breed data in veterinary records is challenging, but computer vision offers an alternative perspective on breed assignment // American Journal of Veterinary Research. Pp. 1–8. DOI:10.2460/ajvr.24.10.0315. [in Eng.].

Rathnayaka R., Anuththara K., Wickramasinghe R., Gimhana P., Weerasinghe L., Wimalaratne G. (2022). Intelligent system for skin disease detection of dogs with ontology based clinical information extraction, 2022 IEEE 13th Annual Ubiquitous Computing // Electronics & Mobile Communication Conference (UEMCON). Pp. 59–66. DOI:10.1109/uemcon54665.2022.9965696. [in Eng.].

Summers J.F., O'Neill D.G., Church D., Collins L., Sargan D., Brodbelt D.C. (2019). Health-related welfare prioritisation of canine disorders using electronic health records in primary care practice in the UK // BMC Veterinary Research. — Vol. 15. — Pp. 1902. DOI:10.1186/s12917-019-1902-0. [in Eng.].

Jackson J., Radford A.D., Belshaw Z., Wallis L.J., Kubinyi E., German A.J., Westgarth C. (2024). Using

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

veterinary health records at scale to investigate ageing dogs and their common issues in primary care // Journal of Small Animal Practice. DOI:10.1111/jsap.13809. [in Eng.].

Yue P., Tang H., Li W., Zhang W., Yan B. (2025). MLKGC: large language models for knowledge graph completion under multimodal augmentation, Mathematics. — Vol. 13. — Pp. 1463. DOI:10.3390/math13091463. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 269–284
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.016

COMPARATIVE EVALUATION MODEL BASED ON EXPERT SYSTEMS FOR ABSTRACTIVE SUMMARIZATION OF SCIENTIFIC TEXTS

D. Kantayeva*, A. Seisenbekkyzy, S. Biloshchytska
Astana IT University, Astana, Kazakhstan.
E-mail: d.kantayeva@astanait.edu.kz

Dina Kantayeva — Master of Technical Science, PhD student, Astana IT University, Astana, Kazakhstan
E-mail: d.kantayeva@astanait.edu.kz, https://orcid.org/0000-0002-4097-5078;
Arailym Seisenbekkyzy — Master of Technical Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0005-4907-8639;
Svitlana Biloshchytska —Doctor of Technical Sciences, Associate Professor, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-0856-5474.

© D. Kantayeva, A. Seisenbekkyzy, S. Biloshchytska

Abstract. The growing number of scientific publications requires effective methods of summarizing scientific texts. Traditional extractive approaches do not allow preserving the logical structure, while abstract generalization methods are difficult to process highly structured documents. In this study, a comparative assessment model is proposed to improve the quality of abstract generalization of scientific texts. This study presents a comparative evaluation of BART-base and BART-large models. A corpus of 17,038 documents from arXiv and PubMed was used for fine-tuning. Evaluation with ROUGE and BLEU metrics shows that BART-large outperforms BART-base, achieving ROUGE-1 = 0.2224, ROUGE-2 = 0.0695, and BLEU = 0.29. In contrast, BART-base reached ROUGE-1 = 0.1932 and BLEU = 0.25. BART-large demonstrated better phrase-level coherence.

Keywords: Abstractive summarization, BART, ROUGE, NLP, text generation, model evaluation

For citation: Kantayeva D., Seisenbekkyzy A., Biloshchytska S. (2026). Comparative evaluation model based on expert systems for abstractive summarization of scientific texts // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 269–284. https://doi. org/10.54309/IJICT.2026.25.1.016. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ҒЫЛЫМИ МӘТІНДЕРДІ АБСТРАКТИВТІ РЕФЕРАТТАУҒА АРНАЛҒАН ЭКСПЕРТТІК ЖҮЙЕЛЕРГЕ НЕГІЗДЕЛГЕН САЛЫСТЫРМАЛЫ БАҒАЛАУ МОДЕЛІ

Д. Кантаева, А. Сейсенбеккызы, С. Билощицкая*
Astana IT University, Астана, Қазақстан.
E-mail: d.kantayeva@astanait.edu.kz

Кантаева Дина — техника ғылымдарының магистрі, докторант, Astana IT University, Астана, Қазақстан
E-mail: d.kantayeva@astanait.edu.kz, https://orcid.org/0000-0002-4097-5078;
Сейсенбекқызы Арайлым — техника ғылымдарының магистрі, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0005-4907-8639;
Билощицкая Светлана — техника ғылымдарының докторы, қауымдастырылған профессор, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-0856-5474.

© Д. Кантаева, А. Сейсенбеккызы, С. Билощицкая

Аннотация. Ғылыми жарияланымдар санының артуы ғылыми мәтіндерді рефераттаудың тиімді әдістерін қолдануды қажет етеді. Дәстүрлі экстрактивті тәсілдер мәтіннің логикалық құрылымын сақтауға мүмкіндік бермейді, ал абстрактивті жалпылау әдістері құрылымы күрделі құжаттарды өңдеу барысында қиындықтарға тап болады. Осы зерттеуде ғылыми мәтіндерді абстрактивті жалпылау сапасын арттыруға бағытталған салыстырмалы бағалау моделі ұсынылады. Зерттеуде BART-base және BART-large модельдеріне салыстырмалы бағалау жүргізілді. Модельдерді қосымша оқыту үшін arXiv және PubMed дерекқорларынан алынған 17 038 құжаттан тұратын корпус пайдаланылды. ROUGE және BLEU метрикалары арқылы жүргізілген бағалау нәтижелері BART-large моделінің BART-base моделінен жоғары нәтиже көрсеткенін анықтады: ROUGE-1 = 0,2224, ROUGE-2 = 0,0695 және BLEU = 0,29. Ал BART-base моделі ROUGE-1 = 0,1932 және BLEU = 0,25 нәтижелеріне қол жеткізді. Сонымен қатар, BART-large моделі сөз тіркестері деңгейінде жоғарырақ мәтіндік үйлесімділік көрсетті.

Түйінді сөздер: абстрактивті рефераттау, BART, ROUGE, NLP, мәтін генерациясы, модельдерді бағалау

Дәйексөздер үшін: Кантаева Д., Сейсенбеккызы А., Билощицкая С. (2026). Ғылыми мәтіндерді абстрактивті рефераттауға арналған эксперттік жүйелерге негізделген салыстырмалы бағалау моделі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 269–284. https://doi.org/10.54309/IJICT.2026.25.1.016. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

СРАВНИТЕЛЬНАЯ МОДЕЛЬ ОЦЕНКИ НА ОСНОВЕ ЭКСПЕРТНЫХ СИСТЕМ ДЛЯ АБСТРАКТИВНОГО РЕФЕРИРОВАНИЯ НАУЧНЫХ ТЕКСТОВ

Д. Кантаева, А. Сейсенбеккызы, С. Билощицкая*
Astana IT University, Астана, Казахстан.
E-mail: d.kantayeva@astanait.edu.kz

Кантаева Дина — магистр технических наук, докторант, Astana IT University, Астана, Казахстан
E-mail: d.kantayeva@astanait.edu.kz, https://orcid.org/0000-0002-4097-5078;
Сейсенбеккызы Арайлым — магистр технических наук, Astana IT University, Астана, Казахстан
https://orcid.org/0009-0005-4907-8639;
Билощицкая Светлана — доктор технических наук, доцент, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-0856-5474.

© Д. Кантаева, А. Сейсенбеккызы, С. Билощицкая

Аннотация. Рост количества научных публикаций требует применения эффективных методов реферирования научных текстов. Традиционные экстрактивные подходы не позволяют сохранять логическую структуру текста, в то время как методы абстрактивного обобщения испытывают трудности при обработке документов со сложной структурой. В настоящем исследовании предложена сравнительная модель оценки, направленная на повышение качества абстрактивного обобщения научных текстов. В исследовании представлена сравнительная оценка моделей BART-base и BART-large. Для дообучения моделей использовался корпус, содержащий 17 038 документов из баз данных arXiv и PubMed. Оценка с использованием метрик ROUGE и BLEU показала, что модель BART-large превосходит BART-base, достигнув значений ROUGE-1 = 0,2224, ROUGE-2 = 0,0695 и BLEU = 0,29. В свою очередь, модель BART-base продемонстрировала значения ROUGE-1 = 0,1932 и BLEU = 0,25. Модель BART-large также продемонстрировала более высокую связность на уровне фраз.

Ключевые слова: абстрактивное реферирование, BART, ROUGE, NLP, генерация текста, оценка моделей

Для цитирования: Кантаева Д., Сейсенбеккызы А., Билощицкая С. (2026). Сравнительная модель оценки на основе экспертных систем для абстрактивного реферирования научных текстов // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 269–284. https://doi.org/10.54309/IJICT.2026.25.1.016. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Introduction.

The rapid growth of digitalization and the introduction of the Internet of Things and intelligent systems in areas such as fitness (Hladkyi et al., 2021), healthcare (Ananthula et al., 2024) and smart offices (Neftissov et al., 2023) have significantly increased the volume of text data. This highlights the need for automated approaches to information processing and synthesis. Existing large-scale datasets for abstracting usually cover small texts. The average length of articles in the CNN/Daily Mail dataset is about 600 words (Hermann et al., 2015). Some studies confirm the effectiveness of hybrid models combining statistical and semantic methods. (Larsen et al., 2010; Babenko et al., 2024; Biloshchytska et al., 2025). In this study, we develop a model for high-quality generalization of long texts.

The relevance of developing a method of abstract abstracting of scientific texts in English is due to the need to optimize scientific information. While research is mainly focused on extractive approaches (Nallapati et al., 2016; Vanderwende et al., 2007), modern neural network models such as seq2seq have shown promising results (Nallapati et al., 2016; See, et al., 2017; Li et al., 2017). However, when processing long and structured texts, they face difficulties in maintaining context (Shao et al., 2017).

The research hypothesis. The use of an integrated approach combining artificial intelligence and deep learning technologies will make it possible to develop an effective method for abstracting scientific texts in English. It is expected that this method will significantly improve the process of extracting key information from scientific texts.

Goals and objectives of the study. The purpose of the research is to develop and test a method of abstract abstracting of scientific texts in English, which provides automatic creation of concise and informative summaries reflecting the key content of the source documents.

Materials and methods.

The focus is on methods using neural networks and statistical approaches. There are extractive and abstract text-based summarization approaches. The output of the first approach is a summary generated from key sentences of the input text, the second generates new sentences that are not present in the input text (Jones et al., 2007).

Extractive statistical methods, such as frequency analysis of the Moon (Luhn, 1958), were effective for short texts. Later improvements with vector models (Yu et al., 2020) increased their accuracy, but they remain limited for long scientific texts.

Graph methods, such as LexRank (Erkan et al., 2004), have improved the quality of abstracting by considering the relationships between sentences, but are limited when working with scientific texts (Gunes, 2004). Later, Plaza (2008) proposed concept graphs with ontologies, which increased efficiency for specialized fields (Plaza et al., 2008). However, such methods require manual tuning that limits scalability. The summarized text generated by extractive methods is not carefully filtered and often contain unnecessary information. It may not catch overall value of the text if key information covers several sentences.

Abstract methods form new sentences, which makes summaries more informative

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and coherent. The first step was the Pointer-Generator model (See et al., 2017), followed by T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), which showed high generation quality, but sometimes created redundant texts. The use of attention mechanisms and hybrid architectures (Song et al., 2019; Bahdanau et al., 2014). improved the processing of semantics, and multilingual models (ruT5, mBART) confirmed their effectiveness for different languages. In general, abstract approaches are superior to extractive ones. This method handles more details in the text, but they require a lot of data for training.

The quality and relevance of the data directly determine the effectiveness of the model (Lewis et al., 2020); therefore, special attention is paid to scientific texts in English. For the study, several datasets were considered, analyzed according to the criteria of structure, relevance and suitability for abstract abstracting.

Table 1. Comparative analysis of datasets

Dataset	Source	Content	Advantages	Limitations	Suitability
Topic Modeling for Research Articles	Kaggle	Titles, Abstracts, Keywords	Structured metadata, thematic analysis	no full text	Low
Latest Research Articles	Kaggle	Titles, Full texts	Current publications	No scientific texts, no standard scientific sections	Average
Wikipedia Promotional Articles	Wikipedia, Kaggle	Promotional articles, Metadata, Full texts	The possibility of stylistic analysis	Non-scientific texts	Low
COVID-19 Vaccine Articles	Kaggle	Titles, Full texts	Full content	Limited to a narrow topic	Average
arXiv Dataset	arXiv, Kaggle	Titles, Abstracts, Links to PDFs, Metadata	Extensive storage, scalability	Text extraction is required	High
The arXive Dataset	arXiv, Mendeley	Full texts Abstracts, Sections, Metadata	Full content, clear structure	-	Very high

As shown in Table 1, most datasets (Kaggle, Wikipedia) are limited by the lack of full texts or narrow topics, whereas «The arXiv Dataset» (Mendeley Data, 2021), created on the basis of arXiv and PubMed, with full texts, annotations and section structure, proved to be the most suitable for abstract abstracting due to its scale and high-quality preparation data.

Table 2. Description of the cleaned datasets

	arXive	PubMed		
	Text length	Abstract length	Text length	Abstract length
Count	17 038			
Mean	263.44	11.75	88.89	6.85
Std	102.57	2.13	60.56	2.88
Min	100.00	10.00	3.00	1.00
25%	179.00	10.00	47.00	5.00
50%	252.00	11.00	77.00	7.00
75%	338.00	13.00	113.00	9.00
Max	500.00	20.00	1887.00	23.00

Combining and cleaning data from these sources resulted in the creation of a

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

final set of 17,038 records (see Table 2).

The data set includes 17,038 records and 35 columns containing identifiers, full texts, annotations, and metrics, as shown in Figure 1. For quality control, a selective validation of the astro-ph0202198 article was performed, confirming a match with the original on the arXiv.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17038 entries, 0 to 17037
Data columns (total 35 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   article_id                  17038 non-null  object
 1   article_text                17038 non-null  object
 2   abstract_text               17038 non-null  object
 3   section_names               17038 non-null  object
 4   sections                    17038 non-null  object
 5   len_text                    17038 non-null  int64
 6   len_abstract                17038 non-null  int64
 7   sent_ext                    17038 non-null  float64
 8   word_ext                    17038 non-null  float64
 9   word_inf                    17038 non-null  float64
 10  best_rouge1_wih             17038 non-null  float64
 11  best_rouge2_wih             17038 non-null  float64
 12  best_summary_wih            17038 non-null  object
 13  best_summary_vns            17038 non-null  object
 14  best_rouge1_vns             17038 non-null  float64
 15  best_rouge2_vns             17038 non-null  float64
 16  voted_summary               17038 non-null  object
 17  voted_method                17038 non-null  object
 18  voted_r1                    17038 non-null  float64
 19  voted_r2                    17038 non-null  float64
 20  best_summary_vns_initgreed  17038 non-null  object
 21  best_rouge1_vns_initgreed   17038 non-null  float64
 22  best_rouge2_vns_initgreed   17038 non-null  float64
 23  len_sum_vns                 17038 non-null  int64
 24  len_sum_wih                 17038 non-null  int64
 25  len_sum_vot                 17038 non-null  int64
 26  len_sum_vns_init_greed      17038 non-null  int64
 27  best_summary_gen            17038 non-null  object
 28  best_rouge1_gen             17038 non-null  float64
 29  best_rouge2_gen             17038 non-null  float64
 30  len_sum_gen                 17038 non-null  int64
 31  best_summary_gen_greedinit  17038 non-null  object
 32  best_rouge1_gen_greedinit   17038 non-null  float64
 33  best_rouge2_gen_greedinit   17038 non-null  float64
 34  len_sum_gen_greedinit       17038 non-null  int64
dtypes: float64(15), int64(8), object(12)
memory usage: 4.5+ MB
```

Fig. 1. Dataset structure.

After a data cleaning the texts were reduced to lowercase, excluded HTML and URLs, normalized and tokenized; signs of length and number of sentences were created.

Data analysis revealed that articles contain an average of 200–250 words, annotations – 10–12 sentences, accounting for 15–25 % of the text (see Figures 2-4).

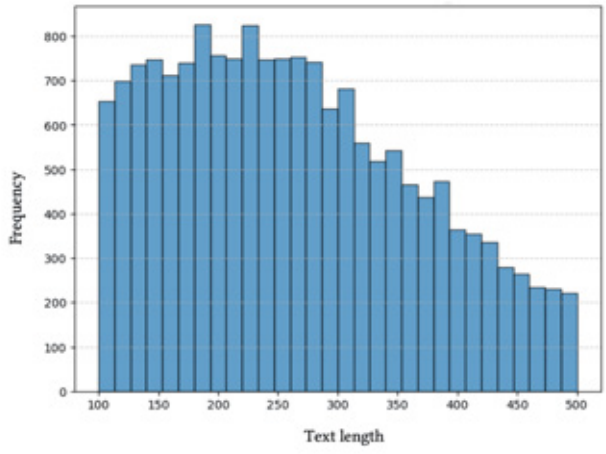


Fig. 2. Text length distribution.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

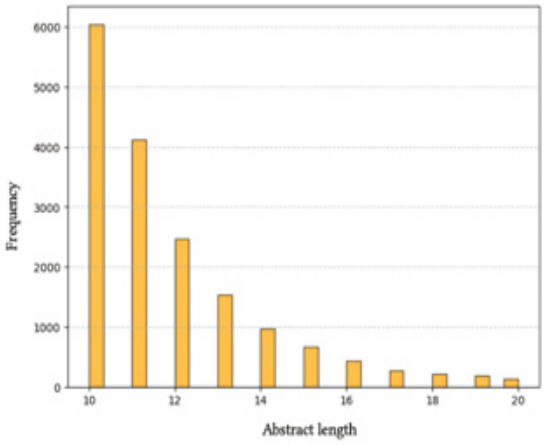


Fig. 3. Abstract length distribution.

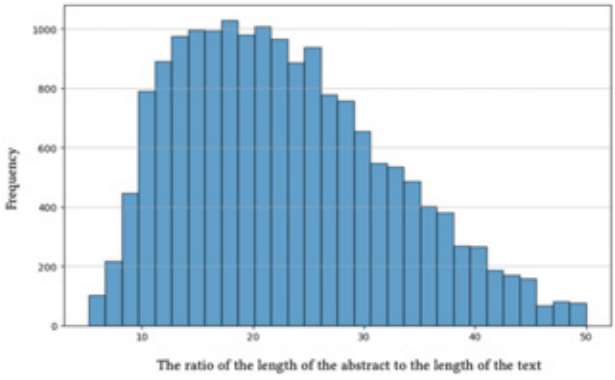


Fig. 4. The ratio of the length of the abstract to the length of the text.

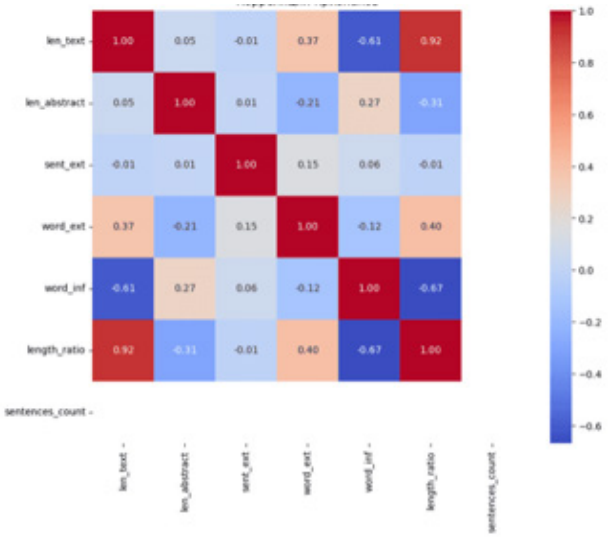


Fig. 5. Correlation of features.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

As shown in Figure 5 long texts contain fewer information-rich words.

Data preparation. To implement the task of automatic abstracting of scientific texts, the main columns were selected: the full texts of articles and their annotations. The data was then divided into training, validation, and test sets in the proportions of 80 %, 10 %, and 10 %. The total number of examples is shown in Table 3.

Table 3. Data separation

Dataset	Number of examples
Train	17 038
Validation	1 704
Test	1 704

Abstractive summarization is one of the key tasks of natural language processing and is aimed at generating annotations that convey the main ideas of the source text using new formulations. The approach is based on maximizing the conditional probability $P(Y|X)$, where X is the source text and Y is the abstract, and is implemented using modern deep learning architectures. The most widely used models are Seq2Seq with the attention mechanism, Transformer, BART, T5, and GPT.

A model of Sequence-to-Sequence (Seq2Seq) is used to convert the input sequence $X = \{x_1, x_2, ..., x_n\}$ in the output sequence $Y = \{y_1, y_2, ..., y_m\}$ even when the difference in their lengths. The architecture includes an encoder and a decoder interacting through a hidden representation.

The encoder generates hidden states:

$$h_t = f(h_t^{-1}, x_t) \tag{1}$$

where f is implemented via GRU or LSTM. The final state of h_n is passed to the decoder, which generates output tokens step by step, updating the state using the formula:

$$s_t = g(s_t^{-1}, y_t^{-1}, h), \tag{2}$$

where g is a nonlinear function like the function f. The probability of predicting the entire output sequence is defined as:

$$P(Y \vee X) = \prod_{t=1}^{m} P(y_t \vee y_1, \dots, y_{t-1}, h), \tag{3}$$

To improve the quality of predictions, an attention mechanism is used, which allows the decoder to focus on the most significant parts of the input sequence when generating each token. The significance weight of the input token α_{ti} is calculated as:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{n} exp(e_{t,i})}, \quad (4)$$

where e_{ti} = score (s_t, h_i) is a function estimating the importance of the i-th input token for generating the t-th output token (Zhang et al., 2024).

The main advantage of BART is the combination of a bidirectional encoder and an autoregressive decoder, which provides context understanding and generation of new formulations. The model is trained on the task of restoring distorted text X from its modified version X~, which uses word deletion, sentence rearrangement, and noise addition.

At the encoding stage, a hidden representation is formed.:

$$H = Encoder(X), \quad (5)$$

where $H = \{H_1, h_2, ..., h_n\}$ represents the hidden states of the encoder. At the decoding stage, the model sequentially generates the output sequence $Y = \{y_1, y_2, ..., y_m\}$. The probability of the output sequence is calculated as follows:

$$P(Y \vee X) = \prod_{t=1}^{m} P(y_t \vee y_1, \dots, y_{t-1}, h), \quad (6)$$

where $P(y_t \mid y_1, ..., y_{t-1}, H)$ is the probability of the next token based on previously generated tokens and the hidden representation of the encoder (Tsvetkov et al., 2008).

Generative Pre-trained Transformer (GPT) is pre-trained on large bodies of texts, which ensures its effectiveness in language processing tasks, including abstract abstracting. The basis is an autoregressive mechanism, where each y_t token is predicted based on the previous $\{y_1, y_2, ..., y_{t-1}\}$:

$$P(Y) = \prod_{t=1}^{m} P(y_t \vee y_1, \dots, y_{t-1}), \quad (7)$$

The model is based on a transformer decoder and an attentional mechanism.:

$$Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d^k}}\right)V, \quad (8)$$

where Q, K, and V are matrices of queries, keys, and values, and d_k is the dimension of the keys. This formula provides the calculation of weighted values, which helps the model to better understand the context of the text (Zhong et al., 2024).

The BART model was chosen for the study, as it combines bidirectional encoding

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and auto-correction, processes context efficiently, is resistant to data skips, and shows stable results with limited training samples.

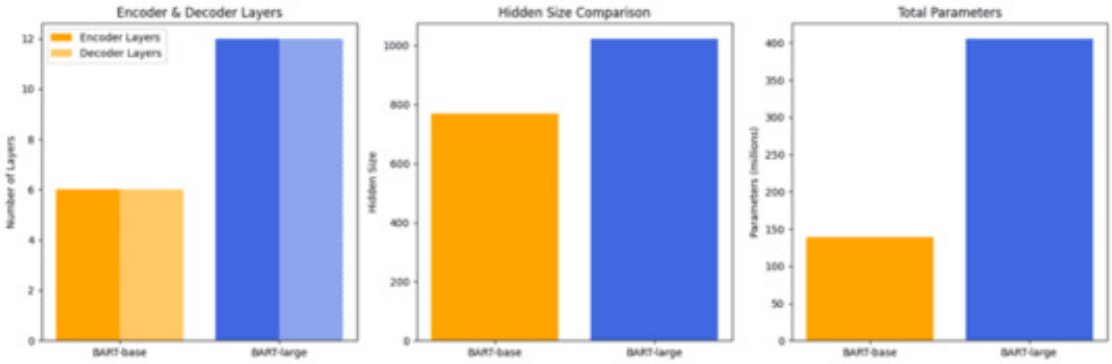


Fig. 6. Comparison of BART-base and BART-large architectures.

Depending on the computing capabilities and the requirements for the quality of text generation, BART is presented in two versions (see Figure 6): BART-base and BART-large. Each of them has its own advantages and limitations, which determines their application in various scenarios of summarization.

Table 4. Comparison of BART models

	BART-base	BART-large
Specifications	12 layers, 768 hidden, 16 heads, 139M parameters	24 layers, 1024 hidden, 16 heads, 406M parameters
Features	Compact version optimized for limited resources	Deep architecture, improved context understanding
Advantages	Balance between quality and cost	Higher accuracy, better connection detection
Application	Text generation, abstract abstracting, Machine translation, QA	Long and complex texts, scientific articles

Due to the fixed size of the input context of the model, the maximum length of sequences was set: 512 tokens for article introductions and 150 tokens for annotations. This limitation is since BART supports up to 1024 tokens, but 512 are enough for introductions of scientific articles, and the average length of annotations does not exceed 150 words, as shown in Table 5.

Table 5. Limiting the length of sequences

Length	Value	Justification
Input data	512 tokens	BART supports up to 1024 tokens, but 512 is enough for introductions of scientific articles.
Output data	150 tokens	The average length of scientific abstracts does not exceed 150 words.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

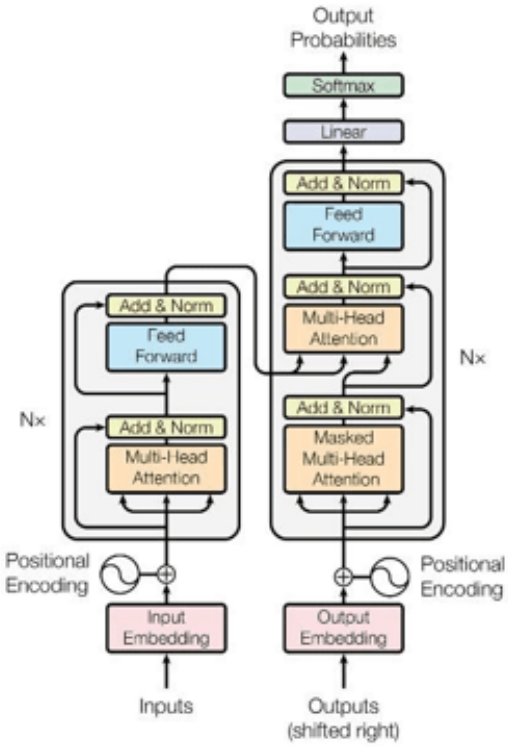


Fig. 7. Encoder-Decoder architecture (Source: (Brown et al., 2020)).

BART-base and BART-large allow you to paraphrase the source text while maintaining its structural integrity. BART-base provides a balance between quality and computational costs, which makes it convenient for practical implementation, while BART-large demonstrates maximum accuracy through deep context processing.

To assess the quality of automatic abstracting, metrics are used that compare generated annotations with reference ones. The main one is ROUGE, a standard in summation that measures match at the level of n-grams and word sequences. Additionally, the BLEU metric was used.

BLEU is one of the main metrics for evaluating the quality of automatically generated text, originally used in machine translation (Brown et al., 2020). It compares the generated text with the reference text based on the coincidence of word sequences (n-grams).

To avoid a situation where the model creates texts that are too short and have a high overlap, a brevity penalty (BP) is introduced:

$$BP = \begin{cases} 1, if c > r \\ e^{\left(1-\frac{r}{c}\right)}, if c \leq r \end{cases}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

where *c* is the length of the generated text (hypothesis), and *r* is the length of the reference text. If the generated text is shorter than the reference text, the penalty reduces the final BLEU score. The final formula for calculating BLEU includes a geometric mean of n-gram accuracy and a brevity penalty:

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} \omega_n \log P_n\right)$$

where P_n is the accuracy of n-gram matches, ω_n are the weights for different n-grams (usually assumed to be equal), and N is the maximum length of n-grams.

BLEU is focused on precision, whereas ROUGE is primarily concerned with completeness. This means that BLEU better evaluates the correspondence of the generated text to the reference in terms of the accuracy of individual phrases, while ROUGE measures how fully the source text was reflected in the final annotation.

Recall-oriented understudy for gisting evaluation (ROUGE)

ROUGE metrics, which are the standard in automatic abstracting tasks, were used to evaluate the quality of annotations (Lewis et al., 2019).

ROUGE-1 measures word matches (unigrams):

$$ROUGE - 1 = \frac{\sum_{\omega \in W} min\left(count_{gen}(\omega), count_{ref}(\omega)\right)}{\sum_{\omega \in W} count_{ref}(\omega)}$$

where $count_{gen}(\omega)$ — is the number of occurrences of the word ω in the generated text, $count_{ref}(\omega)$ — is the number of occurrences of a word in the reference annotation, and *W* is the set of all words in the reference text.

The ROUGE-2 metric evaluates the coincidence of pairs of consecutive words (bigrams), which makes it possible to identify the preservation of local context and sentence coherence:

$$ROUGE - 2 = \frac{\sum_{b \in B} min\left(count_{gen}(b), count_{ref}(b)\right)}{\sum_{b \in B} count_{ref}(b)}$$

where *B* is the set of all bigrams in the reference text.

The ROUGE-L metric is based on the search for the Longest Common Sequence (LCS), which allows you to evaluate the grammatical and structural consistency between the generated and reference text.

This metric considers the word order and is calculated as follows:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

$$ROUGE - L = \frac{LCS(gen, ref)}{length(ref)}$$

where $LCS(gen, ref)$ — is the length of the largest common subsequence between the generated text and the reference annotation.

ROUGE metrics are limited by lexical overlap and do not consider semantics, which may underestimate the quality of annotations. In the future, it is worth using BERTScore and METEOR metrics that consider semantic connections.

Results and discussion.

To assess the quality of the annotations, the pre-trained and pre-trained versions of BART-base and BART-large were compared on the corpus of scientific articles using the ROUGE metric.

Table 6. Results of pre-trained models on the test set

	Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0	BART-base	0.1803	0.0417	0.0960	0.0961
1	BART-large	0.1881	0.0507	0.1088	0.1088

At the first stage of testing, the pre-trained versions of BART-base and BART-large were evaluated without additional training based on scientific articles. The results presented in Table 6 demonstrate the initial performance of the models.

Despite the more complex architecture of BART-large, its performance is only slightly higher than that of BART-base. This indicates that without additional training in a specialized corpus, models are not able to effectively consider the features of a scientific text. The main reason for this limited effectiveness is that pre-trained transformational models, including BART, are most often trained on the corpus of news articles, social networks, and other sources focused on generalized language rather than an academic style of presentation.

To improve the quality of summarization of scientific texts, both models, BART-base and BART-large, were further trained on a specialized corpus of articles. The main stages included marking up and preprocessing data, choosing a learning strategy, determining optimal hyperparameters, monitoring the learning process, and subsequent quality assessment.

The key parameters are the batch size, the learning rate, the optimization strategy, the number of epochs, and the schedule for changing the learning rate.

1. Batch size = 8 – the value is selected based on the balance between learning efficiency and available memory.

2. Learning rate = 3e-5 – the learning rate was set based on empirical observations and recommendations for setting up BART.

3. AdamW optimizer – was chosen as the main algorithm for updating weights because it works efficiently with large models by enabling weight decay.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

4. Weight decay = 0.01 – this parameter prevents overfitting by limiting the amount of weight updates.

5. Number of epochs = 3 and 5 – the model was tested after 3 and 5 epochs, since preliminary experiments showed that with prolonged training, the model ceases to improve significantly.

In addition, a scheduler was used, which allowed the model to gradually reduce the learning rate, stabilizing the process of updating weights.

The results of BART-base retraining are presented in Table 7, which shows the ROUGE metrics after epochs 3 and 5.

Table 7. Results of the retrained BART-base model

	Epochs	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0	3	0.1932	0.0598	0.1234	0.1235
1	5	0.1932	0.0598	0.1234	0.1235

As can be seen from the table, the results after 5 epochs did not change compared to 3 epochs, which indicates that the model has reached its learning limit on this corpus.

Further training of the BART-large model showed a more significant increase in all metrics, which is reflected in Table 8.

Table 8. Results of the retrained BART-large model

	Epochs	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0	3	0.2224	0.0695	0.1355	0.1356

Compared to BART-base, the BART-large model has shown significant improvement, which confirms the effectiveness of a deeper architecture. The increase in ROUGE-2 is especially noticeable, which indicates a better preservation of the coherence of the text and bigram dependencies.

At the first stage, pre-trained BART-base and BART-large models were tested, which showed low ROUGE values due to lack of adaptation to scientific discourse, while BART-large gave slightly better results. After further training on the corpus of scientific articles, the indicators improved significantly: the increase in ROUGE-1 was +1.29% for BART-base and +3.43% for BART-large. Additional training (from 3 to 5 epochs) for BART-base did not increase the metrics, which indicates the efficiency limit.

Table 9. Test results of pre-trained and fine-tuned models

Model	Epochs	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
BART-base (without fine-tuning)	–	0.1803	0.0417	0.0960	0.0961
BART-large (without fine-tuning)	–	0.1881	0.0507	0.1088	0.1088
BART-base (fine-tuned)	3	0.1932	0.0598	0.1234	0.1235
BART-base (fine-tuned)	5	0.1932	0.0598	0.1234	0.1235
BART-large (fine-tuned)	3	0.2224	0.0695	0.1355	0.1356

BART-large showed the largest increase, increasing ROUGE-2 by 1.88%, which

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

confirms its better ability to take into account the context and form coherent annotations. A comparison with BART-base shows that the depth of the architecture and the number of parameters directly affect the quality of the abstraction.

Discussion.

The findings confirm the benefits of fine-tuning pretrained summarization models. BART-large demonstrates higher performance due to its deeper architecture and larger parameter set. BART-base is suitable for environments with limited computational resources. However, for complex scientific texts with structured sections deeper models like BART-large are preferable.

A limitation of this study is the exclusive use of lexical metrics, which are known to underestimate semantic metrics. Although ROUGE and BLEU capture surface overlap, they fail to reflect meaning preservation or factual consistency. Thus, incorporating semantic metrics such as BERTScore or BLEURT in future evaluations is essential.

Additionally, this study does not include human evaluation, which is critical for assessing fluency, coherence, and readability. The effect of summarization on literature review speeding or research indexing also remains unexplored.

Conclusion.

Most abstracting models, including BART, are pre-trained on general-purpose corpora, which limits their effectiveness for scientific texts with a strict structure and specialized vocabulary. The study evaluated BART-base and BART-large, which showed low ROUGE values without adaptation and significant improvement after further training on the corpus of scientific articles. BART-large, surpassed BART-base in all metrics, especially for ROUGE-2, which reflects better text connectivity. The results confirm the need to adapt transformational models to the scientific language and further develop approaches based on semantic metrics, specialized models and postprocessing methods. The results demonstrate that fine-tuning improves the quality of summaries, particularly for BART-large, which achieved a ROUGE-1 score of 0.2224 and BLEU score of 0.29. Future work will focus on expanding comparisons to include other models (e.g., T5, PEGASUS) and conducting human evaluation.

REFERENCES

Ananthula B., Budde N., Baranovskyi O., Babenko T., Bigdan A., Lisnevskyi R. (2024). In-depth Study of Intelligent Healthcare Systems Using 5G Security, CEUR Workshop Proceedings. — Vol. 3680. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192549346.

Biloshchytska S., Tleubayeva A., Kuchanskyi O., Biloshchytskyi A., Andrashko Y., Toxanov S., Mukhatayev A., Sharipova S. (2025). Text Similarity Detection in Agglutinative Languages: A Case Study of Kazakh Using Hybrid N-gram and Semantic Models. Applied Sciences. — Vol. 15. — No. 12. P. 6707. DOI: 10.3390/app15126707.

Bahdanau D., Cho K., Bengio Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate, arXiv preprint arXiv:1409.0473.

Babenko T., Kolesnikova K., Lisnevskyi R., Makilenov S., Landovsky Y. (2024). Definition of Cryptojacking Indicators, CEUR Workshop Proceedings. — Vol. 3680. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192508442

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A.,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Sutskever I., Amodei D. (2020). Language Models are Few-Shot Learners // arXiv.

Batura T.V., Bakieva A.M. (2019). Methods and Systems of Automatic Text Summarization. — Novosibirsk: IPC NSU.

Gritsenko T.Yu. (2020). Analysis of problems of automatic text summarization algorithms, Dnevnik Nauki. No. 5. Pp. 25–25.

Gunes R.R. Dragomir. (1958). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization // Journal of Artificial Intelligence Research. — Vol. 22. — Pp. 457–479.

Hermann K.M., Kocisky T., Grefenstette E., Espeholt L., Kay W., Suleyman M., Blunsom P. (2015). Teaching machines to read and comprehend, Advances in Neural Information Processing Systems, Pp. 1693–1701.

Hladkyi Y., Gladka M., Kostikov M., Lisnevskyi R. (2021). An IoT Solution: A Fitness Trainer, CEUR Workshop Proceedings. — Vol. 3179. — Pp. 215–226. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136276676.

Jones K.S. (2007). Automatic summarising: The state of the art, Information Processing & Management. — Vol. 43. — No. 6. Pp. 1449–1481.

Lin C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proc. ACL Workshop. Pp. 74–81.

Luhn H.P. (1958). The Automatic Creation of Literature Abstracts // IBM Journal of Research and Development. — Vol. 2. — No. 2. Pp. 159–165.

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension // arXiv. — Vol. 36 — Pp. 8037–8050.

Larsen B., Van I. A. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index, Scientometrica. — Vol. 84. — No. 3. Pp. 575–603.

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, Proc. ACL.

Li P., Lam W., Bing L., Guo W., Li H. (2017). Cascaded attention based unsupervised information distillation for compressive summarization, Proc. EMNLP. Pp. 2071–2080.

Neftissov A., Sarinova A., Kazambaev I., Kirichenko L., Lisnevskyi R., Rzayeva L. (2023). Development of the Smart Office Concept, IEEE Int. Conf. on Smart Information Systems and Technologies (SIST). Pp. 382–386. DOI: 10.1109/SIST58284.2023.10223512.

Nallapati R., Zhai F., Zhou B. (2016). SummaRuNNer: A recurrent neural network-based sequence model for extractive summarization of documents. Proc. 31st AAAI Conf. on Artificial Intelligence. Pp. 3075–3081.

Papineni K., Roukos S., Ward T., Zhu W. (2002). BLEU: A method for automatic evaluation of machine translation, Proc. 40th Annual Meeting of the ACL, Pp. 311–318.

Plaza L., Diaz A., Gervas P. (2008). Concept-Graph Based Biomedical Automatic Summarization Using Ontologies, Proc. Int. Conf. on Computational Linguistics. Pp. 53–56.

Raffel C., Shazeer N.M., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P.J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, J. Machine Learning Research. — Vol. 21. — No. 140. Pp. 1–67.

See A., Liu P.J., Manning C.D. (2017). Get to the point: Summarization with pointer-generator networks. Proc. Association for Computational Linguistics (ACL). URL: https://arxiv.org/abs/1704.04368.

Shao Y., Gouws S., Britz D., Goldie A., Strope B., Kurzweil R. (2017). Generating high-quality and informative conversation responses with sequence-to-sequence models, Proc. EMNLP. Pp. 2210–2219.

Song S., Huang H., Ruan T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications. — Vol. 78. — Pp. 857–875.

Tsvetkov V.Ya. & Domnitskaya E.V. (2008). On the methodology of data collection // Fundamental Research. No. 1. Pp. 52–53.

Vanderwende L., Suzuki H., Brockett C., Nenkova A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion, Information Processing & Management. — Vol. 43. — No. 6. Pp. 1606–1618.

Zhang M., Li C., Wan M., Zhang X., Zhao Q. (2024). ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics. Expert Systems with Applications. — Vol. 237.

Zhong Q., Ding L., Liu J. (2024). Encoding-Enhanced Sequence-to-Sequence Pretraining for Language Understanding and Generation // IEEE Transactions on Knowledge and Data Engineering. — Vol. 36 — Pp. 8037–8050.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 285–314
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.017

DEEP LEARNING-BASED VEHICLE MAKE AND MODEL RECOGNITION: DATASET CHALLENGES AND INFLUENCING FACTORS IN URBAN TRAFFIC

N. Mordvintsev, Zh. Oralbekova, A. Bakiyeva*
Astana IT University, Astana, Kazakhstan.
E-mail: Zh.Oralbekova@astanait.edu.kz

Nikita Mordvintsev — Master's student, Astana IT University, Astana, Kazakhstan https://orcid.org/0009-0008-9112-6229;
Zhanar Oralbekova — PhD, Associate Professor, Astana IT University, Astana, Kazakhstan
E-mail: Zh.Oralbekova@astanait.edu.kz, https://orcid.org/0000-0002-3726-8882;
Aigerim Bakiyeva — Candidate of Technical Sciences (PhD), Assistant Professor, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-7186-6408.

© N. Mordvintsev, Z. Oralbekova, A. Bakiyeva

Abstract. This paper presents a deep learning-based approach to vehicle make and model recognition (VMMR). Using ResNet50, AlexNet, VGG19, MobileNet, EfficientNet, VIT b16, ConvNeXt, SimCLR, CLIP we address the challenges of annotation of data sets, class imbalance, and quality of real-world cameras. A data set of 283,899 vehicle images was extracted and tested, achieving validation accuracy of 0.07 to 0.32. We analyze how vehicle size, acceleration, and turning radius affect detection performance and propose directions for optimizing smart city traffic systems.

Keywords: Deep Learning, Vehicle Recognition, VMMR, ResNet50, Dataset Augmentation, Intelligent Transportation Systems, Urban Traffic, Model Classification, Smart Cities

For citation: Mordvintsev N., Oralbekova Z., Bakiyeva A. (2026). Deep learning-based vehicle make and model recognition: dataset challenges and influencing factors in urban traffic // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 285–314. https://doi. org/10.54309/IJICT.2026.26.2.017. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ТЕРЕҢ ОҚЫТУҒА НЕГІЗДЕЛГЕН КӨЛІК ҚҰРАЛДАРЫНЫҢ МАРКАСЫ МЕН МОДЕЛІН ТАНУ: ҚАЛАЛЫҚ КӨЛІК

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ҚОЗҒАЛЫСЫНДАҒЫ ДЕРЕКТЕР ЖИЫНТЫҒЫНЫҢ МӘСЕЛЕЛЕРІ ЖӘНЕ ЫҚПАЛ ЕТУШІ ФАКТОРЛАР

Н. Мордвинцев, Ж. Оралбекова, А. Бакиева*
Astana IT University, Астана, Қазақстан.
E-mail: Zh.Oralbekova@astanait.edu.kz

Мордвинцев Никита — магистрант, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0008-9112-6229;
Оралбекова Жанар — PhD, қауымдастырылған профессор, Astana IT University, Астана, Қазақстан
E-mail: Zh.Oralbekova@astanait.edu.kz, https://orcid.org/0000-0002-3726-8882;
Бакиева Айгерим — техника ғылымдарының кандидаты, PhD, ассистент-профессор, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-7186-6408.

© Н. Мордвинцев, Ж. Оралбекова, А. Бакиева

Аннотация. Мақалада терең оқытуға негізделген көлік құралдарының маркасы мен моделін тану (Vehicle Make and Model Recognition, VMMR) тәсілі ұсынылады. ResNet50, AlexNet, VGG19, MobileNet, EfficientNet, ViT-B/16, ConvNeXt, SimCLR және CLIP модельдерін пайдалана отырып, деректер жиынтығын аннотациялау, сыныптар арасындағы теңгерімсіздік және нақты жағдайларда камералар арқылы алынған кескіндердің сапасы сияқты мәселелер қарастырылады. Көлік құралдарының 283 899 кескінінен тұратын деректер жиынтығы қалыптастырылып, тестіленді; валидациялық деректер жиынындағы дәлдік 0,07-ден 0,32-ге дейінгі аралықта болды. Көлік құралының өлшемі, үдеуі және бұрылу радиусы сияқты факторлардың тану тиімділігіне әсері талданып, ақылды қалалардың көлік жүйелерін оңтайландыру бағыттары ұсынылды.

Түйінді сөздер: терең оқыту, көлік құралдарын тану, VMMR, ResNet50, деректер жиынтығын аугментациялау, интеллектуалды көлік жүйелері, қалалық көлік қозғалысы, модельдерді жіктеу, ақылды қалалар

Дәйексөздер үшін: Мордвинцев Н., Оралбекова Ж., Бакиева А. (2026). Терең оқытуға негізделген көлік құралдарының маркасы мен моделін тану: қалалық көлік қозғалысындағы деректер жиынтығының мәселелері және ықпал етуші факторлар // халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 285–314. https://doi.org/10.54309/IJICT.2026.25.1.017. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАСПОЗНАВАНИЕ МАРКИ И МОДЕЛИ ТРАНСПОРТНЫХ СРЕДСТВ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ: ПРОБЛЕМЫ НАБОРОВ ДАННЫХ

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

И ФАКТОРЫ, ВЛИЯЮЩИЕ НА РАБОТУ СИСТЕМЫ В УСЛОВИЯХ ГОРОДСКОГО ДВИЖЕНИЯ

Н. Мордвинцев, Ж. Оралбекова, А. Бакиева*
Astana IT University, Астана, Казахстан.
E-mail: Zh.Oralbekova@astanait.edu.kz

Мордвинцев Никита — магистрант, Astana IT University, Астана, Казахстан https://orcid.org/0009-0008-9112-6229;
Оралбекова Жанар — PhD, ассоциированный профессор, Astana IT University, Астана, Казахстан
E-mail: Zh.Oralbekova@astanait.edu.kz, https://orcid.org/0000-0002-3726-8882;
Бакиева Айгерим — К.т.н., PhD, Ассистент-профессор, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-7186-6408.

© Н. Мордвинцев, Ж. Оралбекова, А. Бакиева

Аннотация. В статье представлен подход к распознаванию марки и модели транспортных средств (Vehicle Make and Model Recognition, VMMR) на основе глубокого обучения. С использованием моделей ResNet50, AlexNet, VGG19, MobileNet, EfficientNet, ViT-B/16, ConvNeXt, SimCLR и CLIP рассматриваются проблемы аннотирования наборов данных, дисбаланса классов и качества изображений, получаемых камерами в реальных условиях. Был сформирован и протестирован набор данных, содержащий 283 899 изображений транспортных средств; точность на валидационной выборке составила от 0,07 до 0,32. Проанализировано влияние размера транспортного средства, ускорения и радиуса поворота на эффективность распознавания, а также предложены направления оптимизации транспортных систем умного города.

Ключевые слова: глубокое обучение, распознавание транспортных средств, VMMR, ResNet50, аугментация набора данных, интеллектуальные транспортные системы, городской транспортный поток, классификация моделей, умные города

Для цитирования: Мордвинцев Н., Оралбекова Ж., Бакиева А. (2026). Распознавание марки и модели транспортных средств на основе глубокого обучения: проблемы наборов данных и факторы, влияющие на работу системы в условиях городского движения // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 285-314. https://doi.org/10.54309/IJICT.2026.25.1.017. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Deep learning has become a fundamental approach for vehicle make and model recognition (VMMR) within intelligent transportation systems (ITS). Convolutional Neural Networks (CNNs) are widely employed to automatically extract hierarchical features from vehicle images, progressing from low-level visual cues to high-level semantic representations. Various architectures have been proposed to enhance VMMR performance, including semi-supervised CNNs, coarse-to-fine frameworks for fine-grained classification, and hybrid methods integrating dimensionality reduction techniques such as Principal Component Analysis (PCA) to accelerate training. These deep learning models ensure reliable recognition from multiple perspectives (Jabeen et al., 2021). Also current paper made comparison with more advanced models like MobileNet, EfficientNet and Vision Transformer VIT b16, ConvNeXt, SimCLR, CLIP. In deep learning, channel fusion methods play a vital role in improving model performance, mainly through concatenation, summation, and multiplication. Concatenation merges feature maps from different levels along the depth dimension, improving the model's ability to represent features. This is often used in encoder-decoder architectures where low- level and high-level features are combined.

An analyzed study shows that the ITS system includes dynamic data transfer, where the machine model definition will help to provide some data (Vahora et al., 2025). Conventional traffic monitoring systems depend largely on manual observation or basic sensor-based methods, which restrict their accuracy, coverage, and real-time responsiveness. The integration of Deep Learning (DL) and Computer Vision (CV) has emerged as an advanced approach for intelligent traffic surveillance, enabling automated, scalable, and adaptive systems capable of analyzing and responding to traffic situations in real time without human intervention (Siva et al., 2025). Intelligent traffic lights, smart parking guidance systems, and advanced public transportation technologies are employed to enhance traffic flow, alleviate congestion, and boost overall traffic efficiency. By leveraging the centralized learning and decentralized execution framework of the Actor- Critic strategy gradient algorithm, additional information is utilized by critics to streamline the traffic signal control process and achieve optimized traffic signal management (Wei et al., 2024). In the article of the analyzed research, the author includes queue length, average speed, and ambulance data in the system architecture. Deep learning on the semantic data of cars adds data on the model and type of car to the architectures, and therefore more detailed physical data on car models (Girhe et al., 2022). While several studies have applied deep learning to vehicle make and model recognition (VMMR), this research uniquely focuses on adapting and evaluating state-of-the-art CNN architectures (ResNet50, VGG19, AlexNet, MobileNet, EfficientNet, VIT b16, ConvNeXt, SimCLR, CLIP). By incorporating physical characteristics of vehicles (such as acceleration, size, and turning radius) into the analysis, this work provides a multi-dimensional framework that goes beyond image classification – aiming to support smart city traffic optimization efforts.

Materials and methods.

Building a deep learning system architecture component from camera data to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

obtain data on traffic volume, car type, car brand, and hence subsequent aspects about each car and aggregated road line values. Maintaining and overcoming the difficulties of maintaining a deep learning model in the form of creating datasets.

How can deep learning techniques be applied to develop an accurate and efficient car brand detection system for traffic roads, and how can we create a reliable dataset using existing tools for training these models?

Objectives:

● Identifying the factors related to car models that influence city traffic and investigating these factors for specific brands through official vehicle documentation;

● Analyze modern approaches to data processing and analysis, focusing on existing methods and technologies for vehicle detection and identification;

●Train models on the identified datasets and evaluate their performance.

ResNet50 is a widely used deep learning model from the ResNet series, consisting of a structured arrangement of Bottleneck modules–specifically, 3, 4, 6, and 3 modules in its second to fifth layers, respectively. These modules include two distinct block types: Conv Blocks and Identity Blocks. Positioned between shallow and deep networks, ResNet50 balances model complexity and performance, making it well-suited for small-scale image datasets while reducing overfitting risks. While primarily designed for image classification, ResNet50's architecture can be further optimized for specialized tasks. For instance, enhancements such as a detail extraction module, which incorporates differential convolution for improved texture and contour recognition, can refine feature detection. Additionally, multi-channel fusion techniques, such as incorporating a binary mask, help the model focus on essential image regions, improving classification accuracy (Wang et al., 2022).

ResNet50 is a convolutional neural network architecture developed by researchers at Microsoft Research and introduced at the CVPR conference in 2015. It consists of 50 layers, with batch normalization applied after each convolution to ensure more stable and faster training. The network uses the ReLU (Rectified Linear Unit) activation function. Following the initial layer, a Max Pooling layer is applied to reduce the dimensionality of the data. At the end of the model, Global Average Pooling is used to reduce the number of parameters and prevent overfitting. ResNet50 is widely used for image classification, object detection, and segmentation tasks. In this work, a pre-trained ResNet50 model from torchvision.models was used, with the final layer replaced to predict one of 43 classes. The model was fine-tuned, freezing all layers except the last one. The CrossEntropyLoss function was used for training, with the Adam optimizer set at a learning rate of 0.001. The model was trained for 10 epochs (Sharma et al., 2023).

ResNet (Residual Network) is a deep convolutional neural network (CNN) architecture introduced by He et al. in 2015. It introduced residual learning, which uses skip connections (shortcuts) to bypass certain layers, addressing the vanishing gradient problem in very deep networks. This innovation allowed for the training of extremely deep networks, such as ResNet-50, ResNet-101, and ResNet-152, without performance degradation. ResNet achieved state-of-the-art results on the ImageNet

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

dataset, significantly improving classification accuracy (Elgendy, 2020).

ResNet introduced residual connections to address the vanishing gradient problem in deep networks, allowing for deeper and more accurate models. However, its complexity and depth can lead to increased computational costs and training time. Additionally, ResNet is known to be vulnerable to adversarial attacks and overfitting in small datasets if not properly regularized. Research by Li et al. (2022) highlights that Although ResNet reduces the issue of gradient vanishing, it is computationally intensive and may struggle to generalize to unconventional image distributions (Wang et al., 2022).

Hyperparameter tuning, particularly for learning rate, batch size, and the temperature value t is critical for enhancing the SupCon ResNet model's performance. In the referenced work, t was fixed at 0.07, while different combinations of learning rates and batch sizes were tested, revealing that a batch size of 512 with a learning rate of 0.05 achieved the lowest average false negative rate (FNR), outperforming larger batch sizes (Hoang et al., 2024). The Residual Neural Network (ResNet) architecture to address vanishing gradient issues in deep CNNs by using skip connections, enabling the training of very deep models like ResNet- 50 and ResNet-18 efficiently and with improved performance (Naaman, 2022). The ResNet-50 architecture, applied to vehicle brand detection as reported in the cited research, boosted classification accuracy from an initial 34 % to a final 95.31 %, highlighting its effectiveness in handling diverse vehicle scales and improving overall model performance (Mohite et al., 2025). The FastAI-driven ResNet-32 model demonstrated exceptional diagnostic performance in classifying ductal carcinoma tissue slides, achieving a classification confidence of 98.6 % and a high Diagnostic Odds Ratio (DOR) of 196.75 across multiple cross-validation folds (Praveen et al., 2022). Based on the reported results, the improved ResNet-18 model achieved the highest overall classification accuracy of 96.50 % on the MIT-BIH dataset, outperforming state-of- the-art models in ECG heartbeat classification, particularly by leveraging slicing techniques and compensating for class imbalance (Jing et al., 2021). Based on the results reported by author the iSPLInception model, built upon the Inception-ResNet architecture, outperformed CNN, CNN-LSTM, and LSTM variants across four UCI HAR datasets, demonstrating superior adaptability and scalability in human activity recognition despite dataset imbalances (Ronald et al., 2021). To improve ResNet-50 performance, Lin and Wu introduced adaptive learning rates and both L1/ L2 regularization to address overfitting and accuracy fluctuations. Additionally, by combining feature maps from intermediate layers (conv5 block1 out and conv5 block2 out), they enhanced the model's ability to detect critical patterns such as bleeding in diabetic retinopathy grading (Lin et al., 2023). The ResNet- 50 method achieves the highest overall accuracy (OA) due to its residual modules, which effectively solve the network degradation problem. The deep architecture, combined with convolutional layers, enhances feature extraction capabilities and training performance. ResNet-50's success is attributed to its ability to maintain robust performance even with deeper networks, setting it apart from other methods like BBO and CSO in this experiment.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Future work will explore different ResNet variants, such as Wide Residual Networks (WRN), ResNeXt, and MobileNet, to further investigate facial emotion recognition (Li et al., 2021). In the work cited by the authors, ResNet models, including ResNet-18 and ResNet-50, were applied to facial emotion recognition after preprocessing with CLAHE, using convolution, batch normalization, and residual blocks to effectively extract features and mitigate degradation in deep architectures (Sarwinda et al., 2021). Recent advancements in training procedures, as implemented in the timm library, have enhanced ResNet-50's performance, offering improved training methods and data augmentations that contribute to better results compared to the original model (Wightman et al., 2021). Figure 1 shows ResNet model summary in this study.

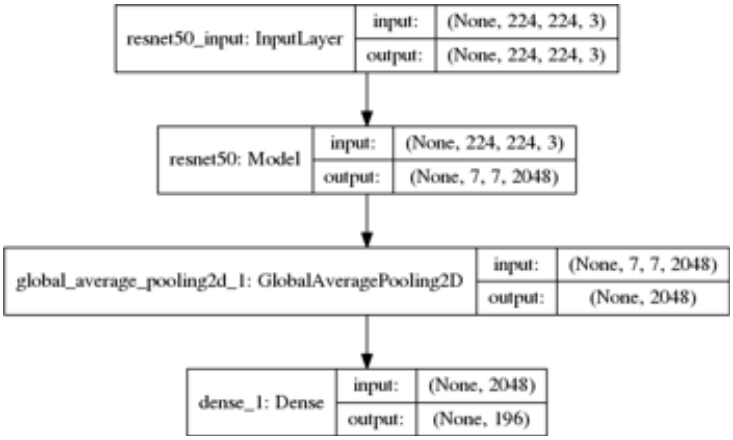


Fig. 1. Used pretrained ResNet model summary in this study.

VGG19

VGG19 is a deep convolutional neural network (CNN) developed by Simonyan and Zisserman as an extension of VGGNet. It consists of 19 weight layers, including 16 convolutional layers with 3 × 3 filters and three fully connected layers, followed by a softmax classifier. The network follows a uniform architecture, where small convolutional filters are stacked in multiple layers, increasing depth while maintaining a consistent structure. VGG19 improves upon AlexNet by using smaller 3 × 3 convolutional filters, which increase the network's capacity to capture complex features while reducing the number of parameters per layer. Despite its greater depth and over 144 million parameters, VGG19 converges faster due to implicit regularization from its deeper structure. It achieved a top-5 error rate of approximately 7.4 % on the ImageNet dataset, making it more accurate than its predecessor, AlexNet (Elgendy, 2020).

VGG-19 is widely used for benchmarking due to its deep architecture, which enhances classification accuracy. It employs dropout layers to mitigate overfitting and max-pooling for downsampling. Compared to sigmoid activation, ReLU activation is used for efficient training and gradient stability. The model has been successfully applied in medical image classification, achieving 96% accuracy and 82% sensitivity on the Kaggle fundus image dataset (Sudha et al., 2020).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Sudha and Ganeshbabu (2020) noted that VGG19 suffers from high overfitting risk, especially on small or imbalanced datasets, which is a common issue in VMMR when car brands are unevenly represented (Sudha et al., 2020). Additionally, the lack of residual connections makes it harder for gradients to flow, and training can become unstable in deeper networks compared to ResNet.

The VGG19 model demonstrated excellent performance with a test accuracy of 97.51 % and a test loss of 0.06244, indicating accurate predictions and high classification effectiveness across the dataset (Chowdhury et al., 2024). VGG19, a deep convolutional neural network architecture, was used in referenced work with spatial attention and class balance techniques to achieve an accuracy of 97.38% for recognizing benign malware. Despite its high performance, the model faces challenges due to dataset imbalance, where some malware classes have significantly fewer examples than others (Awan et al., 2021). Based on the findings of the cited study, VGG-19 with transfer learning achieved an outstanding accuracy of 99.72 % for classifying tomato leaf diseases, utilizing HSV color space for image segmentation to improve classification performance. The model's superior results, including high precision, sensitivity, and F1 score, were achieved by effectively segmenting leaf regions and eliminating background, in contrast to previous methods that did not apply such preprocessing (Nguyen et al., 2022). The VGG-19 network offers significant advantages in image recognition, including high verification accuracy and efficient training and verification times due to its use of consistent 3x3 convolution kernels and a deeper network structure (Yang et al., 2021). In the article, VGG-16 and VGG-19 both show impressive performance across varying epoch numbers. VGG-16 achieves high accuracy on both training and test data, with a peak test accuracy of 0.9656 at epoch 1, while VGG-19 shows slightly better results, reaching a test accuracy of 0.9744 at epoch 15 and maintaining high performance throughout (Mascarenhas et al., 2021). In the discussion, the VGG- 19 model is used as the base architecture for the proposed model, with only the classifier layer being updated. Despite overfitting during training, careful tuning of hyperparameters helped improve the model's performance, though the FER2013 dataset showed lower accuracy compared to the other two datasets (Meena et al., 2022). The VGG-19 model applied for cataract detection achieved a test accuracy of 97.47 % on the Kaggle dataset, demonstrating its effectiveness in detecting cataracts from unfiltered fundus photographs. Despite this success, further research into feature extraction techniques and classification algorithms is needed for more in-depth and precise detection, including severity grading (Khan et al., 2021). The use of transfer learning with the VGG19 network for strip steel defect recognition is effective. The hierarchical learning rate adjustment, improved feature extraction module, and the Adam optimizer contribute to significant improvements in the model's accuracy, robustness, and convergence speed. The model's real-time performance is also competitive, with a detection speed of 54 images per second (Wan et al., 2021). In these results, VGG19 is used as a feature extraction method in the proposed algorithm, offering a nonlinear transformation approach, unlike the linear transformations (DWT, DCT, and DFT). This distinction highlights VGG19's ability to capture more complex, deep features for

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

medical image watermarking (Han et al., 2021). The VGG-19 model, when pretrained, demonstrated exceptional performance with an accuracy of 97.3 %, recall of 99.2 %, and precision of 96.7 %, making it highly effective for classifying pneumonia from chest X-ray images. The review article shows the good quality of the model in the analysis of medical images (Ikechukwu et al., 2021). Figure 2 shows the pretrained VGG19 model summary.

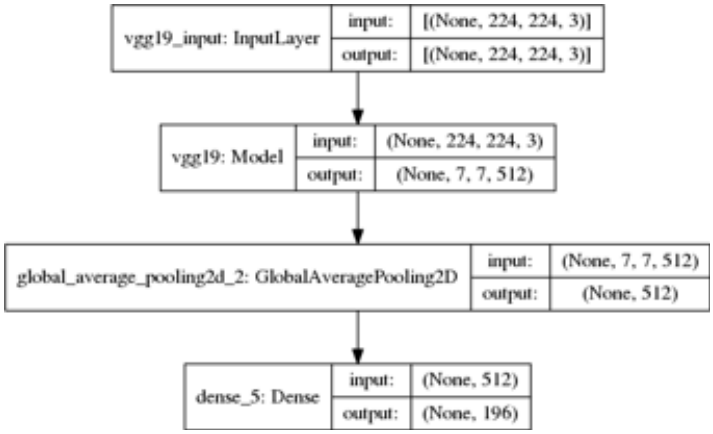


Fig. 2. Used pretrained VGG19 model summary in this study.

AlexNet is a deep neural network architecture developed in 2012 for image classification. It includes convolutional and pooling layers. The model uses activation, dropout, and overlapping pooling (ReLU) to reduce overfitting. AlexNet won the ImageNet competition (Li et al., 2021). AlexNet is a deep neural network that significantly improves computer vision, optimized for large-sized images. It was the winning model of the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and was developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. The architecture includes five convolutional layers with subsequent max-pooling layers, three fully connected layers, and a final softmax layer for classification. Unlike earlier models like LeNet, AlexNet incorporates Rectified Linear Unit (ReLU) activation functions to speed up training, dropout layers to reduce overfitting, and overlapping max pooling for improved feature extraction. Its depth and complexity, with approximately 60 million parameters and 650,000 neurons, enabled it to learn intricate patterns and significantly outperform previous approaches in image recognition tasks (Elgendy, 2020).

AlexNet, though revolutionary when introduced, is now considered relatively shallow and outdated for complex tasks like fine-grained vehicle classification. It has only eight layers and lacks the depth and modularity of newer architectures. AlexNet tends to underfit when applied to large scale, high resolution datasets and does not include advanced mechanisms for extracting fine-grained features (Elgendy, 2020).

Moreover, its use of large filter sizes in early layers and fewer total layers reduces its ability to capture fine details required for tasks like distinguishing between visually similar car models.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

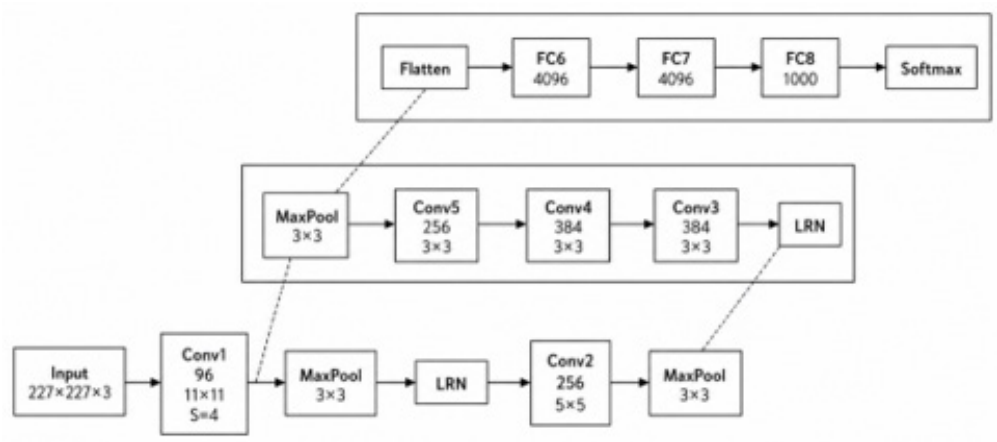


Fig. 3. Used reconstructed AlexNet model summary in this study.

AlexNet demonstrated effective clothing classification, achieving a 74 % accuracy on clean images and 68.28% accuracy on a combined dataset of clean and loss images, showcasing its ability to classify various clothing types using deep learning (Noh, 2021). AlexNet, a deep learning model, is highly efficient in facial expression recognition (FER) due to its computational speed, providing faster results in comparison to other models. It performs well in real-time applications, offering a balance between accuracy and time consumption, making it suitable for dynamic environments such as human- computer interaction and healthcare (Paulchamy et al., 2025). In the comparison of different deep learning models for 3D image segmentation, 3D AlexNet demonstrated superior performance, with higher accuracy and faster training times compared to models like ResNet50 and Inception-V4. Its optimized architecture, including reduced layers and channels, resulted in a classification performance with an AUC value of 0.964, outperforming the other models in terms of accuracy, sensitivity, specificity, and F1-score (Chen et al., 2021). For instance, the modified AlexNet architecture has shown to be a strong and efficient CNN model for the early identification of crop diseases (Chen et al., 2022). It attained an accuracy of 81.07 %, a precision of 82.70 %, and a recall of 80.24 %, indicating competitive effectiveness among similar models (Nugraha et al., 2023). AlexNet, when applied to the CEW dataset for eye state recognition, achieved an accuracy of 97.11% and an AUC of 99.42 %, demonstrating competitive performance. Although it did not outperform the proposed DCNN in terms of both accuracy and AUC, it remains a strong alternative for eye state detection in real-world applications (Kayadibi et al., 2022). Based on the results, the proposed 3D AlexNet achieved competitive performance in brain tumor classification, surpassing existing methods in precision for LGG, HGG, and healthy subjects, though slightly underperforming in overall accuracy compared to a deep CNN approach (Rani, et al., 2022). Based on the results, AlexNet demonstrated superior recognition accuracy (87.8 %) and faster detection performance compared to VGG16 and ResNet18, making it a suitable choice for real-time traffic signal recognition, especially when combined with YOLOv5s and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

low-light enhancement techniques (Niu et al., 2022). Based on the results, introducing deconvolution layers into AlexNet significantly reduces model parameters, especially in the fully connected layers, and improves test accuracy by 1–3%, making the enhanced AlexNet both more efficient and more accurate for small-scale image classification (Li et al., 2021). The improved AlexNet model, modified for tomato leaf disease recognition using transfer learning, achieved superior performance with 98.72 % accuracy by fine-tuning layers, adjusting the output to 8 classes, and employing weighted fusion of HOG and LBP features. Compared to traditional models like VGG and SVM based methods, it demonstrated faster convergence, higher recognition accuracy, and stronger generalization (Qiu et al., 2024).

Vehicle Make and Model Recognition (VMMR) refers to the task of identifying a vehicle's make (manufacturer) and model (specific product) from an image or video input. This process is essential for applications in Intelligent Transportation Systems (ITS), where it is used for tasks like traffic management, surveillance, behavior analysis, and monitoring. VMMR helps in automatic vehicle detection at toll stations, real-time traffic condition analysis, and locating stolen vehicles (Lyu et al., 2022). The proposed solution is deeply embedded in the Internet of Things (IoT) ecosystem, linking physical vehicle tags with cloud-based data systems for seamless identification and monitoring (Mufti et al., 2023). Driving style (DS) can be classified as either short term – based on situational behavior or long term – reflecting consistent driving patterns over time; classification methods include both continuous indices for precision and discrete categories for simplicity, influenced by driver, vehicle, and environmental factors (Mei et al., 2025). A hierarchical cloud platform architecture for intelligent vehicle systems, organized into four key layers application, network, platform, and terminal. The terminal layer handles data collection from vehicles and sensors. The network layer ensures communication between vehicles and the cloud. The platform layer is central to the system, offering powerful computing and storage for processing large-scale vehicle data. At the top, the application layer supports services such as driving style recognition, vehicle coordination, and speed planning–leveraging the data processed in the lower layers to deliver intelligent transportation solutions (Mei et al., 2023) (Figure 3).

This study concludes that the proposed TCN-ATM model significantly outperforms traditional models like LSTM, SVM, and TCN in recognizing lane change intentions, while multi-task learning approaches, especially MTL-LSTM, enhance prediction accuracy by leveraging correlations between driving indicators, offering promising applications for improving autonomous vehicle perception and control (Yuan et al., 2023). The reviewed literature provides a detailed overview of energy management strategies (EMSs) for hybrid electric vehicles and driving cycles (DCs) (Liu et al., 2020). This aspect affects road traffic in terms of electric vehicles.

The article (Li et al., 2023) reviews an integrated control strategy for mechanical-electrohydraulic hybrid vehicles (EHHVs) with real-time driving mode recognition. By using PCA for dimensionality reduction, K-means clustering for offline classification, and LVQ neural networks for online mode recognition, the proposed LVQ-fuzzy EMS

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

improves energy allocation and recovery, optimizing consumption without compromising dynamic performance. The study highlights the potential of EHHVs for efficient power and energy recovery, offering a foundation for future hybrid vehicle EMS and intelligent mode switching.

According to the analysis conducted in the cited study, we developed an algorithm for recognizing and correcting vehicle maneuvers in vibration-based road quality measurement systems. The algorithm demonstrated a 93 % classification accuracy and improved road surface classification by 7 % in suburban areas and 10 % in urban environments, where aggressive maneuvers are more common. By compensating for the impact of these maneuvers, the approach significantly enhanced the accuracy of road quality measurements, offering a scalable and low-cost solution that integrates seamlessly into existing systems (Nagy et al., 2025). This paper introduces an adaptive energy management strategy for hybrid electric tracked vehicles (HETVs), combining deep reinforcement learning (DRL) and transfer learning (TL) to address the challenges of long training times. The proposed bi-level control framework optimizes energy management by integrating DDPG and TL, with a focus on improving adaptability and performance across different driving cycles. Experimental results demonstrate the effectiveness of the approach, offering significant improvements in training efficiency and energy savings (Guo et al., 2020). This paper proposes a multi-sensor feature fusion classification method for vehicle micro-motor noise recognition, inspired by auditory perception principles. By transforming multi-sensor time-domain signals into Mel spectrograms, the method effectively extracts and optimizes features, achieving a high recognition accuracy of 94.90%. The approach also employs SpecAugment for data augmentation, increasing sample volume and improving model training. Experimental results demonstrate its effectiveness, especially in resisting interference in vehicular environments, with an accuracy of 94.27 % under low signal-to-noise ratio conditions (Zhao et all., 2023).

This paper discusses the development of Automatic License Plate Recognition (ALPR) systems, which are widely used in intelligent transportation systems for tasks such as traffic law enforcement, toll collection, and security monitoring. It highlights the transition from traditional methods to deep learning-based approaches, which offer superior accuracy and speed for both license plate detection and recognition. The paper also emphasizes the need for large, diverse datasets to train these models, and it introduces the challenge of character wise annotation for European license plates, which the study aims to address. Furthermore, the work describes a bimodal re-identification system that employs YOLOv5 for detecting vehicles and license plates (Ramajo-Ballester et al., 2024). This paper introduces a GPS trajectory-based machine learning method to better understand vehicle mobility patterns by capturing both travel actions and intentions. It uses DBSCAN clustering to identify high-frequency access locations and a graph-based PCA method to infer the intended destinations of vehicles. The study, based on data from 11,590 vehicles in Beijing, demonstrates that the method can reveal drivers' travel intentions beyond their actual parking locations. The findings show significant

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

differences between intended destinations and parking behaviors, with drivers heading home more likely to park near residential areas, while those heading to work are less likely to park near office buildings. The method also identifies commuters and non-commuters by analyzing frequent access points, revealing that commuters tend to travel shorter distances compared to non-commuters (Su et al., 2025). VMMR has recently attracted growing interest in AI and computer vision. However, issues like low lighting, bad weather, occlusion, and angle variation still hinder detection accuracy (Berwo et al., 2023).

A standard convolutional network, such as ResNet, can be progressively modernized toward the hierarchical architecture of vision Transformers like Swin, while excluding any attention-based modules. Model accuracies within the ResNet-50 / Swin-T computational regime are represented by the foreground bars, whereas results in the ResNet-200 / Swin-B regime are shown in gray. Hatched bars indicate architectural modifications that were not applied. Detailed quantitative results for both regimes are provided in the appendix.

Incorporation of Transformer-inspired architectural components, such as patchified convolutional stages, normalization refinements, and depth scaling, consistently improves performance. Through these cumulative enhancements, the resulting architecture, known as ConvNeXt, emerges as a purely convolutional model that matches or exceeds the accuracy of the Swin Transformer while maintaining comparable computational efficiency (Liu et al., 2022).

SimCLR, or Simple Framework for Contrastive Learning of Visual Representations, is a self-supervised learning method designed to learn image representations without using labeled data. It operates by maximizing the agreement between two differently augmented views of the same image while minimizing the similarity between views of different images. The approach relies on strong data augmentation techniques–such as random cropping, color distortion, and Gaussian blur–to create correlated image pairs used for training.

A convolutional neural network, typically a ResNet, serves as the base encoder to extract feature representations, which are then mapped to a latent space through a small projection head composed of a multilayer perceptron. In this latent space, the contrastive loss, known as the normalized temperature-scaled cross-entropy (NT-Xent) loss, is applied. This objective encourages high cosine similarity between positive pairs and low similarity between negatives drawn from the same mini-batch.

Through this contrastive framework, SimCLR effectively learns robust and transferable visual representations. It achieves 76.5 % top-1 accuracy on ImageNet under the linear evaluation protocol, surpassing previous self-supervised approaches and establishing a new state of the art in self- and semi-supervised learning (Chen et al., 2020).

CLIP is a language and image based pretraining model that represents and compares visual and textual embeddings. The model jointly encodes textual and visual embeddings on large data sets, optimizing the classification process using cosine

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

similarity between vectors.

This neural network differs from computer vision models in that it processes natural language, enabling it to generalize to a wide range of visual concepts without explicitly training it on a specific task. During inference, the text encoder can transform class names or descriptions into embeddings, forming a zero-shot linear classifier that allows the model to recognize new categories directly from text.

Through this contrastive training paradigm, CLIP acquires highly transferable and robust representations. It demonstrates strong performance across diverse tasks, including object recognition, OCR, action understanding, and geo-localization, while surpassing supervised ImageNet models in both efficiency and zero-shot generalization capability (Radford et al., 2021).

Two-step detectors (e.g., R-CNN, Fast R-CNN, Faster R-CNN, R-FCN) generally offer higher accuracy but are computationally heavier. One-step detectors (e.g., SSD, YOLO variants, RetinaNet) are faster and better suited for real-time use (Berwo et al., 2023). This paper introduces an enhanced activation function combining ReLU and sigmoid, which shows a significant improvement in recognition accuracy – 8.58 % higher than ReLU and 16.55 % better than LreLU, when tested on the University of Berkeley's road scene dataset. The model also outperforms sigmoid and Tanh functions on large road datasets, addressing the oscillation issues of LReLU and the inability of ReLU to handle the negative half-axis effectively. While the model still has room for improvement in certain categories, future work will focus on expanding the dataset and optimizing network parameters to boost accuracy. The paper highlights that deep learning models, closely linked to traditional computer vision systems, are improving rapidly in structure, design, and training, with significant advancements expected in the coming years (Li et al., 2023). This paper (Liu et al., 2024) introduces a view-independent vehicle model recognition framework using object detection and image retrieval techniques, designed for edge-cloud systems. Tested on the Stanford Cars dataset, the approach outperforms most previous methods in terms of accuracy and efficiency. The image screening and component detection processes reduce search time and calculation costs, making the model more efficient. Unlike traditional end-to-end methods, this two-stage approach avoids repeated training when the sample set changes, offering adaptability for large-scale and multi-category vehicle recognition. The results suggest strong potential for real-time applications, with future work focusing on improving feature descriptors and developing an adaptive vehicle model recognition system (Liu et al., 2024). ViTs have gained attention for their strong performance in image tasks, thanks to the MSA module that models global interactions. However, their generalization can be limited due to a lack of inductive bias for local relations. To overcome this, CNN-Transformers (HVTs) combine self-attention and convolution, enhancing performance by learning both local and global features. Various methods, like early and parallel integration, have been proposed to incorporate convolutional bias into transformers. Despite challenges, HVTs show great potential for real-world computer vision applications (Khan et al., 2023).

Convolutional Neural Networks (CNNs) have become a popular choice for

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Acoustic Target Recognition (ATR) in acoustic images, offering superior performance over traditional template-matching techniques. Unlike template matching, which struggles with generalizing patterns and handling multi-scale objects, CNNs are capable of learning relevant features directly from the data. The hierarchical structure of the CNN allows for the extraction of simple edge features and complex patterns in deep layers. This is suitable for ATR tasks such as object detection and classification in acoustic images. In the paper, the models demonstrated their effectiveness even with computational limitations and in an underwater environment (Bresciani et al., 2024). The reviewed study discusses the issues of hyperparameter tuning for deep reinforcement learning (DRL), in particular the impact of network capacity on performance (Palomeras et al., 2024).

This study focuses on the detection and recognition of vehicles and their license plates using deep learning methods, specifically leveraging pre-trained YOLO models for vehicle detection. The task involves identifying regions of interest (ROIs) within images that correspond to vehicles and license plates. The work also highlights the importance of large, diverse datasets for training vehicle recognition models, with several datasets mentioned, such as Stanford-Cars, VehicleID, and VeRi-776, which contain extensive image collections with varying levels of camera coverage and annotations. These datasets are critical for improving the performance of vehicle re-identification systems, ensuring the models can handle different vehicle types, camera angles, and environmental conditions (Ramajo-Ballester et al., 2024). The study under review examines the use of long-short-term memory (LSTM) models for recognizing vehicle engine dynamics. Instead of k-means methods, the paper uses an autoencoder for data dimensionality reduction and spectral clustering. Nonlinear dimensionality reduction improves the model's robustness to noisy data (Chen et al., 2024). The TSDR-KNN model was created for character recognition in clear images. It processes inputs through a convolutional layer to extract relevant features and applies a max-pooling operation to enhance its efficiency. On the other hand, HRU-Net focuses on image recovery by combining a compact CNN with a self-adaptive image processing (SAIP) module for enhanced image clarity post-defogging. This approach incorporates filters for white balance, contrast, gamma correction, and sharpening, optimizing the image by reducing chromatic aberration and restoring lost details in darker regions (Rani et al., 2024). This study compares five driving condition recognition methods, highlighting the navigation-based method's high accuracy in most driving cycles, though it struggles with blended conditions like WLTP. Clustering showed the best fuel economy improvement, while fuzzy logic had the lowest. The study recommends using the navigation based method in the vehicle concept phase and clustering in production. Future research may focus on hardware testing and the impact of autonomous driving technologies (Xu et al., 2023). The proposed representative sample selection method significantly improves recognition rates compared to random and regular methods, offering a foundation for targeting and obtaining samples that optimize recognition algorithm performance in limited conditions (Zhang et al., 2024). This article explores different motion segmentation

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and appearance-based methods used for vehicle recognition. The text emphasizes the importance of these techniques in applications like traffic surveillance, urban planning, and security (Kanagamalliga et al., 2024).

MobileNet is built on depthwise separable convolutions, a form of factorized convolution that divides a standard convolution into two stages: a depthwise convolution and a pointwise convolution. In this structure, the depthwise convolution applies a single filter to each input channel independently, while the pointwise convolution, implemented as a 1×1 convolution, combines the resulting feature maps across channels. This separation of spatial filtering and channel mixing drastically reduces computational cost and model size compared to conventional convolutions (Khasoggi et al., 2019). MobileNets are lightweight convolutional neural networks designed to minimize computational latency while maintaining accuracy. They achieve this efficiency through depthwise separable convolutions, which separate spatial and channel operations to reduce parameters and operations. This architecture enables fast inference on resource-limited devices such as mobile and embedded systems (Howard et al., 2017). Thus, it is a more advanced neural network that optimizes and creates new features to classify using adaptive depthwise and pointwise convolutions. The mobile network structure used in the study is also shown in Figure 4.

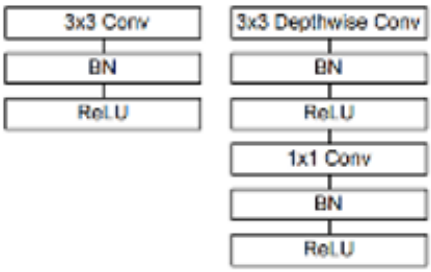


Fig. 4. "Left: Standard convolutional layer with batchnorm and ReLU. Right: Depthwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU." (Howard et al., 2017; p. 4).

EfficientNet is a convolutional neural network architecture designed to optimize both accuracy and computational efficiency through compound scaling of network depth, width, and resolution. It achieves high performance with fewer parameters by uniformly scaling these dimensions using a simple, principled method (Pu et al., 2024). EfficientNet is a family of convolutional neural networks developed through neural architecture search and compound scaling, which jointly optimizes network depth, width, and resolution. This design achieves a balance between accuracy and efficiency, outperforming previous convolutional models in both performance and computational cost. EfficientNet architectures also generalize well across diverse datasets, demonstrating strong transfer learning capabilities with compact model sizes (Tan et al., 2020). EfficientNet scales, allocates resources, and adapts to feature classification using 5 scaling methods.

The Vision Transformer (ViT) is a deep learning architecture that applies the Transformer model to image analysis by dividing images into patches and processing

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

them as sequences using self-attention mechanisms. This approach enables ViT to effectively capture spatial relationships within images without relying on traditional convolutional operations (Zhang et al., 2023). The ViT-B/16 model demonstrates high effectiveness in automated medical image analysis, achieving near-perfect accuracy in identifying elbow fractures from X-ray images, thereby significantly improving the reliability of diagnostic systems (Alam et al., 2024). The Vision Transformer (ViT) uses only the encoder of the Transformer architecture. It divides an image into fixed-size patches, projects each patch into a vector, and processes them as a sequence. A special classification token is added to the sequence, and its output represents the final class prediction (Huo et al., 2023). Visual transformers are closely related to the theory of text transformers and encoders, they use these methods and divide the main image into a sequence of embeddings.

Table 1. Comparison of Strengths and Weaknesses of CNN Models

Model	Strengths	Weaknesses
ResNet50	Mitigates vanishing gradients; enables deep training; fast convergence; good on small datasets; customizable for tasks	High computational cost; struggles with unusual distributions; overfitting without regularization; adversarial vulnerability; redundant features in deep layers
VGG19	Simple, uniform architecture; strong feature extraction with small filters; high accuracy; stable training	Very large parameter count; overfits small datasets; lacks skip connections; inefficient gradient flow
AlexNet	Pioneering deep learning model; includes dropout, ReLU, and overlapping pooling; efficient early training	Shallow depth; underfits high-res data; lacks fine feature extraction; not suited for fine-grained recognition
MobileNet	Uses fewer resources, more chances for feature recognition due to architecture.	Performance drops on large-scale or complex datasets due to limited representational capacity.
	Focus on scaling to achieve the desired accuracy and classification features	Compound scaling is computationally expensive to design and retrain for new domains.
ViT	Improved efficiency in object detection. An approach different from the CNN in defining objects	Requires large-scale datasets for effective training due to the absence of inductive biases like locality and translation invariance.
ConvNeXt	Increased convolution layers to 7x7 for better accuracy, increased kernel size. Fewer activation functions. Replacing ReLU with GELU. Fewer normalization layers	High attention cost. Needs more resources to input.
SimCLP	A simple and effective self-supervised framework that learns powerful visual representations without labels.	It achieves high accuracy but requires large batch sizes and extensive computational resources.
CLIP	A multimodal model that aligns images and text in a shared embedding space, enabling strong zero-shot generalization.	Its broad versatility comes at the cost of reliance on massive pretraining data and potential bias from web-scale text sources. Lack of text embeddings for models.

This work is a continuation of the development of the VMMR application for analyzing vehicle traffic.

Literature analysis of existing systems. Finding the values of car models that affect city traffic, searching for these values of certain brands in official vehicle documentation.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Training existing models to classify car brands and detect cars, determine their type.

In this article, the models described above were trained under identical conditions and analyzed using metrics, LIME, SHAP, and GradCam analyzers. The output data will help compare the performance of the models in identifying a car model from an image.

An existing dataset of car models was used for classification training, which contains 291,752 images of cars of the brand from 1950 to 2016 (Tafazzoli et al., 2017). This dataset is publicly available. Most vehicle classes refer to the year 2015 when the image was made. Most of the samples in the dataset are rendered in 3D free (complex angles). After filtering and dividing the dataset into the training and testing ones, 283,899 images and 7,982 classes of car models (1,188 skipped classes) remained. A cloud resource with an NVIDIA RTX 4000 Ada Generation GPU was used. The images were transformed to a size of 256x256 and randomly resized and cropped to 224x224. Some images were horizontally mirrored. Color Jitter was used to improve robustness, converting some images to grayscale p=0.03. Standardizes pixel values using ImageNet normalization parameters mean = [0.485,0.456,0.406], std = [0.229,0.224,0.225]. Preprocessing also used horizontal flipping and zooming methods. The models used the Adam optimizer. All models were trained under the same enviroment and conditions: epochs = 10, batch size = 32.

Jupiter notebook, saved models and other resources of this study are available in cloud storage (Mordvintsev, 2025).

Results and discussion.

Identifying the factors related to car models that influence city traffic and investigating these factors for specific brands through official vehicle documentation.

The size of a vehicle affects its maneuverability and the space it occupies on the road. Large vehicles require more room to maneuver, which reduces traffic speed and increases congestion. Turning radius also plays a significant role, affecting maneuver times at intersections.

The analyzed modern approaches to data processing, and the focus on vehicle identification and subsequent interpretation of characteristics.

Re-identification is the process of matching the appearance of objects within and between camera networks. Typically, this involves a user searching for all occurrences of a queried object across the cameras in a network. For example, a traffic officer may be trying to locate a specific car in the city camera system. Using of a deep CNN learning model to identify same car model images (Elgendy, 2020).

The classification of different machine learning approaches is based on the varying angles of the datasets utilized in these studies. The main objective of this analysis is to provide an overview of the current achievements of the VMMR and to highlight potential areas for development. This classification includes frontal, rear, multiview and miscellaneous perspectives (Elgendy, 2020). Current research working with frontal and back side car views.

A deep CNN ResNet50 was used as the backbone network to extract object-level features. In another approach, a hybrid approach combining the deep VGG16 model

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

with a Genetic Algorithm (GA) was proposed. In addition, a new loss function was introduced to improve the performance of deep CNN models (Elgendy, 2020).

There are three main learning techniques for training a model: supervised, unsupervised, and reinforcement learning. Hyperparameter tuning is also crucial to improve the performance of a model. Transfer learning uses pre-trained models to improve task performance. The model adaptation process uses model coefficients and trains on a large dataset specific to the task.

Data transformation methods transform images to focus on identifying patterns and features. This implementation included horizontal flipping, random rotation, and image normalization.

Accuracy and loss metrics of models.

The evaluation metrics of the trained models are presented in Figures 5–13.

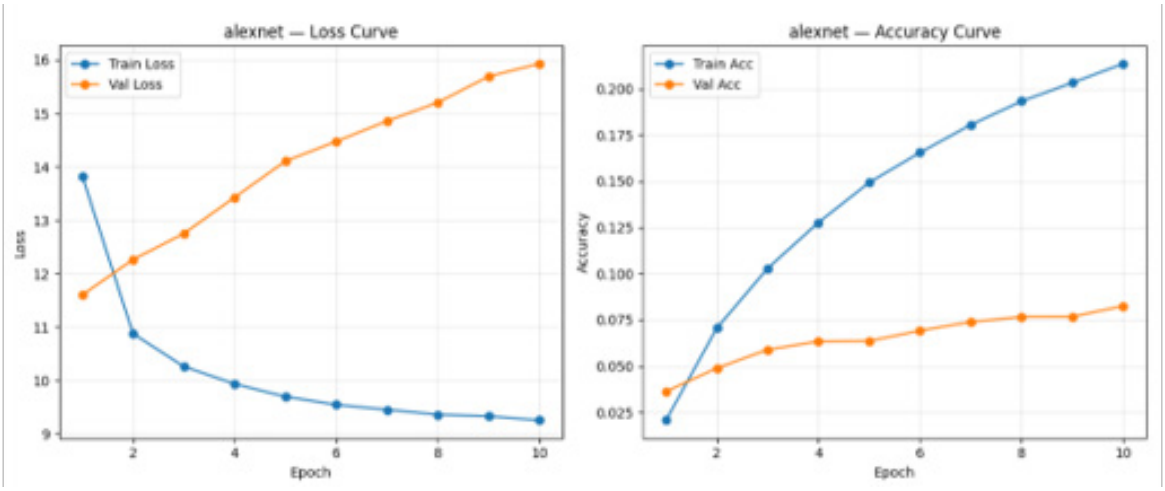


Fig. 5. AlexNet accuracy/loss.

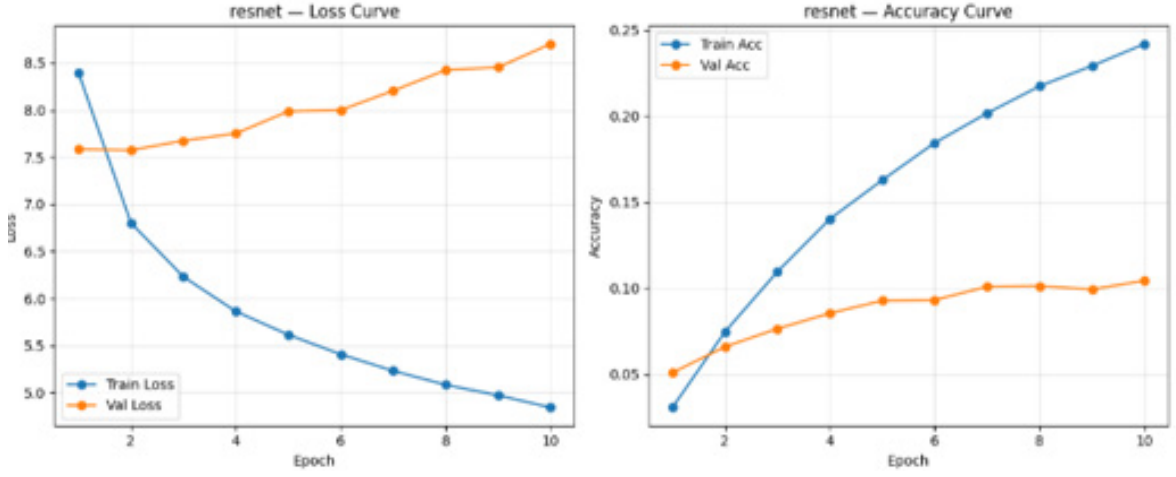


Fig. 6. ResNet accuracy/loss.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

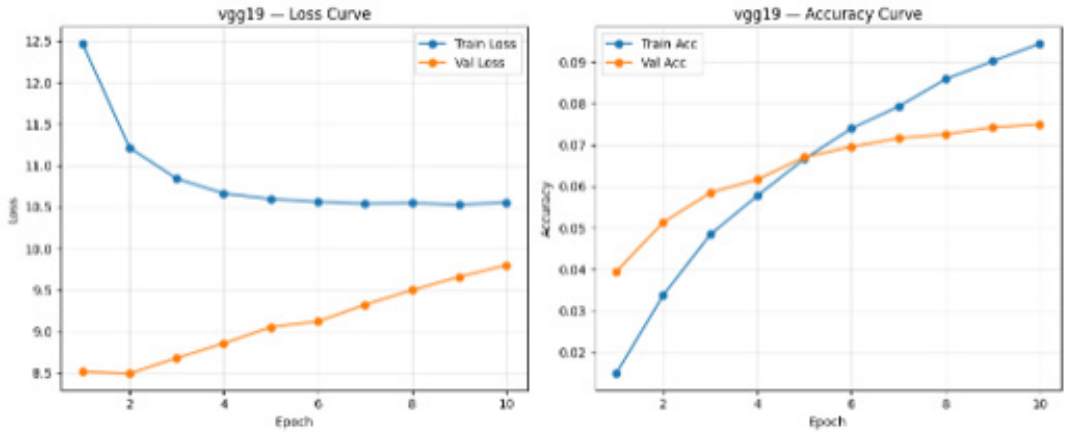


Fig. 7. VGG19 accuracy/loss.

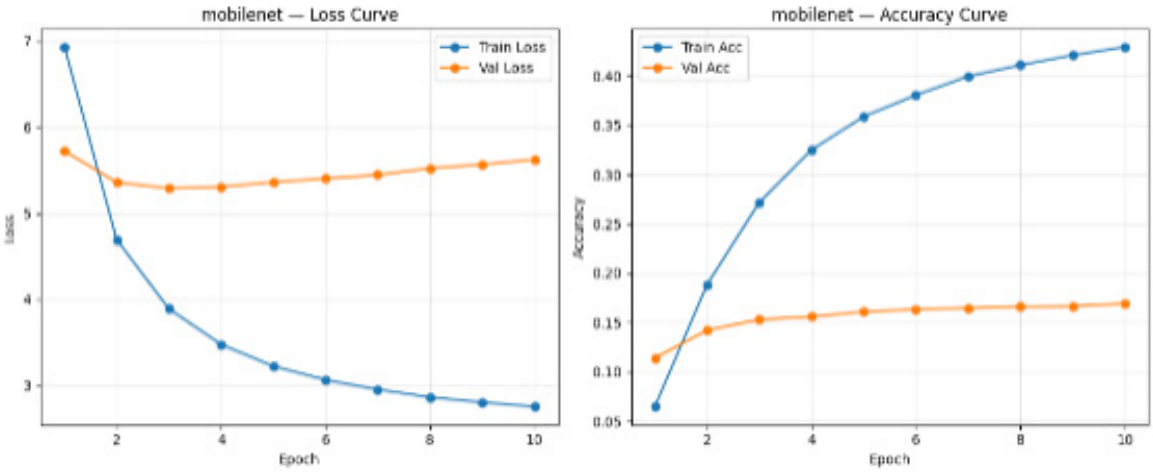


Fig. 8. MobileNet accuracy/loss.

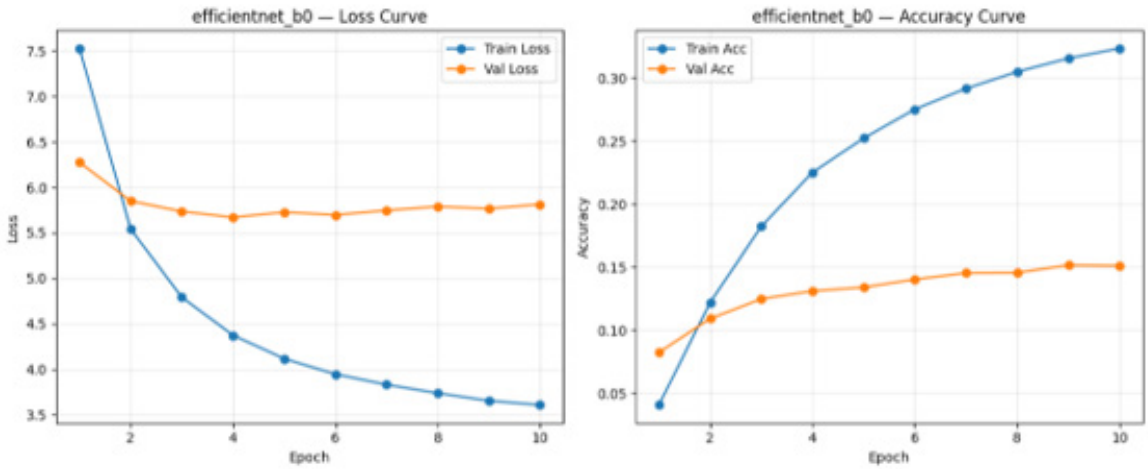


Fig. 9. EfficientNet accuracy/loss.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

For the current training, the EfficientNet layers were scaled in depth and finally in breadth for classification.

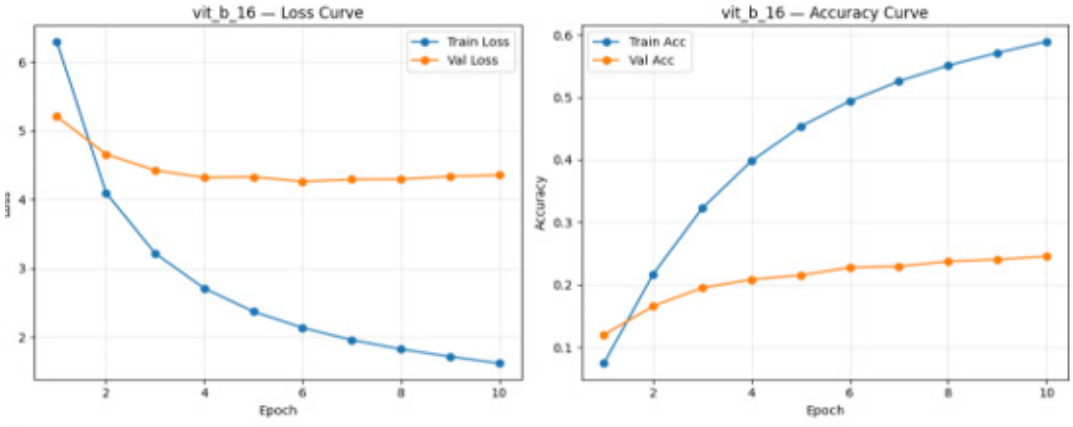


Fig. 10. ViT b16 accuracy/loss.

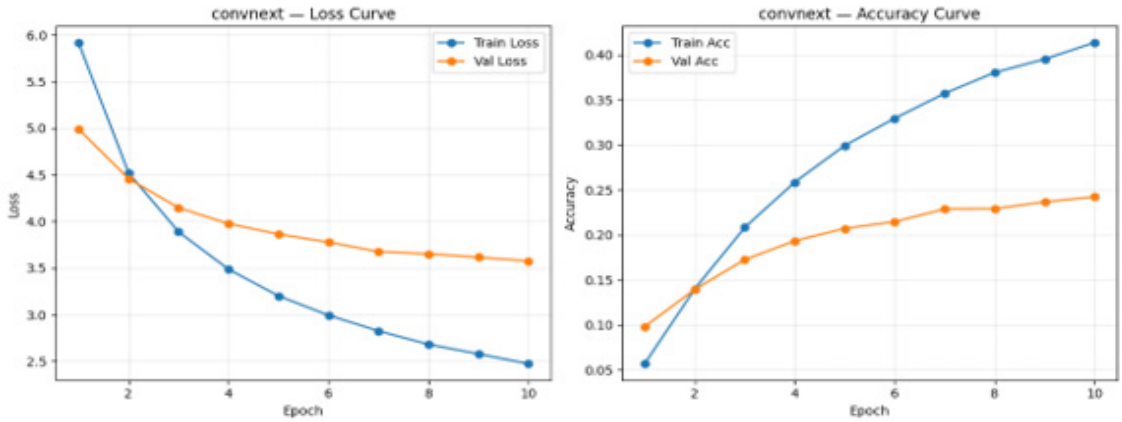


Fig. 11. ConvNeXT accuracy/loss.

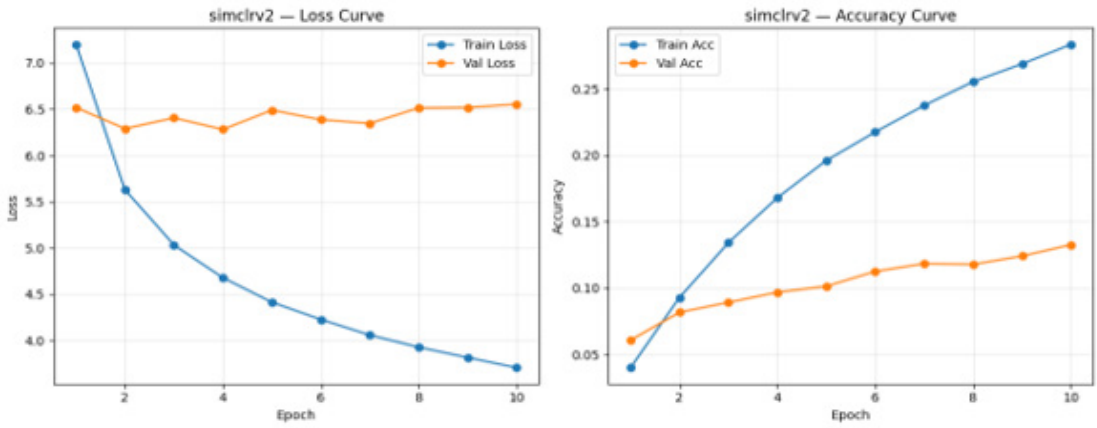


Fig. 12. SimCLR accuracy/loss.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

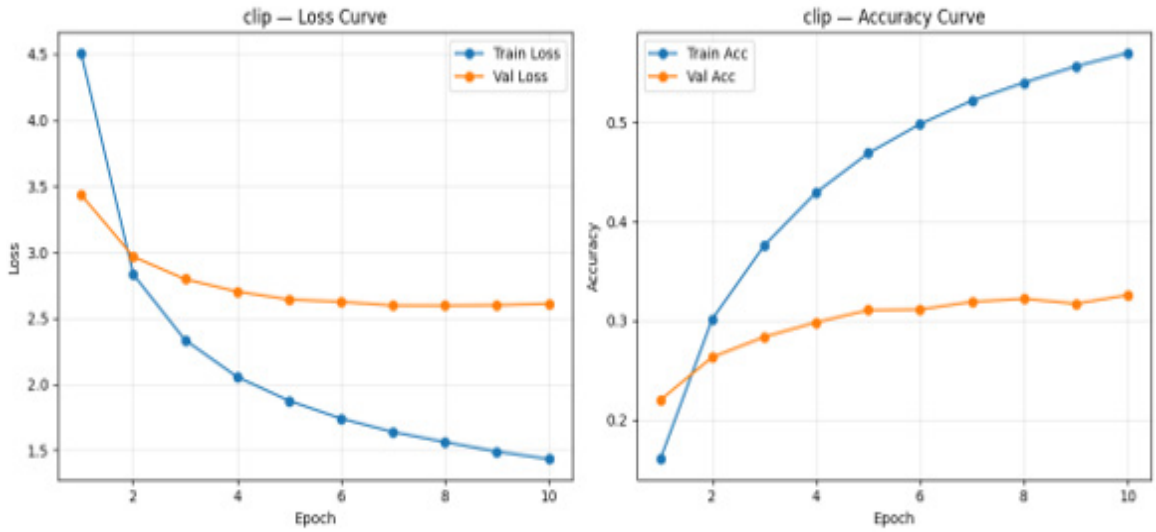


Fig. 13. CLIP accuracy/loss.

The comparison of model accuracy and loss is presented in Table 2.

Table 2. Comparison of Accuracy and Loss of models from Figures 5-13.

Model	Final train loss	Final validation loss	Final train accuracy	Final validation accuracy
AlexNet	9.2563	15.9313	21.36 %	8.26 %
ResNet	4.8434	8.7011	24.19 %	10.44 %
VGG19	10.5568	9.8006	9.44 %	7.49 %
MobileNet	2.7533	5.6254	42.95 %	16.91 %
EfficientNet	3.6078	5.8118	32.36 %	15.09 %
ViT b16	1.6194	4.3545	58.95 %	24.55 %
ConvNeXt	2.4711	3.5753	41.34 %	24.21 %
SimCLR	3.7059	6.5539	28.35 %	13.27 %
CLIP	1.4322	2.6097	56.98 %	%

As can be seen in Table 2, scalable networks have an advantage in classifying machine models, especially ConvNeXt, ViT b16 and CLIP.

Locally Interpretable Model-Agnostic Explanations analysis

From the visualization of 32 main features, it can be seen that the EfficientNet contains more accurate 2 negative features, while the negative features of the others are all over the picture Table 3. From the properties of convolutional networks, it can be seen that all models distinguish between the boundaries of the machine and the environment. All models except AlexNet mark the components of the car. As you can see, trained CLIP highlights complex areas grouped together and has greater semantic analysis.

Gradient weighted Class activation mapping analysis.

Table 3 presents Locally Interpretable Model-Agnostic Explanations (LIME) analysis of models.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Table 3. LIME analysis of models.

Model	Correct Top-k First	Top-1 accuracy	Positive features on car shape	Negative features on car shape	Mean absolute difference
AlexNet	No	0.996	11	1	0.7966
ResNet	No	0.168	12	1	0.0580
VGG19	No	0.080	10	1	0.0096
MobileNet	Yes	0.857	14	1	0.6702
EfficientNet	Yes	0.567	12	2	0.4127
ViT b16	Yes	0.331	11	1	0.1880
ConvNeXt	No	0.074	9	1	0.0159
SimCLR	Yes	0.663	11	1	0.4939
CLIP	Yes	0.990	5 (2 complex)	1	0.7908

Figures 14–22 show the results of gradient weighted class activation mapping.



Fig. 14. AlexNet GradCam.



Fig. 15. ResNet GradCam.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Fig. 16. VGG19 GradCam.



Fig. 17. MobileNet GradCam.



Fig. 18. EfficientNet GradCam.



Fig. 19. ViT b16 GradCam.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Fig. 20. ConvNeXt GradCam.



Fig. 21. SimCLR GradCam.

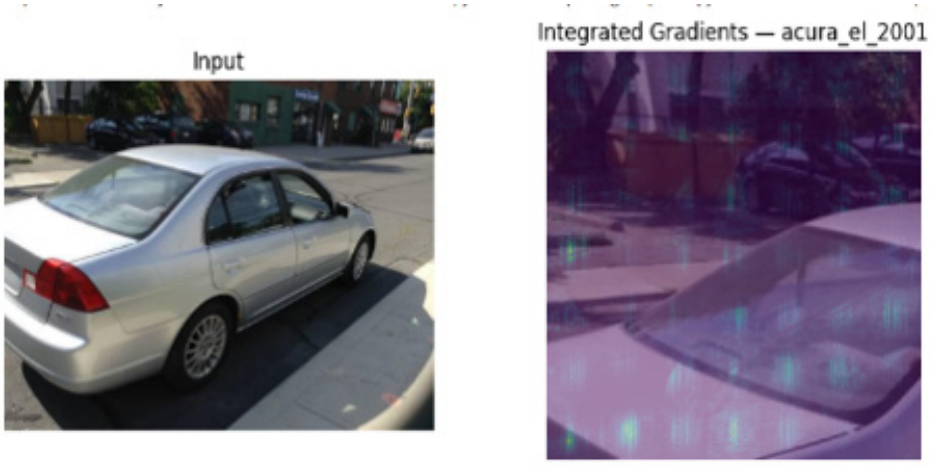


Fig. 22. CLIP GradCam.

The scaled models have a larger gradient on the machine, while in Figures 14-19 they are noisier. Also the gradient being divided into squares in the visual transformer

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and CLIP Figures 19 and 22. SHapley Additive exPlanations analysis

Table 4 presents SHAP analysis of models.

Table 4. SHAP analysis of models

Model	Correct Top-k First	Top-1 accuracy	Mean absolute difference
AlexNet	No	0.2	0.048
ResNet	No	0.2	0.048
VGG19	No	0.1	0.000
MobileNet	Yes	0.9	0.716
EfficientNet	Yes	0.6	0.422
ViT b16	Yes	0.4	0.238
ConvNeXt	No	0.1	0.022
SimCLR	Yes	0.65	0.476
CLIP	Yes	0.99	0.791

Although the visual transformer has greater accuracy, MobileNet, SimCLR and EfficientNet shows a contrast in classification (average absolute difference between top k) Table 4. Also, the MobileNet does not show attribution in the center of the hit map, unlike other models.

Discussion.

This work opens up new possibilities for the implementation of road analysis in a smart city. The semantic capabilities of current deep learning networks for determining complex information about a car brand model were considered. New possibilities for dynamic analysis of cars on the road open up a new look at formulas for calculating traffic lights and other road regulations. This research also observes unpredictable different behavior of similar deep learning models on the same dataset. The results show that scalable networks have potential for semantic recognition, and it is also noted that image tokenization also shows high performance in identifying different models of similar objects (cars).

The validation accuracy is below 30 %, which may indicate accuracy on new images in production mode. The evaluation was based on static test sets; model robustness under varying conditions (motion blur, night lighting, etc.) remains untested. Production mode of training requires multi-view dataset for recognition. Some highlighted LIME or SHAP regions may not correspond to true discriminative car features (e.g., background influence in first 32 features).

Creating simulations of rain, night, blur, and seasons using images from the dataset. Supplementing the dataset with other points of view on the machine. Increasing learning epochs.

Conclusion.

In this article, certain factors were found that are related to both the car model and the user experience of the road and the infrastructure of smart city roads. It was also found that not all factors are recorded in the documentation of the models with a direct value, for example, the size of the rotation of the car. Various modern approaches to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

classifying a car model by a two-dimensional image were analyzed. Scalable and non-scalable models were compared.

This study reviewed current data processing and analysis methods for vehicle detection and classification based on deep learning architectures. ResNet50, VGG19, AlexNet, MobileNet, EfficientNet, Vision Transformer (ViT-B16), ConvNeXt, SimCLR, and CLIP models have been trained and tested on a massive dataset comprised of 283,899 vehicle images.

The experimental results showed that all the models succeeded in learning a rough appearance-based model of vehicles with a validation accuracy range of 0.07-0.32, revealing that there remain significant issues regarding dataset complexity, inter-class similarity, and possible imbalance of the data distribution. Classical convolutional architectures like VGG19 and ResNet50 were average performers, while more recent architectures like ConvNeXt and ViT-B16 have superior feature extraction capabilities, but they are increasingly demanding in terms of fine-tuning and computational power. Self-supervised and contrastive methods (SimCLR and CLIP) emphasized that transfer learning from large pre-trained representations is promising, but they are no match for domain-specific adaptations.

In conclusion, the study confirms that deep learning is still a promising approach for vehicle recognition but it is still far from trivial to obtain high accuracy with them without performing proper data pre-processing, class balancing, and model tuning. The results highlight the relevance of domain adaptation, augmentation techniques, and architecture dependent tuning for improving generalization. Future works may investigate hybrid or ensemble methods, transformer CNN based architectures, contrastive pretraining on domain related vehicle datasets for further enhancements in the reliability of detection and identification for practical real-life ITSs.

REFERENCES

Awan M.J., Masood O.A., Mohammed M.A., Yasin A., Zain A.M., Damasˇevicˇius R. and Abdulkareem K.H. (2021). Image-based malware classification using VGG19 network and spatial convolutional attention // Electronics. — Vol. 10. — No. 19.

Alam, Taukir & Yeh, Wei-Cheng & Hsu, Fang & Shia, Wei-Chung & Robert Singh, A. & Hassan, Taimoor & Lin, Wenru & Yang, Hong-Ye & Hussain, Tahir (2024). An Integrated Approach using YOLOv8 and ResNet, SeResNet & Vision Transformer (ViT) Algorithms based on ROI Fracture Prediction in X-ray Images of the Elbow. Current medical imaging. 20. 10.2174/0115734056309890240912054616.

Berwo M.A., Khan A., Fang Y., Fahim H., Javaid S., Mahmood J., Abideen Z.U. and MS S. (2023). Deep learning techniques for vehicle detection and classification from images/videos: A survey // Sensors. — Vol. 23. — No. 10. Pp. 4832.

Bresciani M., Zacchini L., Topini A., Ridolfi A., and Costanzi R. (2024). Automatic target recognition and geolocalisation of natural gas seeps using an autonomous underwater vehicle. Control Engineering Practice. — Vol. 145. — P. 105864.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning. Pp. 1597-1607. PmLR.

Chen J., Wan Z., Zhang J., Li W., Chen Y., Li Y., and Duan Y. (2021). Medical image segmentation and reconstruction of prostate tumor based on 3d alexnet // Computer methods and programs in biomedicine. — Vol. 200. — P. 105878.

Chen H.-C., Widodo A.M., Wisnujati A., Rahaman M., Lin J.C.-W., Chen L., and Weng C.-E. (2022). Alexnet convolutional neural network for disease detection and classification of tomato leaf // Electronics. — Vol. 11. — No. 6. P. 951.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Chowdhury M.T., Rahman H., Sumon M.I., and Talha A. (2024). Classification of satellite images with vgg19 and convolutional neural network (CNN), in 2024 2nd International Conference on Advancement in Computation Computer Technologies (InCACCT). Pp. 397–402.

Chen Y., Yang C., Wang W., and Gao P. (2024). Energy management for hybrid electric vehicles using an intelligent recognition method of engine characteristics. IFAC-PapersOnLine. — Vol. 58. — No. 29. Pp. 379–384.

Hoang T.N. and Kim D. (2024). Supervised contrastive resnet and transfer learning for the in-vehicle intrusion detection system. Expert Systems with Applications. — Vol. 238. — Article 122181.

Hongwen Pu. & Keyi Yi. (2024). A Comparative Analysis of EfficientNet and MobileNet Models' Performance on Limited Datasets: An Example of American Sign Language Alphabet Detection. Highlights in Science, Engineering and Technology. No. 94. Pp. 558–564. 10.54097/yh5d3s04.

Han B., Du J., Jia Y., and Zhu H. (2021). Zero-watermarking algorithm for medical image based on VGG19 deep convolution neural network // Journal of Healthcare Engineering. — Vol. 2021. — No. 1. P. 5551520.

Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., and Adam H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huo, Yingzi & Jin, Kai & Cai, Jiahong & Xiong, Huixuan & Pang, Jiacheng. (2023). Vision Transformer (ViT)-based Applications in Image Classification. Pp. 135–140. 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033.

Ikechukwu A.V., Murali S., Deepu R., and Shivamurthy R. (2021). Resnet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images // Global Transitions Proceedings. — Vol. 2. — No. 2. Pp. 375–381.

Girhe N., Patel F., Kale P., Mahure A., and Bhute A. (2022). "Smart traffic system to optimize city traffic. YMER 21.12. Pp. 972–994.

Jabeen S., Jabeen A.J., Adnan S.M., and Rao W.A. (2021). Vehicle Make and Model Recognition Using Structural and Pattern Based Feature Descriptors, 2021 // International Conference on Communication Technologies (ComTech). Pp. 73–78. DOI:10.1109/ComTech52583.2021.9616857.

Jing E., Zhang H., Li Z., Liu Y., Ji Z., and Ganchev I. (2021). ECG heartbeat classification based on an improved ResNet-18 model // Computational and Mathematical Methods in Medicine 2021. Article 6649970.

Khan A., Rauf Z., Sohail A., Khan A.R., Asif H., Asif A. and Farooq U. (2023). A survey of the vision transformers and their CNN-transformer based variants. Artificial Intelligence Review. No. 56. Pp. 2917–2970.

Khan, M.S.M. Ahmed M., Rasel R.Z., and Khan M.M. (2021). Cataract detection using convolutional neural network with VGG-19 model, in 2021 IEEE World AI IoT Congress (AIIoT). Pp. 0209–0212. IEEE.

Kayadibi I., Guraksın G.E., Ergun U., and Ozmen Suzme N. (2022). An eye state recognition system using transfer learning: Alexnet-based deep convolutional neural network // International Journal of Computational Intelligence Systems. — Vol. 15. — No. 1. P. 49.

Kanagamalliga S., Kovalan P., Kiran K., and Rajalingam S. (2024). Traffic management through cutting-edge vehicle detection, recognition, and tracking innovations // Procedia Computer Science. — Vol. 233. — Pp. 793–800.

Khasoggi, B. & Ermatita, E. & Samsuryadi, S. (2019). Efficient mobilenet architecture as image recognition on mobile and embedded devices // Indonesian Journal of Electrical Engineering and Computer Science. — Vol. 16. — No. 389. 10.11591/ijeecs.v16.i1. Pp. 389-394.

Li S., Wang L., Li J., and Yao Y. (2021). Image classification algorithm based on improved AlexNet // in Journal of Physics: Conference Series. — Vol. 1813. P. 012051. IOP Publishing.

Lyu Y., Schiopu I., Cornelis B., and Munteanu A. (2022). Framework for vehicle make and model recognition–a new large-scale dataset and an efficient two-branch–two-stage deep learning architecture // Sensors. — Vol. 22. — P. 8439.

Liu Z., Mao H., Wu C.Y., Feichtenhofer C., Darrell T. & Xie S. (2022). A convnet for the 2020s., In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Pp. 11976–-11986.

Li B. and Lima D. (2021). Facial expression recognition via resnet-50 // International Journal of Cognitive Computing in Engineering. — Vol. 2. Pp. 57–64.

Liu C., Pu Z., Li Y., Jiang Y., Wang Y., and Du Y. (2024). Enabling edge computing ability in view-independent vehicle model recognition // International Journal of Transportation Science and Technology. No. 14. Pp. 73–86.

Li S., Wang L., Li J., and Yao Y. (2021). Image classification algorithm based on improved AlexNet // Journal of Physics: Conference Series. — Vol. 1813. P. 012051.

Li L., Zhang T., Lu L., Zhang H., Yang J., and Zhang Z. (2023). An energy active regulation management strategy based on driving mode recognition for electro-hydraulic hybrid vehicles. Energy. — Vol. 285. — P. 129548.

Lin C.-L. and Wu K.-C. (2023). Development of revised Resnet-50 for diabetic retinopathy detection. BMC

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

bioinformatics 24.1. P. 157.

Mascarenhas S. and Agarwal M. (2021). A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification, in 2021 International conference on disruptive technologies for multi- disciplinary research and applications (CENTCON). — Vol. 1. Pp. 96–99. IEEE.

Mohite, Sagar & Jadhav, Swati & Mohite, Snehal & Jagadale-Patil, Devyanee- Netra & Gayakwad, Milind & Singh, Kartikey & Joshi, Rahul. (2025). Vehicle identification using ResNet-50: CNN approach, in 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS). Pp. 1–6. IEEE.

Meena G., Mohbey K.K., Indian A., and Kumar S. (2022). Sentiment analysis from images using vgg19 based transfer learning approach // Procedia Computer Science. — Vol. 204. — Pp. 411–418.

Mufti N., Ullah S., Sharif A., Nawaz M.W., Alkhayyat A., Imran M.A., Abbasi Q.H. (2023). IoT-enabled vehicle recognition system using inkjet-printed windshield tag and 5 G cloud network // Internet of Things. — Vol. 23. — Pp. 100873.

Mordvintsev, N. (2025). Jupyter notebooks and resources [Data set]. Astana IT University. https://drive.google.com/drive/folders/1qmm7Pnmpgxoab7s3fpog4b-3zaawzvzQ?usp=sharing.

Noh S.K. (2021). "Recycled clothing classification system using intelligent IoT and deep learning with AlexNet // Computational Intelligence and Neuroscience. — Vol. 2021. No. 1. P. 5544784.

Nagy R. and Szalai I. (2025). Vehicle maneuver recognition and correction algorithm for road quality measurement system optimization. Measurement: Sensors. Article. 101816.

Naaman O. (2022). Resnet and LSTM-based accurate approach for license plate detection and recognition. Traitement du Signal. — Vol. 39. — No. 5. Pp. 1577–1583.

Nguyen T.H., Nguyen T.N., and Ngo B.V. (2022). A VGG-19 model with transfer learning and image segmentation for classification of tomato leaf disease. AgriEngineering. — Vol. 4. — No. 4. Pp. 871–887.

Niu C. and Li K. (2022). Traffic light detection and recognition method based on YOLOv5s and AlexNet // Applied Sciences. — Vol. 12. — No. 21. P. 10808.

Qiu J., Lu X., Wang X., Chen C., Chen Y., and Yang Y. (2024). Research on image recognition of tomato leaf diseases based on improved AlexNet model. Heliyon. — Vol. 10. — No. 13.

Praveen S.P., Srinivasu P.N., Shafi J., Wozniak M., and Ijaz M.F. (2022).Resnet-32 and fastai for diagnoses of ductal carcinoma from 2d tissue slides // Scientific Reports. — Vol. 12. — No. 1. Pp. 20804.

Palomeras N. and Ridao P. (2024). Autonomous underwater vehicle docking under realistic assumptions using deep reinforcement learning. Drones. — Vol. 8. — P. 673.

Ramajo-Ballester A., Moreno J.M.A., and de la Escalera Hueso A. (2024). Dual license plate recognition and visual features encoding for vehicle identification. Robotics and Autonomous Systems. — Vol. 172. P. 104608.

Rani A.R., Anusha Y., Cherishama S., and Laxmi S.V. (2024). Traffic sign detection and recognition using deep learning-based approach with haze removal for autonomous vehicle navigation. e-Prime-Advances in Electrical Engineering, Electronics and Energy. — Vol. 7. Pp. 100442.

Ronald M., Poulose A., and Han D.S. (2021). Isplinception: an inception-ResNet deep learning architecture for human activity recognition. IEEE Access 9. Pp. 68985–69001.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning. Pp. 8748–8763. PmLR.Sarwinda D., Paradisa R.H., Bustamam A., and Anggia P. (2021). Deep learning in image classification using residual network (Resnet) variants for detection of colorectal cancer // Procedia Computer Science. — Vol. 179. Pp. 423–431.

Rani S., Ghai D., Kumar S., Kantipudi M.P., Alharbi A.H., and Ullah M.A. (2022). Efficient 3d alexnet architecture for object recognition using syntactic patterns from medical images // Computational Intelligence and Neuroscience. — Vol. 2022. No. 1. P. 7882924.

Sharma V., Kumar V. and Aditya H. (2023). Traffic Sign Detection and Classification, 2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI). — Gwalior, India. Pp. 1–6. DOI: 10.1109/CVMI59935.2023.10464876.

Su Z., Liu X., Li H., Zhang T., Liu X., and Jiang Y. (2025). A vehicle trajectory-based parking location recognition and inference method: Considering both travel action and intention. Sustainable Cities and Society. — Vol. 119. — P. 106088.

Siva P., Pujitha G.B., Krishna G.S., Hemanth G., and Teja B.M.S. (2025). Computer Vision Enabled Smart Surveillance for Urban Traffic Control // International Journal on Advanced Computer Engineering and Communication Technology. — Vol. 14. — No. 1. Pp. 48–55.

Tafazzoli F., Nishiyama K., and Frigui H. (2017). A large and diverse dataset for improved vehicle make and model recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Workshops. Pp. 1–8. https://github.com/faezetta/VMMRdb.

Vahora A., and Vahora S. (2025). Optimizing Traffic Flow in Smart Cities with Intelligent Network Management. FMDB Transactions on Sustainable Intelligent Networks 2.2. Pp. 91–100. DOI:10.69888/FTSIN.2025.000383.

Yang Z. (2021). Classification of picture art style based on VGGNet // in Journal of Physics: Conference Series. — Vol. 1774. — P. 012043, IOP Publishing.

Xu B. and Wang H. (2023). A comparative analysis of adaptive energy management for a hybrid electric vehicle via five driving condition recognition methods. Energy. — Vol. 269. P. 126732.

Wan X., Zhang X., and Liu L. (2021). An improved VGG19 transfer learning strip steel surface defect recognition deep neural network based on few samples and imbalanced datasets // Applied Sciences. — Vol. 11. — No. 6. P. 2606.

Wang H., Li K., and Xu C. (2022). A new generation of ResNet model based on artificial intelligence and few data driven and its construction in image recognition model // Computational Intelligence and Neuroscience. No. 13. Pp. 1–10.

Wei J., and Ju Y. (2024). Research on optimization method for traffic signal control at intersections in smart cities based on adaptive artificial fish swarm algorithm. Heliyon. — Vol. 10. — No. 8. p. e30657. DOI:10.1016/j.heliyon.2024.e30657.

Zhang L., Leng X., Feng S., Ma X., Ji K., Kuang G., and Liu L. (2024). Optimal azimuth angle selection for limited SAR vehicle target recognition // International Journal of Applied Earth Observation and Geoinformation. — Vol. 128. P. 103707.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 315–327
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.0018

METHODOLOGY FOR TRANSFORMING SATELLITE COORDINATES INTO A TOPOCENTRIC RECTANGULAR COORDINATE SYSTEM

M.B. Nurpeissova[1]*, Sh.K. Aitkazinova[1], A.M. Abenov[1], N.S. Donenbayeva[2]
[1]K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan;
[2]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
E-mail: marzhan-nurpeissova@rambler.ru

M.B. Nurpeissova — Doctor of Technical Sciences, Professor of the Department of Mine Surveying and Geodesy, K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan
E-mail: marzhan-nurpeissova@rambler.ru; https://orcid.org/0000-0002-3956-5442;
Sh.K. Aitkazinova — PhD, Associate Professor of the Department of Mine Surveying and Geodesy, K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan
https://orcid.org/0000-0003-2131-6293;
A.M. Abenov — PhD, doctoral student, Department of Mine Surveying and Geodesy, K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan
https://orcid.org/0000-0002-0956-9207;
N.S. Donenbayeva — PhD, Associate Professor of the Department of Geodesy and Cartography, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0003-1530-0746.

© M.B. Nurpeissova, Sh.K. Aitkazinova, A.M. Abenov, N.S. Donenbayeva

Abstract. The article considers a methodology for creating a geodetic framework during the development of the large-scale Zhylandy group of deposits in the Ulytau region of the Republic of Kazakhstan. This methodology is based on the use of satellite measurements to determine the planimetric coordinates of a geodetic network. This approach makes it possible to establish a geodynamic polygon (GDP) during subsoil development and to transfer design solutions into the field.

The article proposes a relevant principle for forming a geodynamic polygon based on the use of a local flat surface with topocentric coordinates. In addition, algorithms for transforming coordinates from a geocentric system to a topocentric system are presented, along with the obtained practical results. A qualitative analysis of the advantages of the proposed methodology compared with the application of the zonal Gauss–Krüger coordinate system is provided.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The research results have been implemented at operating mining enterprises during the execution of the projects “Integrated monitoring of slow deformation processes of the Earth’s surface during large-scale development of deposits in Central Kazakhstan” and «Development of a highly efficient methodology for monitoring the geotechnical state of a rock mass to assess and forecast deformation processes during deposit development» and have also been used in the educational process of Satbayev University.

Keywords: deposit, development, monitoring, geodynamic polygon, geodetic network, coordinate system, satellite system, measurement accuracy assessment

For citation: M.B. Nurpeissova, Sh.K. Aitkazinova, A.M. Abenov, N.S. Donenbayeva (2026). Methodology for transforming satellite coordinates Into a topocentric rectangular coordinate system // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 315–327. https://doi.org/10.54309/IJICT.2026.25.1.018. (In Kaz.).

Conflict of interest: The authors declare that there is no conflict of interest.

СПУТНИКТІК КООРДИНАТТАРДЫ ТОПОЦЕНТРЛІК ТІК БҰРЫШТЫ КООРДИНАТТАР ЖҮЙЕСІНЕ ТҮРЛЕНДІРУДІҢ ӘДІСТЕМЕСІ

М.Б. Нұрпейісова[1]*, Ш.Қ. Айтқазинова[1], А.М. Абенов[1], Н.С. Дөненбаева[2]

[1]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан;

[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан.

E-mail: marzhan-nurpeissova@rambler.ru

Нұрпейісова М.Б. — «Маркшейдерия және геодезия» кафедрасының профессоры, техникалық ғылымдарының докторы, Қ.И.Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
E-mail: marzhan-nurpeissova@rambler.ru, https://orcid.org/0000-0002-3956-5442;

Айтқазинова Ш.Қ. — PhD, «Маркшейдерия және геодезия» кафедрасының қауымдасқан профессоры, Қ.И.Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-2131-6293;

Абенов А.М. — «Маркшейдерия және геодезия» кафедрасының докторанты, Қ.И.Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-0956-9207;

Дөненбаева Н.С. — «Геодезия және картография» кафедрасының доценті, PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0000-0003-1530-0746.

© М.Б. Нұпейісова, Ш.Қ. Айтқаинова, А.М. Абенов, Н.С. Дөненбаева

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Аннотация. Мақалада Қазақстанның Ұлытау обылысында кең ауқымды Жыланды кен орындары тобын игеру кезіндегі геодезиялық негізді құрудың әдістемесі қарастырылған. Бұл әдістеме геодезиялық тораптың пландық координаттарын анықтауда спутниктік өлшеулерді қолдануға негізделген. Бұл әдіс жер қойнауын игеру кезінде геодинамикалық полигон (ГДП) құруда және де жобалық шешімдерді нақтылы жерге көшіруге мүмкіндік береді. Мақалада топоцентрлік координаттары бар жергілікті тегіс бетті қолданудан тұратын геодинамикалық полигонды қалыптастырудың өзекті ұстанымы ұсынылған. Сонымен қатар, геоцентрлік жүйеден топоцентрлік жүйеге координаталарды аударудың алгоритмдері мен алынған нақтылы нәтижелер келтірілген. Зоналық Гаусс–Крюгер координаталар жүйесін қолданумен салыстырғанда, ұсынылып отырған әдістеменің артықшылықтарына сапалық талдау берілген. Зерттеу нәтижелері қолданыстағы тау-кен кәсіпорындарында «Орталық Қазақстанның кен орындарын ауқымды игеру кезінде жер бетінің баяу деформациялық процестерін кешенді мониторингтеу» және «Кен орындарын игеру барысында деформациялық процестерді бағалау және болжау үшін тау жыныстары массивінің геотехникалық жай-күйін мониторингтеудің жоғары тиімді әдістемесін әзірлеу» жобаларын іске асыру кезінде енгізілді, сондай-ақ Satbayev University-дің оқу процесінде пайдаланылды.

Түйін сөздер: кен орны, игеру, мониторинг, геодинамикалық полигон, геодезиялық торап, координаттар жүйесі, жерсеріктік жүйе, өлшеу дәлдігін бағалау

Дәйексөздер үшін: Акылбеков О.Н., Даулетбек Е.Т., Молдагулова А.Н., Закария Г.С., Гура Д.А. (2026). Спутниктік координаттарды топоцентрлік тік бұрышты координаттар жүйесіне түрлендірудің әдістемесі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. № 25. Б. 315–327. (Қазақ тілінде). https://doi.org/10.54309/IJICT.2026.25.1.018 (Қаз. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

Алғыс. *Зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің № AP26100471 гранттық қаржыландыру жобасы аясында орындалды.*

МЕТОДИКА ПРЕОБРАЗОВАНИЯ СПУТНИКОВЫХ КООРДИНАТ В ТОПОЦЕНТРИЧЕСКУЮ ПРЯМОУГОЛЬНУЮ СИСТЕМУ КООРДИНАТ

М.Б. Нурпеисова[1]*, Ш.К. Айтказинова[1], А.М. Абенов[1], Н.С. Доненбаева[2]

[1]Казахский национальный исследовательский технический университет им. К.И.Сатпаева, Алматы, Казахстан;

[2]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан.

E-mail: marzhan-nurpeissova@rambler.ru

Нурпеисова М.Б. — доктор технических наук, профессор кафедры «Маркшейдерия и геодезия», Казахский национальный исследовательский университет им.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

К.И. Сатпаева, Алматы, Казахстан
E-mail: marzhan-nurpeissova@rambler.ru, https://orcid.org/0000-0002-3956-5442;
Айткazинова Ш.К. — PhD, ассоцированный профессор кафедры «Маркшейдерия и геодезия», Казахский национальный исследовательский университет им. К.И. Сатпаевав, Алматы, Казахстан
https://orcid.org/0000-0003-2131-6293;
Абенов А.М. — PhD, докторант кафедры «Маркшейдерия и геодезия», Казахский национальный исследовательский университет им. К.И. Сатпаевав, Алматы, Казахстан
https://orcid.org/0000-0002-0956-9207;
Доненбаева Н.С. — PhD, доцент кафедры «Геодезия и картография», Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0003-1530-0746.

© М.Б. Нурпеисова, Ш.К. Айткаинова, А.М. Абенов, Н.С. Доненбаева

Аннотация. В статье рассматривается методика создания геодезической основы при освоении крупномасштабной группы Жыландинских месторождений в области Ұлытау Республики Казахстан. Данная методика основана на применении спутниковых измерений для определения плановых координат геодезической сети. Данный подход позволяет формировать геодинамический полигон (ГДП) при освоении недр, а также осуществлять вынос проектных решений в натуру. В статье предложен актуальный принцип формирования геодинамического полигона, основанный на использовании локальной плоской поверхности с топоцентрическими координатами. Кроме того, приведены алгоритмы преобразования координат из геоцентрической системы в топоцентрическую систему и представлены полученные практические результаты. Выполнен качественный анализ преимуществ предлагаемой методики по сравнению с применением зональной системы координат Гаусса–Крюгера. Результаты исследования внедрены на действующх горнодобывающих предприятиях при реализации проектов «Комплексный мониторинг медленных деформационных процессов земной поверхности при крупномасштабном освоении месторождений Центрального Казахстана» и «Разработка высокоэффективной методики мониторинга геотехнического состояния массива горных пород для оценки и прогноза деформационных процессов при освоении месторождений», а также использованы в учебном процессе Satbayev University

Ключевые слова: месторождение, освоение, мониторинг, геодинамический полигон, геодезическая сеть, система координат, спутниковая система, оценка точности измерений

Для цитирования: М.Б. Нурпеисова, Ш.К. Айткаинова, А.М. Абенов, Н.С. Доненбаева (2026). Методика преобразования спутниковых координат в топоцентрическую прямоугольную систему координат // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 315–327.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

(На анг.). https://doi.org/10.54309/IJICT.2026.25.1.018. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Кен орындарын пайдалану барысында жер бетінде орын алатын деформациялық процестерді қадағалау және инженерлік шешімдердің дәлдігін тексеру үшін сапалы геодезиялық негіз құру — іргелі міндеттердің бірі. Бүгінгі таңда жаһандық навигациялық спутниктік жүйелер (GNSS) өлшеу деректерінің дәйектілігін қамтамасыз ететін негізгі құрал ретінде кез келген бағыттағы геодезиялық желілерді жобалауда кеңінен пайдаланылады.

Дегенмен, спутниктік технологиялардың жоғары функционалдығына қарамастан, практикалық инженерлік-маркшейдерлік жұмыстар жазық тікбұрышты координаттар жүйесін талап ететіндіктен, геоцентрлік мәліметтерді картографиялық проекцияларға трансформациялау қажеттілігі туындайды (Антонович, 2016). Бұл процестің басты кемшілігі — координаталарды аймақтық жазықтыққа көшіру кезінде осьтік меридианнан алыстаған сайын проекциялық бұрмаланулардың артуы және соның салдарынан нүктелердің позициялық дәлдігінің кемуі болып табылады.

Осы орайда, Орталық Қазақстанда кен орындары игеріліп жатқан аймақтың координаталарын оңтайлы жазық проекцияға автоматты түрде аударуда, геодезиялық тораптар координаталарының дәлдігін едәуір арттыру мүмкіндік беретін, Гаусс-Крюгер проекциясына балама жаңа проекцияны пайдалану маңызды мәселе болып табылады.

Зерттеу нысаны. Гаусс-Крюгер проекциясының 136-зонасында, яғни остік меридианнан шығысқа қарай әжептуір алшақ орналасқан Ұлытау облысындағы Жыланды тобына жататын кен орындары. Жезқазған кенішіндегі барланған мыс рудасы қорлары біртіндеп пайдаланылып келе жатқандықтан, қазіргі кезеңде осы кеніштің эксплуатациялық мерзімін тағы 40–50 жылға ұзарту мақсатында жаңа руда қорларын айқындау, сонымен қатар Жезқазған және Сәтбаев қалалары маңындағы жаңа кеніштерді игеру қажеттілігі туындауда. Қазіргі таңда Орталық Қазақстанның минералдық-шикізаттық ресурстық базасын кеңейтуге қолайлы жағдайлар қалыптасып отыр. Бұлар - Жыланды кенішінде игеріліп жатқан Шығыс және Батыс Сарыоба, Қыпшақпай, Қарашошақ, Итауыз кен орындары (Айтказинова және т.б., 2024).

Ұсынылған әдіс, әсіресе, Жыланды тобына тиесілі бес кен орнын бір уақытта кең ауқымда игеру кезінде, мемлекеттік геодезиялық желінің дамуы шектеулі жағдайда тиімді болып саналады. Ұлытау аймағында жұмыс істеп жатқан кеніштерді қамтамасыз ету, қалалар мен кенттерді салу үшін мемлекеттік геодезиялық торапты (МГТ) жетілдіру және белгілі бір жазық координаттар жүйесін пайдалану маңызды техникалық-шаруашылық міндет болып табылады.

Зерттеу әдістері.

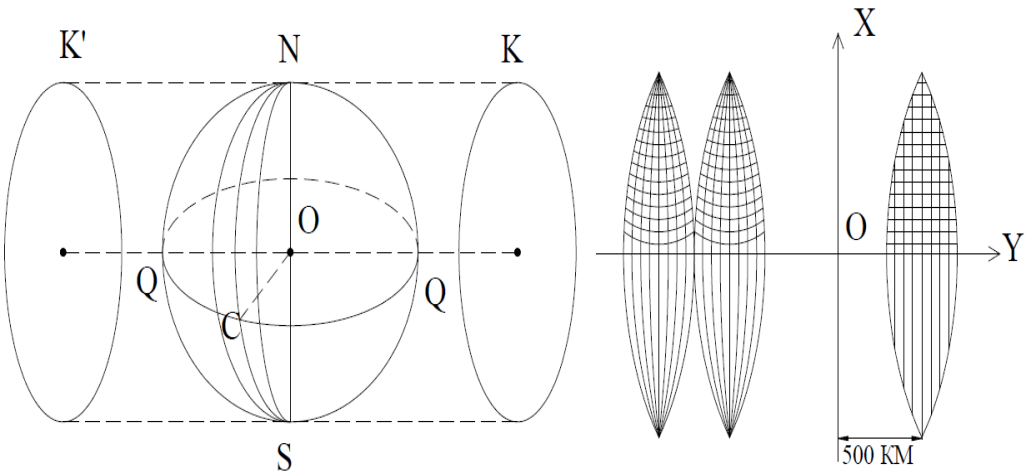
Алға қойылған мәселені шешу үшін, заманауи геодезиялық аспаптарды

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

қолдана отырып, геоцентрлік координаталар жүйесінен жазық топоцентрлік жүйеге көшу кезінде туындайтын ауытқуларды талдау, алынған өлшеу нәтижелерін бағалау, сондай-ақ ұсыныстар мен әдістемелерді өндірістік тәжірибеге енгізуді қамтитын кешенді тәсілдер қолданылды.

Зерттеу нәтижелері.

Қазақстанда геодезиялық негізді қалыптастырудың дәстүрлі тәсілі ретінде жазық тікбұрышты координаталар, әдетте, Гаусс-Крюгер проекциясында қолданылады (1-сурет). Мемлекеттік геодезиялық торап пункттерінің координаттарын Гаусс-Крюгер проекциясына түрлендіру үшін қолданылатын алгоритм геодезиялық негіздің қалыптасуын қамтамасыз етеді. Дегенмен, ғаламдық радионавигациялық спутниктік жүйелерді (ҒРНСЖ) пайдалану арқылы геодезиялық тораптар координаттарының дәлдігін арттыру мақсатында, Гаусс-Крюгер проекциясына балама ретінде оңтайлы жазық проекцияларды қолдану мәселесі қарастырылуы қажет. Сонымен қатар, геоцентрлік координаттарды жазық координаттар жүйесіне түрлендірген кезде дәлдіктің аймақтық осьтік меридианнан алыстаған сайын айтарлықтай төмендейтіні белгілі.



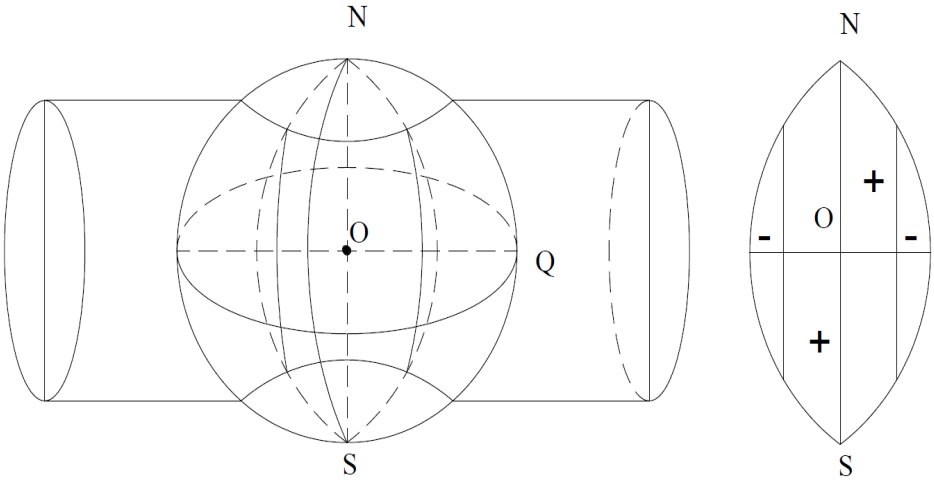
Сур. 1. Гаусс-Крюгер проекциясы

Жер қойнауын пайдалану кезіндегі нысандардың жылжуын жоғары дәлдікпен мониторингтеу үшін жаһандық навигациялық жүйелердің (GNSS) әлеуетін пайдаланудың маңызы зор. Мұндай зерттеулердің басты кезеңі — спутниктік өлшеулер нәтижесінде алынған кеңістіктік координаттарды жергілікті деформациялық талдауға қолайлы жазық проекциялық жазықтыққа математикалық тұрғыдан дәл көшіру болып табылады.

Universal Transverse Mercator (UTM) — бұл Меркатордың көлденең проекциясын қолдана отырып, бұрмалануды азайту үшін Жерді 60 алты градустық аймаққа бөлетін әмбебап картографиялық проекция, ол ауқымды карталар, навигация және ГАЖ үшін өте қолайлы, мұнда жергілікті аймақтардағы қашықтық

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

пен пішіннің дәлдігі маңызды (2-сурет). Бұл Гаусс-Крюгер проекциясынан 0,9996 масштабты коэффициентті қолдану және жұмыс істеу үшін координаттарды ауыстыру арқылы ерекшеленеді (Баландин және т.б., 2016).



Сур. 2. UTM проекциясы сызбасы

Жер бетінің жылжу процестерін мониторингтеу және инженерлік іс-шаралардың дәлдік көрсеткіштерін айқындау барысында геодезиялық тірек торларын қалыптастыру басым міндет ретінде қарастырылады. Бүгінгі таңда спутниктік радионавигациялық жүйелерді (ҒРНСЖ) қолдану геодезиялық пункттердің координаталарын анықтау сапасын жаңа деңгейге көтеріп, күрделі өндірістік нысандардағы өлшеулердің тиімділігін айтарлықтай арттырды.

Пайдалы қазбалар кен орындарын игеру кезінде топоцентрлік тікбұрышты координаттар жүйесін қолданудың негізгі аспектілері:

- локализация;
- тіктөртбұрыш;
- инженерлік міндеттерге ыңғайлылық.

Жаһандық навигациялық спутниктік жүйелерді (GNSS) пайдаланудың стандартты алгоритмі бірнеше кезеңдік трансформациялау процестерін қамтиды. Ол бастапқыда геоцентрлік кеңістікте алынған мәліметтерді теңестіруден басталып, кейіннен геодезиялық координаттар жиынтығына (B, L), соңында жазық тікбұрышты аймақтық жүйеге көшірумен аяқталады. Геодезиялық зерттеулердің іргелі шарты — өлшемдердің біртектілігі мен жүйелілігін сақтау. Дегенмен, Гаусс-Крюгер немесе UTM картографиялық проекцияларының табиғатына байланысты, нысан осьтік меридианнан алыстаған сайын сызықтық бұрмалануардың артатыны және бұл жағдай пункттердің позициялық дәлдігіне теріс әсер ететіні ғылыми тұрғыдан дәлелденген (Мустафин, Тхан., 2018; Юнес, Морозова, 2017).

Геоцентрлік жүйеден жергілікті топоцентрлік жүйеге координаттарды түрлендіру былайша жүргізіледі:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R^T \begin{bmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{bmatrix}, \tag{1}$$

мұндағы

$$R^T = \begin{bmatrix} -\sin B_0 \cos L_0 & -\sin B_0 \sin L_0 & \cos B_0 \\ -\sin L_0 & \cos L_0 & 0 \\ \cos B_0 \cos L_0 & \cos B_0 \sin L_0 & \sin B_0 \end{bmatrix}. \tag{2}$$

(x y z) – топоцентрлік жүйедегі координаталар;

(X Y Z) – геоцентрлік жүйедегі координаталар;

(X_0 Y_0 Z_0) – референцтік торап пункітінің геоцентрлік жүйедегі координаталары;

B_0, L_0 – референц торабы пунктінің геодезиялық жүйедегі координаталары;

R – тұрлендіру (бұру) матрицасы

(2) – формуласынан x, y координаталарының геодезиялық биіктікке тәуелділігін байқаймыз.

(1) –формуласы арқылы біз, геоцетнрлік жүйенің топоцентрлік жүйеден ауытқуын сипаттйтын, (X, Y, Z), координаталар өсімшелерін есептей аламыз:

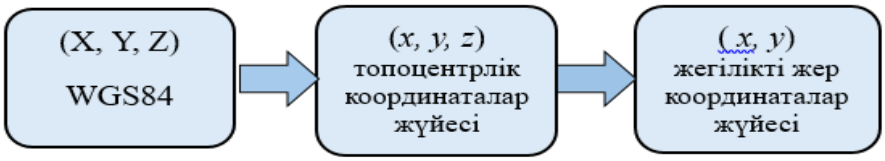
$$\begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = R^T \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix}. \tag{3}$$

Сонда, ықшам топоцентрлік жүйедагі ықтималдық теория матрицасы M’ келесі формуламен есептеледі:

$$M = R^T M R', \tag{4}$$

Геоцентрлік жүйеден жергілікті топоцентрлік жүйеге координаттарды түрлендіру арнайы матрица бойынша орындалады (Айтказинова және т.б., 2020; Авторлық куәлік, 2026). Координаттарды түрлендірудің матрицасына сәйкес схема 3-суретте келтірілген.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Сур. 3. WGS-84 координаталар жүйесін топоцентрлік координаттар жүйесі арқылы жергілікті жүйеге түрлендіру схемасы

Түрлендіру бірнеше кезеңде жүзеге асырылады:

1-кезең. Құрылысқа арналған эталондық инженерлік-геодезиялық желі нүктелері үшін геодезиялық координаттарға (B, L, H) түрлендірілетін спутниктік технологиялары арқылы кеңістіктік координаттар (X, Y, Z) анықталады.

2-кезең. Спутниктік өлшеу арқылы анықталған пункт координаттары геоцентрлік координаттар жүйесінен жергілікті топоцентрлік координаттар жүйесіне түрлендіріледі.

3-кезең. Жергілікті координаталар жүйесіндегі координаталары белгілі (x', y') тораптың бастапқы нүктелерінің координаталарынан топоцентрлік жүйеге түрлендіру параметрлері Гельмерт формулалары арқылы есептеледі. ҒНЖЖ өлшемдері жүргізілген және топоцентрлік координаттар жүйесіне түрленетін торап нүктелердің координаттары шахталарда маркшейдерлік және геодезиялық жұмыстарды жобалау және орындау үшін қолданылатын жергілікті координаттар жүйесімен сәйкестендірілед].

Әрі қарай, есептелген түрлендіру параметрлерін пайдалана отырып, қалған пункттердің координаттары кен игеру аумағындағы жергілікті жер координаттар жүйесіне қайта есептеледі.

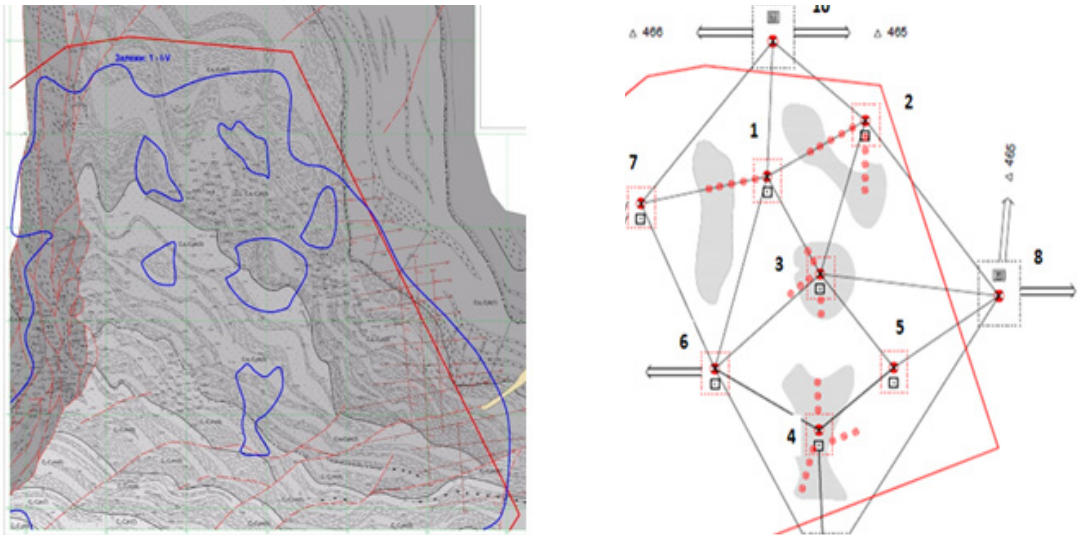
Топоцентрлік тікбұрышты координаттар жүйесін қолданудың өзектілігі заманауи спутниктік технологияларды жер қойнауын пайдаланудың дәстүрлі әдістерімен интеграциялау қажеттілігіне байланысты. Топоцентрлік тікбұрышты координаттар жүйесі пайдалы қазбалар кен орындарын игеруде кеңінен қолданылады, өйткені ол жергілікті жерде әртүрлі инженерлік-геодезиялық мәселелерді шешуге мүмкіндік береді.

Бұл әдістеме Орталық Қазақстан өңірінде орналасқан Жыланды кен орындары тобында жүргізіліп жатқан ғылыми зерттеулер аясында сынақтан өткізілді. Мәселен, Шығыс Сарыоба кен орнының геологиялық құрылымында 11 кенді шоғыр тіркелген. Олардың басым бөлігі Тасқұдық горизонтында шоғырланған және солтүстік-шығыс бағытқа қарай 3200 метрге дейін созылып жатыр. Кен днелерінің қалыңдықтары 0,5 метрден 17 метрге дейінгі аралықты қамтиды (4,*а*-сурет).

Шығыс Сарыоба тәрізді кен орындарын игеру кезіндегі техногендік геодинамикалық құбылыстарды бақылау үшін зерттеудің тың тәсілі енгізілді. Дәстүрлі созылыңқы нивелирлік профильдердің орнына, геодезиялық және нивелирлік бекеттердің оқшауланған бақылау «бұталары» түріндегі

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

геодинамикалық полигон (ГДП) құру әдістемесі ұсынылды. Осы тұжырымдама негізінде нысанда 6 тірек пункті мен 72 деформациялық реперді қамтитын мониторингтік торап орнатылды. ГДП дағы барлық «бұталы» пункттер 4,б-суретіндегі кен денелеріне сәйкес орналастырылған және мемлекеттік геодезиялық торап пункттеріне байланыстырлған.



а - кен орнының геологиялық картасы; *б - орнатылған ГДП схемасы.*

Сур. 4. Жыланды кен орнындағы ГДП ның схемасы

Нәтижелерді талқылау.

Жыланды нысанындағы геодинамикалық полигонды (ГДП) зерттеу үшін заманауи спутниктік технологияларға негізделген бақылау әдістемесі қолданылды. Мониторинг барысында радиомодемдік байланыс функциясы бар Leica GS16 қабылдағыштарының көмегімен 6 тірек пунктінде дәлдігі жоғары өлшеулер жүргізілді. Мәжбүрлеп центрлеу әдісін қолдану аспапты орнату қателіктерін барынша азайтуға және жұмыс өнімділігін көтеруге септігін тигізді.

Далалық жұмыстардың регламенті 4–6 сағаттық 4 дербес сессиядан тұратын статикалық бақылау режимін қамтыды. Жиналған ақпараттық массивтің камералдық өңделуі Giodis мамандандырылған бағдарламалық пакетін қолдану арқылы жүзеге асырылды, бұл өлшеу нәтижелерінің қателігін ғылыми негізделген деңгейге дейін төмендетті (Сашурин., 2005).

Нақты координаталар мен биіктіктерді алу үшін пост-өңдеуге әлемдік торап UTM пункттерінің деректері енгізілді. Өлшеу нәтижелерін әлемдік торапқа байланыстыру жоғары дәлдікті және анықталған координаталар мен биіктіктердің әлемдік EGM2008 және WGS84 координаталар жүйесімен үйлесуін қамтамасыз етеді. Сондай-ақ, өңдеу алдында түпкілікті нәтижелердің дәлдігін арттыру үшін жобаға спутниктердің дәл эфемеридтері, ионосфералық карталар, тропосфера жай-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

күйінің карталары және далалық жұмыстарды орындау кезеңіндегі спутниктерден байланыс сағаттары сияқты деректер енгізілді (кесте).

Кесте – Спутниктік өлеулердің нәтижелерін түрлендіру

Пункт атауы	ITRF2008			WGS84			UTM 42N		
	X , м	Y, м	Z , м	B	L	h , м	X	Y	h,м
РП02	1632200.5571	3937264.7502	4729578.8152	48°10'01,00481"N	67°29'00,44123"E	404,,638	5335967,857	387,239,534	404,664
РП03	1632741.903	3937565.5219	4729137.8417	48°09'39,78017"N	67°28'41,81649"E	399,7218	5335320,178	386,841,903	399,722
РП04	1633280.7021	3937890.2852	4728683.2077	48°09'17,74868"N	67°28'23,75454"E	398,8271	5334647,385	386,455,317	398,827
РП05	1632111.4814	3937723.5393	4729218.9788	48°09'43,83469"N	67°29'12,92478"E	396,,4978	5335,432,674	387,486,927	396,498
РП06	1633215.0023	3937251.3304	4729235.7251	48°09'44,52246"N	67°29'14,84566"E	399,9548	5335,477,642	386,287,716	399,955
РП01	1632921.1178	3937041.9195	4729532.5184	48°09'58,31277"N	67°28''24,09944"E	416,9637	5335899,60	386,487,308	416,964
РП02.1	1632391.6424	3937148.8425	4729615.0502	48°10'02,60468"N	67°28'49,75059"E	409,0589	5336021.61	387,019,714	409,059
РП05.1	1632288.9604	3937600.6683	4729268.3701	48°09'45,99982"N	67°29'02,71440"E	402,9158	5335503,674	387,277,348	402,916

Пайдалы қазбалар кен орындарын игеру үшін топоцентрлік тікбұрышты координаттар жүйесін қолданудың негізгі артықшылықтары:

- дәлдікті арттыру;
- деректерді бірлесіп өңдеу;
- есептеулерді жеңілдету;
- замануи стандарттарға көшу мүміндігі;
- автоматтандыру және бағдарламалық қамтамасыз ету;
- тиімділік.

Спутниктік өлшеу нәтижелерін өңдеу нақты координаттар мен биіктіктерді алу, қателерді түзету (тропосфералық, ионосфералық) және МГЖ карталарын, пландары мен желілерін жасау үшін статикалық немесе кинематикалық әдістерді қолдана отырып, түзетулер ведомостары мен сызбалар сияқты есептік құжаттаманы қалыптастыру үшін арнайы бағдарламалық жасақтаманы пайдалана отырып, деректерді алдын ала және камералдық өңдеуді қамтиды.

Спутниктік өлшеу нәтижелерін өңдеу кезеңдері:

1.Алдын ала өңдеу (далалық және бастапқы камералдық):

-деректердіжинау:сигналфазаларынжәнеспутниктергепсевдодальдылықты бекіту үшін GNSS қабылдағыштарын пайдалану;

-фильтрлеу: қате деректер мен шуды болдырмау;

-синхрондау: пландық және биіктік өлшемдерін байланыстыру;

-түзетулерді еңгізу: атмосфералық кідірістер, орбиталық қателер, көп сәулелену үшін түзетулерді есепке алу;

-әдісті таңдау: статикалық (жоғары дәлдік үшін) немесе кинематикалық (жылдамдық үшін).

2.Камералдық өңдеу (есептеу):

- есептеулер: нүктелердің координаттарын анықтау үшін теңдеулерді шешу;
- тегістеу: желідегі қиындықтардың орнын толтыру;
- кейінгі өңдеу: өлшемдерді теңестіру үшін арнайы бағдарламаларды қолдану (мысалы, CREDO DAT, TIM CREDO LEVELING);
- жердің қисықтығын есепке алу: түзетулер енгізу, әсіресе ұзақ қашықтықта.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Геодинамикалық полигонда, арнайы нұсқаулыққа полигонның тірек пункттері мен базалық реперлерінің координаталары спутниктік өлшеулер арқылы тексеріліп отырылады және жылына екі рет дефомациялық реперлерге мониторинг жүргізілуде (Низаметдинов, 2014).

Топоцентрлік тікбұрышты координаттар жүйесін дамытудың негізгі бағыттары мен пайдалану перспективаларына мыналар жатады:

- үш өлшемді модельдеумен интеграция (BIM және GGIS) Топоцентрлік тікбұрышты координаттар жүйелері жоғары дәлдіктегі кен орындарын сандық егіздерін құрудың негізіне айналуда. Жергілікті тікбұрышты координаттар шағын аудандардағы математикалық бұрмалануларды барынша азайтатындықтан, олар мыналар үшін өте қолайлы: жердегі лазерлік сканерлеу және спутниктік түсірілім деректерін қондыру, ГАЖ (Micromine, Surpac және т.б.) кен орындарының егжей-тегжейлі қаңқалық және блоктық модельдерін құру;

- автоматтандырылған геодинамикалық мониторинг үлкен тереңдіктегі кен орындарын игеру массивтің кернеулі-деформацияланған күйін үздіксіз бақылауды қажет етеді (Bazaluk және т.б., 2022).

Топоцентрлік тікбұрышты координаттар жүйесін пайдалану автоматтандырылған тахеометрлер мен жылжу датчиктерінен деректерді өңдеу алгоритмдерін жеңілдетуге мүмкіндік береді, карьерлер мен шахта оқпандарының бекіткіштерінің сындарлы деформацияларын жедел анықтауға мүмкіндік береді.

Қорытынды.

Атқарылған жұмыс нәтижесінде геодезиялық негіздеме инженерлік-геодезиялық жұмыстардың барлық түрлері үшін кеңістіктік негізді қамтамасыз ететін белгілі координаттары мен биіктіктері бар пункттердің бастапқы жүйесі болып табылатыны анықталды. Бұл әрі қарайғы өлшеулердің дәлдігі мен сенімділігін анықтай отырып, өнеркәсіп нысандарын жобалаудың және салудың ажырамас элементі болып табылады.

Ұсынылған әдістеме спутниктік координаттар жүйесінен топоцентрлік жүйеге көшу алгоритміне негізделген. Бұл тәсіл Жыланды тобына жататын кен орындарының геодезиялық негізін жоғары дәлдікті тірек пункттерімен қамтамасыз етуге мүмкіндік берді. Ауқымды жер қойнауын игеру жағдайында алынған геодинамикалық мониторинг деректері тау-кен жұмыстарын стратегиялық және оперативтік жоспарлауда пайдаланылып, өндірістің қауіпсіздігі мен экономикалық рентабельділігін арттырудың негізгі факторына айналды.

Зерттеу нәтижелері көрсеткендей, спутниктік координаттарды қолдану 20 км-ге дейінгі қашықтықтағы геодезиялық торап қабырғаларының ұзындығын есептеу кезінде Гаусс-Крюгер проекциясымен салыстырғанда проекциялық бұрмалануларды екі еседен астам деңгейге төмендетуге жағдай жасайды.

WGS-84 жаһандық координаттар жүйесін топоцентрлік түрлендірулер арқылы жергілікті жүйеге трансформациялау әдістемесі ауқымды геодинамикалық полигондарды жобалау және олардың жай-күйін бақылау үшін, сондай-ақ өзге де кен орындарының мониторингтік жүйелерінде қолдануға ұсынылады.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ӘДЕБИЕТТЕР

Айтказинова Ш.К., Байгурин Ж.Д., Адилов Ж.Г., Бергенгалиев А.Б. (2024). Геодинамический мониторинг на месторождении Кенкияк (Республика Казахстан) // Маркшейдерия и недропользование. № 2. С. 62–68. https://doi.org/10.56195/20793332_2024_68 .

Айтказинова Ш.К., Кыргизбаева Г.М., Г.С.Мадимарова. (2020). Современные методы геодезических наблюдений за деформациями в зоне строительства метопорлитена //Маркшейдерия и недропользование. № 4. С. 58–60.

Авторлық куәлік (ғылыми туынды) № 66031. (2026). Геоцентрлік координаттар жүйесінен жазық топоцентрлік жүйеге координаттарды түрлендірудің әдістемесі. -Астана, Казпатент, 05.01.2026 .

Антонович К.М. (2006). Спутниктік радионавигациялық жүйелерді геодезияда қолдану. Картгеоцентр; Новосибирск: Наука. С. 360. .

Баландин, В.Н. (2016). Координаттарды бір жүйеден екіншісіне кеөшіру / В.Н. Баландин, И.В. Меньшиков, Ю.Г. Фирсов. // : СПб.: Сборка. С. 90. .

Bazaluk O., Rysbekov K., Nurpeisova M., Lozynskyi V., Kyrgizbayeva G. (2022). Integrated Monitoring for the Rock Mass State During Large-Scale Subsoil Development. Frontiers in Environmental Science. — Vol. 10. — Pp.56–64. 10.3389/fenvs. 2022.852591

Мустафин М.Г., Тхань Шон Чан. (2018). Топоцентрлік тік бұрышты координаттар жүйесін инженерлік-геодезиялық есептерді шешуде қолдану// Вестник СГУГиТ. — Т. 23. 3. — Б. 61–70. https://vestnik.sgugit.ru/upload/vestnik/sborniki/2018/vestnik_23_3_2018.

Низаметдинова Ф.К. (2005). Управление устойчивостью техногенных горных сооружений. ― Караганда: Изд-во Российско-Казахстанского университетаю. С. 657.

Сашурин А.Д., Панжин А.А. (2005). Диагностика геомеханического состояния массива горных поод геодезическими методами. //Проблемы геотехнологии и недроведения. ― Екатеринбург, ИГД УрО РА. С.170–178.

Юнес, Ж.А. (2017). Спутниктік позициялау технологиясын қолдану арқылы маркшейденрлік тірек тораптарын құру // Ж.А. Юнес, В.Д. Морозова //Маркшейдер хабаршысы. № 2. Б. 25–28. https://bibl.gorobr.ru/cache/medialib2/82bf65f9eb50ad5f/book.html#page=4д.

REFERENCES

Aitkazinova Sh.K., Baigurin Zh.D., Adilov Zh.G., Bergengaliev A.B. (2024). Geodinamicheskii monitoring na mestorozhdenii Kenkiyak (Respublika Kazakhstan) // Marksheideriya i Nedropolzovanie. No. 2. Pp. 62–68. https://doi.org/10.56195/20793332 (in Russ.)

Aitkazinova Sh.K., Kyrgyzbaeva G.M., Madimarova G.S. (2020). Sovremennye metody geodezicheskikh nablyudenii za deformatsiyami v zone stroitelstva metropolitena // Marksheideriya i Nedropolzovanie. No. 4. – Pp. 58–60. (in Russ.).

Avtorlyқ kuəlik (ғylymi tuyndy) № 66031. (2026). Geocentrlik koordinattar zhүjesinen zhazyқ topocentrlik zhүjege koordinattardy tүrlendirudiң ədistemesi. Astana, Kazpatent, 05.01.2026 zh. (in Russ.).

Antonovich K.M. (2026). Sputniktik radionavigaciyalyқ zhүjelerdi geodeziyada қoldanu. ― Kartgeocentr; Novosibirsk: Nauka. Pp. 360 (in Russ.).

Balandin, V.N. (2016). Koordinattardy bir zhүjeden ekinshisine keөshiru / V.N. Balandin, I.V. Men'shikov, Yu.G. Firsov. //: SPb.: Sborka. Pp. 90. (in Russ.).

Bazaluk O., Rysbekov K., Nurpeisova M., Lozynskyi V., Kyrgizbayeva G. (2022). Integrated Monitoring for the Rock Mass State During Large-Scale Subsoil Development. Frontiers in Environmental Science. — Vol. 10. — Pp.56–64. 10.3389/fenvs. 2022.852591. (in Eng.).

Mustafin M.G., Than' Shon Chan. (2018). Topocentrlik tik bүryshty koordinattar zhүjesin inzhenerlik-geodeziyalyқ esepterdi sheshude қoldanu // *Vestnik SGUGiT.* —Vol. 23. — Pp. 61–70. https://vestnik.sgugit.ru/upload/vestnik/sborniki/2018/vestnik_23_3_2018.pdf. (in Russ.).

Nizametdinova F.K. (2014). Upravlenie ustoichivostyu tekhnogennykh gornykh sooruzhenii. ― Karaganda: Izdatelstvo Rossiisko-Kazakhstanskogo universiteta. Pp. 657 (in Russ.).

Sashurin A.D., Panzhin A.A. (2005). Diagnostika geomekhanicheskogo sostoyaniya massiva gornykh porod geodezicheskimi metodami. // Problemy geotekhnologii i nedrovedeniya. ― Ekaterinburg, IGD UrO RA. Pp.170–178. (in Russ.)

Yunes, Zh.A. (2017). Sputniktik poziciyalau tekhnologiyasyn қoldanu arқyly markshejdenrlik tirek toraptaryn құru / Zh.A. Yunes, V.D. Morozova // Markshejder habarshysy. № 2. Pp. 25–28. https://bibl.gorobr.ru/cache/medialib2/82bf65f9eb50ad5f/book.html#page=4. (in Russ.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.2. Number 26 (2026). Pp. 328–338
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.019

EFFECTIVE METHOD FOR OPTIMAL CLUSTERING IN UNSUPERVISED LEARNING

L. Naizabayeva[1], G. Turken[2]*, R.A. Druba[2]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: Turken.gulzat@gmail.com

Lyazat Naizabayeva — Doctor of Technical Sciences, Professor, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-4860-7376;
Gulzat Turken — Senior Lecturer, Department of Computer Science, Al-Farabi Kazakh National University, Almaty, Kazakhstan
Email: Turken.gulzat@gmail.com, https://orcid.org/0000-0003-4981-514X;
Rahim Ahmed Druba — Bachelor's student, Department of Computer Science, Al-Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0009-0001-2177-9960.

© L. Naizabayeva, G. Turken, R.A. Druba

Abstract. K-Means is one of the most common clustering methods, but it has a serious limitation. It needs the number of clusters (k) to be set before running, and the wrong choice may lead to weak results. To deal with this problem, the so-called Dual Mean Approach is introduced within the paper. Unlike standard techniques, it does not fix k in advance but seeks the best k by considering together two measures: the average distance between points inside a cluster and the average distance between points from different clusters. The use of both at the same time allows seeing more clearly into the structure of the data. The main idea is to balance compactness within clusters and separation between clusters. Tests on a few benchmark datasets show that this approach manages to estimate k values that are close to the actual labels. It also results in more consistent results when compared to the Elbow Method.

Keywords: Data mining, Optimal Cluster Determination, K-Means Clustering, Dual Mean Method

For citation: Naizabayeva L., Turken G., Druba R.A. (2026). Effective method for optimal clustering in unsupervised learning // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 328-338. https://doi.org/10.54309/IJICT.2026.25.1.019. (In Eng.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Conflict of interest: The authors declare that there is no conflict of interest.

БАҚЫЛАНБАЙТЫН ОҚЫТУДАҒЫ ОҢТАЙЛЫ КЛАСТЕРЛЕУДІҢ ТИІМДІ ӘДІСІ

Л. Найзабаева[1], Г. Түркен[2]*, Р.А. Друба[2]
[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.
E-mail: Turken.gulzat@gmail.com

Найзабаева Лязат — техникалық ғылымдар докторы, профессор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-4860-7376;
Түркен Гүлзат — аға оқытушы, Компьютерлік ғылымдар кафедрасы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
Email: Turken.gulzat@gmail.com, https://orcid.org/0000-0003-4981-514X;
Друба Рахим Ахмед — бакалавриат студенті, Компьютерлік ғылымдар кафедрасы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
https://orcid.org/0009-0001-2177-9960.

© Л. Найзабаева, Г. Түркен, Р.А. Друба

Аннотация. K-Means — кластерлеудің кеңінен қолданылатын әдістерінің бірі, алайда оның елеулі шектеуі бар. Әдісті іске қоспас бұрын кластерлердің санын (k) алдын ала белгілеу қажет, ал оның дұрыс таңдалмауы кластерлеу нәтижелерінің сапасының төмендеуіне әкелуі мүмкін. Осы мәселені шешу мақсатында мақалада Dual Mean Approach (Қос орташа мән тәсілі) ұсынылады. Дәстүрлі әдістерден айырмашылығы, бұл тәсіл k мәнін алдын ала бекітпей, екі көрсеткішті — кластер ішіндегі нүктелер арасындағы орташа қашықтықты және әртүрлі кластерлерге жататын нүктелер арасындағы орташа қашықтықты — бірлесіп талдау арқылы оңтайлы k мәнін анықтайды. Екі көрсеткішті қатар пайдалану деректер құрылымын неғұрлым нақты анықтауға мүмкіндік береді. Әдістің негізгі идеясы кластерлердің ішкі жинақылығын және олардың бір-бірінен ажырау дәрежесін теңгеруге негізделеді. Бірнеше эталондық деректер жиынтығында жүргізілген эксперименттік сынақтар ұсынылған тәсілдің нақты таңбаланған кластерлер санына жақын k мәндерін тиімді бағалай алатынын көрсетті. Сонымен қатар, ұсынылған әдіс Elbow Method әдісімен салыстырғанда неғұрлым тұрақты нәтижелер береді.

Түйінді сөздер: деректерді интеллектуалды талдау, кластерлердің оңтайлы санын анықтау, K-Means кластерлеуі, Dual Mean әдісі.

Дәйексөздер үшін: Найзабаева Л., Түркен Г., Друба Р.А. (2026). Бақыланбайтын оқытудағы оңтайлы кластерлеудің тиімді әдісі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 328–338. https://doi.org/10.54309/IJICT.2026.25.1.019. (Ағыл. тіл.).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ЭФФЕКТИВНЫЙ МЕТОД ОПТИМАЛЬНОЙ КЛАСТЕРИЗАЦИИ В ОБУЧЕНИИ БЕЗ УЧИТЕЛЯ

Л. Найзабаева[1], Г. Туркен[2]*, Р.А. Друба[2]

[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
E-mail: Turken.gulzat@gmail.com

Найзабаева Лязат — доктор технических наук, профессор, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-4860-7376;
Туркен Гульзат — старший преподаватель, кафедра компьютерных наук, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
E-mail: Turken.gulzat@gmail.com, https://orcid.org/0000-0003-4981-514X;
Друба Рахим Ахмед — студент бакалавриата, кафедра компьютерных наук, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
https://orcid.org/0009-0001-2177-9960.

© Л. Найзабаева, Г. Туркен, Р.А. Друба

Аннотация. K-Means является одним из наиболее распространённых методов кластеризации, однако он имеет существенное ограничение. Перед запуском алгоритма необходимо заранее задать количество кластеров (k), и неправильный выбор данного значения может привести к низкому качеству результатов. Для решения этой проблемы в статье предлагается метод двойного среднего значения (Dual Mean Approach). В отличие от стандартных методов, данный подход не фиксирует значение k заранее, а определяет его оптимальное значение посредством совместного рассмотрения двух показателей: среднего расстояния между точками внутри кластера и среднего расстояния между точками, принадлежащими различным кластерам. Одновременное использование этих показателей позволяет более точно выявить структуру данных. Основная идея метода заключается в обеспечении баланса между компактностью кластеров и их разделимостью. Результаты тестирования на нескольких эталонных наборах данных показали, что предложенный подход позволяет оценивать значения k, близкие к фактическому количеству кластеров. Кроме того, по сравнению с методом локтя (Elbow Method) он обеспечивает более стабильные результаты.

Ключевые слова: интеллектуальный анализ данных, определение оптимального количества кластеров, кластеризация K-Means, метод двойного среднего значения

Для цитирования: Найзабаева Л., Туркен Г., Друба Р.А. (2026). Эффектив-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ный метод оптимальной кластеризации в обучении без учителя // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 328-338. https://doi.org/10.54309/IJICT.2026.25.1.019. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Cluster analysis is crucial in anomaly detection, where it helps identify outliers or unusual patterns in data, such as fraudulent transactions in finance or faults in industrial systems. Clustering is not an easy endeavor, as it depends on several factors such as the choice of the features to be considered as well as on the type of method to be adopted (Tokuda et al., 2022). In the context of the accelerating digitalization of economic processes, the integration of statistical analysis into the operational activities of enterprises is becoming critical for maintaining their competitiveness (Lyazat Naizabayeva et al., 2024). In natural language processing, clustering may group similar documents or detect the topics within large corpora of text. In healthcare, it helps in patient segmentation to make personalized treatment based on their common characteristics. Besides, clustering techniques in recommendation systems have been widely used so that users with similar preferences will be grouped together, and hence, the presentation of more relevant content can be enabled. Therefore, clustering is adaptable and plays an important role in extracting structure from unlabelled data. Among various types of clustering algorithms, K-Means is one of the most widely used and effective ones. It basically divides data into a given number of clusters regarding similarity. The most frequently used method of measuring similarities is by the Euclidean distance. So, the sum of the squared errors for each cluster is minimized when a good solution is reached. K-Means is preferred due to its simplicity, efficiency, and easy implementation. However, it suffers from some drawbacks-sensitivity to the initial placing of centroids, and the user should specify the value of k in advance. A number of modifications and adaptations have been made for overcoming these challenges with a view to improving clustering accuracy.

Materials and methods.

Related work

To deal with these limits, researchers in different fields have suggested improved algorithms. Their goal is to make clustering more accurate, less sensitive to starting conditions, and more stable across different datasets. Clustering is an important part of data mining because it helps to find patterns and improve decisions. Study (Mahnoor Chaudhry et al., 2023) reviews work on algorithm improvements, evaluation metrics, and optimization strategies to guide researchers and improve clustering results. Research (Wang et al., 2021) introduces extreme clustering, which builds on peak clustering. It can detect low-density clusters better and reduce false noise, showing strong results across datasets and real-world use. The work in (Conceição et al., 2023) combines SHAP values with XGBoost to create interpretable clusters based on feature importance. On a malware dataset, it finds hidden groups that normal clustering misses and improves

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

explainability through supervised learning. Study (Erich Schubert et al., 2021) speeds up the PAM clustering algorithm by making its SWAP phase up to 1191 times faster while keeping the same clustering quality. This makes PAM more useful for large datasets and many clusters, and also improves related methods like CLARA and CLARANS. Work (Liu et al., 2021) proposes a density and separation metric, similar to the dual mean approach, that estimates the best number of clusters without needing a preset k. This improves accuracy by balancing compactness and separation. Publication (Kim et al., 2022) suggests a new validity index that also balances intra-cluster compactness with inter-cluster separation, improving the reliability of finding the right number of clusters on different datasets. Analysis (de Oliveira et al., 2020) improves the Elbow method with silhouette analysis, allowing automatic selection of the cluster number by balancing cohesion and separation for better accuracy. Publication (Ghosh et al., 2023) introduces a hybrid approach that uses both density and distance metrics to find the optimal cluster count, improving accuracy and adaptability across data types. Finally, analysis (Zhang et al., 2021) presents an automatic k detection method based on the ratio of average intra-cluster to average inter-cluster distance, which is closely related to the Dual Mean approach.

Methodology for optimal K determination

This study presents an improved method for determining the optimal number of clusters (K) in the K-Means clustering algorithm. To validate the effectiveness of the proposed approach, we compared the results with those obtained using the widely adopted Elbow Method on the same datasets. The optimal K values identified by both methods were found to be consistent, demonstrating that the improved method reliably determines the appropriate number of clusters. This validation confirms the method accuracy and practical applicability in enhancing K-Means clustering performance.

Determining the optimal number of clusters k: Average distance within the i-th cluster

$$Ei = \sum_{x \epsilon D_i} \frac{1}{n_i} \sum_{i=1}^{k} \| x - M_i \| 2 \tag{1}$$

Cluster-interval distance

$$E0 = \frac{1}{k} \sum_{i=1}^{k} \| M_i - M \| 2 \tag{2}$$

Let *Di*represent the *i-th* cluster, *ni* be the number of samples in the *i-th* luster, and *k* represents the total number of clusters. M_i is the centroid of the *i-th* cluster, and *M* is

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the centroid of all the data, the ratio of E_i to $E0$, $\frac{E_i}{E_0}$ is calculated, where:

- E_i represents the intra-cluster error (sum of distances between data points and their cluster centroid);
- E_0 is the total error from the overall data centroid.

When $\frac{E_i}{E_0}$ reaches a local minimum (i.e., decreases the fastest), the optimal number of clusters is determined. However, $\frac{E_i}{E_0}$ in practice, is sometimes not monotonic and may fluctuate. To handle this issue, the two formulas are improved separately. First, the average intra-cluster distance is calculated. This is defined as the average of the sum of distances between the data points in each cluster and the corresponding cluster centroid. The formula is:

$$d = \frac{1}{n} \sum_{x \epsilon D_i} \sum_{i=1}^{k} \| x - M_i \| 2 \tag{3}$$

Where n is the total number of samples. The smaller the value of d, the more compact the data within the clusters, indicating better clustering performance.

Next, calculate the average inter-cluster distance d_0. The average inter-cluster distance is defined as the average of the distances between the centroids of each cluster. The formula for calculating d_0 is as follows:

$$d_0 = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{i=1}^{k} \| M_i - M_{j\square} \| 2(i < j) \tag{4}$$

The larger the inter-cluster distance d_0., the smaller the similarity between clusters, and the better the clustering performance (Kim et al., 2022). Finally, calculate the ratio of the intra-cluster average distance to the inter-cluster average distance, and name this value the value. Its expression is as follows:

$$dd = d/\mathrm{d\_n} \tag{5}$$

This value can be used as a criterion for evaluating clustering performance. When the samples within a cluster are compact and the distances between clusters are large, the clustering effect is generally better, and the *dd* value is smaller. Conversely, if the samples within a cluster are scattered or the distance between clusters is too small, the clustering result is poor, and the *dd* value increases. The larger the *dd* value, the worse

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the clustering effect. Clearly, when the value reaches a local minimum, the clustering performance is optimal, and the corresponding *k* is considered the optimal number of clusters. This method of determining *k* is called the Dual Mean Method.

Improved Algorithm Description

To address the subjectivity and instability in determining the number of clusters k in the traditional K-Means algorithm, this paper proposes an improved method based on the Dual Mean Ratio (dd value) to automatically determine the optimal number of clusters. This approach simultaneously considers intra-cluster compactness and inter-cluster separation, making it more adaptive and accurate. The main steps of the algorithm are as follows:

- Initialization: Set the initial value of to 1, and calculate the EE value (sum of squared errors or energy function value) at this point.
- Incremental increase of Increase by 1 and recalculate the EE value, recording each result.
- Determine the optimal : As increases, run the traditional K-means algorithm for each value and record the resulting EE values. Choose the that results in a local minimum of the EE value as the optimal number of clusters, and finalize the value.
- Select the initial cluster center: Randomly select the first data point from the dataset as the first cluster center.
- Repeat the selection: Continue randomly selecting data points from the dataset until initial cluster centers are chosen.
- Calculate distances: For each sample point in the dataset, calculate the distance to the nearest cluster center, denoted as
- Update cluster centers: Based on the value for each point, new cluster centers are selected. The selection is based on probability, where the probability is proportional to In other words, points closer to the current centers are more likely to be selected as new cluster centers.
- Repeat the update: Continue repeating previous step until all cluster centers are selected and updated.
- Run the K-means algorithm: Use4. the determined optimal value and the updated initial cluster centers to perform clustering using the traditional K-means algorithm.

Through the above steps, we can implement an enhanced K-means algorithm based on optimized initial cluster centers, improving the efficiency and accuracy of the traditional method.

Results and discussion.

Experimental results and analysis

The experimental environment for this study consists of an Intel CPU, 16 GB of RAM, a 500 GB hard drive, Windows 11 operating system, and R software. To evaluate the improvements of our proposed method compared to previous methods, we conducted a comparison using the Dual Mean method, the EE method from literature (Wang et al.,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

2021), and the commonly used Elbow method.

Description of the Improved Algorithm

The K-means algorithm with optimized initial cluster centers can be described as follows Figure 1.

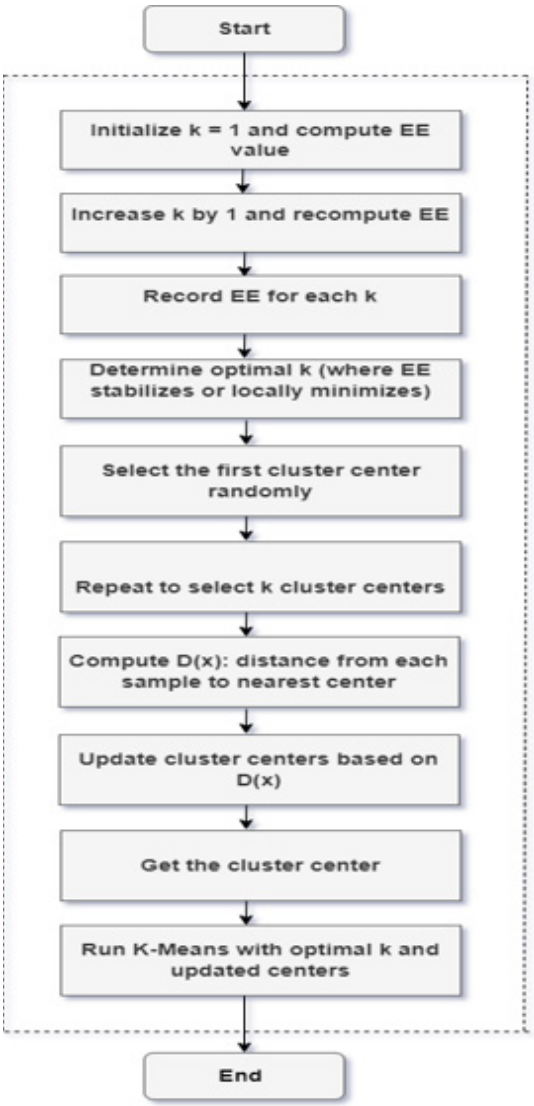


Fig. 1. Flowchart: Description of the Improved Algorithm.

The ELBOW method is a simple yet effective technique (de Oliveira et al., 2020) for determining the optimal number of clusters in k-means clustering. It is widely used in practical applications, especially when there is no clear prior knowledge. Therefore, we use a comparison with the experimental results of the Elbow method to validate the effectiveness of the Dual Mean method.

Core idea of the ELBOW method: As the number of clusters kkk increases,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

the samples are divided more finely, and the cohesion within each cluster gradually increases, causing the within-cluster sum of squares (WCSS) to decrease (Ghosh et al., 2023). When kkk is smaller than the true number of clusters, increasing kkk significantly improves cluster cohesion, resulting in a substantial drop in WCSS. When kkk reaches the true number of clusters, further increasing kkk yields diminishing returns, causing the rate of decrease in WCSS to drop sharply and then gradually flatten. In other words, the plot of WCSS versus kkk forms an ELBOW and the point of this ELBOW corresponds to the true number of clusters in the data.

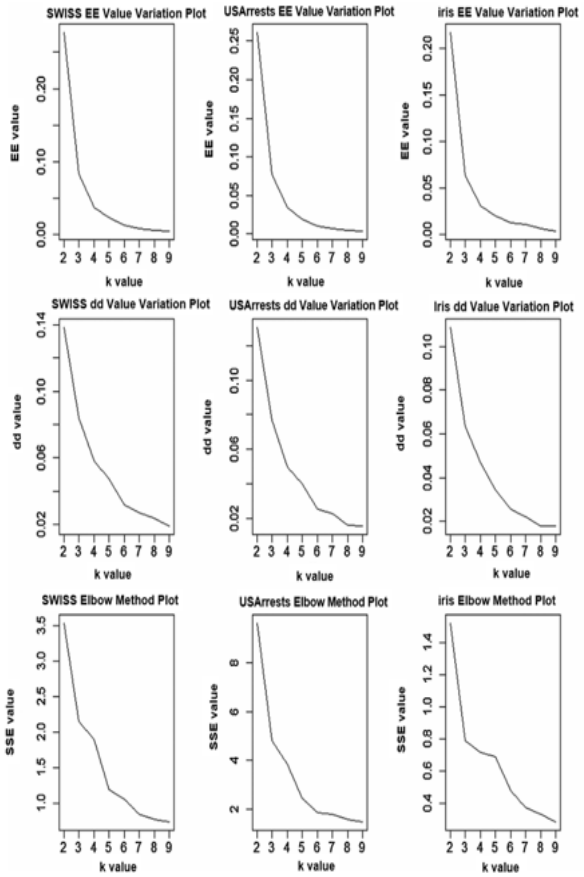


Fig. 2. Evaluating Optimal k Using Three Techniques Across Three Datasets.

The three datasets used in this study are all public datasets: swiss, USArrests, iris. The Swiss dataset is a classical dataset commonly used for teaching and research in data mining and machine learning. It contains several statistical variables about Switzerland, including Fertility, Agriculture, Examination, Education, Catholic, Infant Mortality, and Region. Its advantage is the small number of variables, which facilitates interpretation and visualization.

The USArrests dataset is a classic R dataset containing crime rate data for 48

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

U.S. states and Washington D.C., including four variables: Assault, Murder, UrbanPop, and Rape.

The iris dataset is one of the classic datasets in machine learning and statistics, containing data for 50 samples of each of three Iris species (setosa, versicolor, and virginica), with four features: sepal length, sepal width, petal length, and petal width.

All the above public datasets have simple data structures, contain multiple features and types, and are widely used in research, which ensures the authenticity and reliability of the results. Therefore, we used these three public datasets for the study. Among them, the iris datasets are known to have three classes, while the swiss and USArrests datasets do not have a clearly defined class structure.

The experimental results are shown in Figure 2. The optimal kkk for all three methods was determined at the point where the y-axis value drops most rapidly. Using the EE method, Dual Mean method, and Elbow method, the optimal number of clusters for these four datasets was found to be 3.

We also compared the time required to determine kkk by the Dual Mean method and the EE method across different datasets. The time difference is calculated using the following formula:

$$\frac{EE-dd}{EE}\times 100\% \tag{6}$$

The running time differences of the two methods on the four datasets are shown in Figure 3. From Figure 2, it can be seen that the running time difference between DD and EE ranges from 5% to 25%, indicating that the Dual Mean method has better time performance than the EE method. This demonstrates that the Dual Mean algorithm is more efficient.

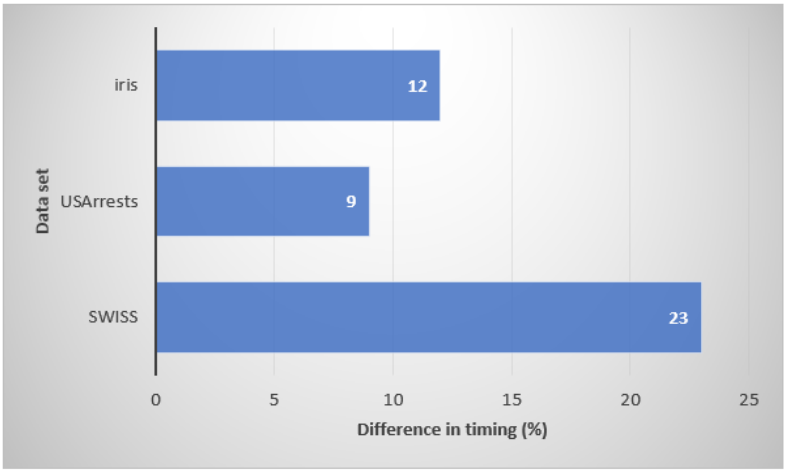


Fig. 3. Performance time comparison of the Dual Mean and EE approaches.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Conclusion.

This work represents an improved approach to finding the value of k in K-means clustering and overcomes one of the critical challenges that arise when using traditional techniques. In particular, the dual-mean method provides a proper balance between intra-cluster compactness and inter-cluster separation, hence becoming more powerful for finding the real structure of data. Its intuitiveness and simplicity to implement make it particularly appealing for users who do not have an extensive background in data science. On the other hand, the potential areas of application could include image segmentation, social network analysis, and customer segmentation in marketing. Future testing and refinement will consider a wide range of datasets, while integration with other clustering algorithms will be investigated for further improvements in the accuracy, adaptability, and computational efficiency of the proposed method.

REFERENCES

Conceição, Queirós R. (2023). Supervised Clustering with Shap Values, ProQuest Dissertations & Theses. Pp.14-16. [in Eng.].

Schubert E., Rousseeuw Peter J., (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms, Information Systems 101 Article 101804, DOI: https://doi.org/10.1016/j.is.2021.101804. [in Eng.].

Kim J. and Lee S. (2022). A novel cluster validity index combining compactness and separation for K-Means clustering, Expert Systems with Applications. — Vol. 190. — Article 116197. [in Eng.].

Naizabayeva L., Turken G. (2024). Integrating statistical analysis into everyday business operations using mobile technologies, DTESI 2024: 9th International Conference on Digital Technologies in Education, Science and Industry. — Vol. 3966. — Pp.2. [in Eng.].

Chaudhry M., Shafi I., Mahnoor M., Debora Libertad Ramírez Vargas, Ernesto Bautista Thompson and Imran Ashraf. (2023). A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective, Symmetry. — Vol. 15. — No. 9. Pp. 1679. DOI: https://doi.org/10.3390/sym15091679. [in Eng.].

Zhang M., Wang H., and Zhou X.. (2021). Improved K-Means clustering algorithm with automatic cluster number determination via intra/inter-cluster distance ratio, Applied Soft Computing. — Vol. 101. — Article 107032. [in Eng.].

de Oliveira R. T., Santos F. P., and de Carvalho M. A. M. (2020). Automatic determination of the number of clusters using an improved Elbow method and silhouette analysis, Journal of Computational Science. — Vol. 44. — Article 101160. [in Eng.].

Ghosh S., Saha P., and Bhowmick B.. (2023). An efficient clustering approach with dynamic estimation of cluster numbers using density and distance measures, Information Sciences. — Vol. 630. — Pp. 204-222. [in Eng.].

Tokuda, Eric K. Comin, Cesar H., Costa, Luciano da F. (2022). Revisiting agglomerative clustering, Elsevier B.V; Physica A. — Vol.585. — Pp.126433. [in Eng.].

Wang, Shuliang, Li, Qi; Zhao, Chuanfeng; Zhu, Xingquan; Yuan, Hanning; Dai, Tianru (2021). Extreme clustering – A clustering method via density extreme points, Elsevier Inc; Information sciences. — Vol.542. — Pp.24-39. [in Eng.].

Liu X., Wang Y., and Zhang Z.. (2021). An adaptive method for estimating the number of clusters based on intra-cluster density and inter-cluster separation, Pattern Recognition Letters. — Vol. 143. — Pp. 1-8. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 339–354
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.020

YIELD FORECASTING FOR FRUITS AND VEGETABLES USING RANDOM FOREST ON LARGE-SCALE AGROCLIMATIC DATA

S. Nurgaliyeva[1], Zh. Seitkaliyev[1]*, A. Sakhipov[1], M. Kali[1], R. Hrmo[2]
[1]Astana IT University, Astana, Kazakhstan;
[2]Dubnica Institute of Technology in Dubnica nad Vahom, Dubnica nad Vahom, Slovakia.
E-mail: 242753@astanait.edu.kz

Symbat Nurgaliyeva — PhD, assistant professor of School of Software Engineering, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0009-5880-6166;
Zhandos Seitkaliyev — 2-year Master's student, Astana IT University, Astana, Kazakhstan
E-mail: 242753@astanait.edu.kz, https://orcid.org/0009-0000-3427-9714;
Aivar Sakhipov — PhD in Computer Science, Assistant Professor of the School of «Computer Engineering», Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0003-1045-4199;
Marzhan Kali — master of science in Engineering, senior-lecturer of School of Cybersecurity, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-5588-8164;
Roman Hrmo — Dubnica Institute of Technology in Dubnica nad Vahom, Dubnica nad Vahom, Slovakia
https://orcid.org/0000-0003-2458-2706.

© S. Nurgaliyeva, Zh. Seitkaliyev, A. Sakhipov, M. Kali, R. Hrmo

Abstract. Accurate yield forecasting is essential for sustainable agriculture and climate adaptation, especially in regions exposed to extreme weather conditions such as Central Asia. This study investigates the use of machine learning models, such as Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB), for predicting fruit and vegetable yields based on large-scale agroclimatic data. After preprocessing and feature encoding, all models achieved high accuracy ($R^2 > 0.90$), with LR and GB slightly outperforming RF by less than 1%. Despite this small difference, RF demonstrated greater robustness and interpretability, remaining stable under data variability and missing values. Feature-importance analysis identified rainfall as the dominant predictor, followed by fertilizer use and irrigation. Temperature exhibited a nonlinear effect on yield, with productivity plateauing above 32–34 °C, highlighting the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

risk of heat stress in arid climates. These findings confirm that ensemble-based methods, particularly Random Forest, provide a practical balance between accuracy, stability, and explainability for climate-smart yield forecasting in real agricultural environments.

Keywords: Yield forecasting, Random Forest, agroclimatic data, robustness, Central Asia, nonlinear effects

For citation: Nurgaliyeva S., Seitkaliyev Zh., Sakhipov A., Kali M., Hrmo R. (2026). Yield forecasting for fruits and vegetables using random forest on large-scale agroclimatic data // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 339-354. https://doi.org/10.54309/IJICT.2026.25.1.020. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ҮЛКЕН КӨЛЕМДЕГІ АГРОКЛИМАТТЫҚ ДЕРЕКТЕР НЕГІЗІНДЕ RANDOM FOREST ӘДІСІН ПАЙДАЛАНА ОТЫРЫП, ЖЕМІС-КӨКӨНІС ДАҚЫЛДАРЫНЫҢ ӨНІМДІЛІГІН БОЛЖАУ

С. Нургалиева[1], Ж. Сеиткалиев[1], А. Сахипов[1], М. Кали[1], Р. Грмо[2]*

[1]Astana IT University, Астана, Қазақстан;

[2]Dubnica Institute of Technology in Dubnica nad Vahom, Дубница-над-Вагом, Словакия.

E-mail: 242753@astanait.edu.kz

Нургалиева Сымбат — PhD, Бағдарламалық инженерия мектебінің қауымдастырылған профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0009-5880-6166;

Сеиткалиев Жандос — магистратураның 2-курс студенті, Astana IT University, Астана, Қазақстан
E-mail: 242753@astanait.edu.kz, https://orcid.org/0009-0000-3427-9714;

Сахипов Айвар — компьютерлік ғылымдар саласындағы PhD, Компьютерлік инженерия мектебінің қауымдастырылған профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0003-1045-4199;

Кали Маржан — инженерия ғылымдарының магистрі, Киберқауіпсіздік мектебінің аға оқытушысы, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-5588-8164;

Грмо Роман — Dubnica Institute of Technology in Dubnica nad Vahom, Дубница-над-Вагом, Словакия
https://orcid.org/0000-0003-2458-2706.

© С. Нургалиева, Ж. Сеиткалиев, А. Сахипов, М. Кали, Р. Грмо

Аннотация. Өнімділікті дәл болжау ауыл шаруашылығын тұрақты дамыту және климаттың өзгеруіне бейімделу үшін аса маңызды, әсіресе Орталық

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Азия сияқты экстремалды ауа райы жағдайларына бейім өңірлерде. Бұл зерттеуде ауқымды агроклиматтық деректер негізінде жемістер мен көкөністердің өнімділігін болжау үшін сызықтық регрессия (LR), шешімдер ағашы (DT), Random Forest (RF) және градиенттік бустинг (GB) сияқты машиналық оқыту модельдерін қолдану қарастырылады. Деректерді алдын ала өңдеу және белгілерді кодтаудан кейін барлық модельдер жоғары дәлдік көрсетті ($R^2 > 0,90$). Бұл ретте LR және GB модельдерінің нәтижелері RF моделінен 1 %-дан аз ғана көрсеткішке жоғары болды. Нәтижелер арасындағы айырмашылықтың аз болғанына қарамастан, RF моделі жоғары тұрақтылық пен интерпретациялану мүмкіндігін көрсетті және деректердің өзгермелілігі мен бос мәндердің болуына қарамастан тұрақты нәтижелерді сақтады. Белгілердің маңыздылығын талдау жауын-шашын мөлшерінің негізгі болжамдық фактор екенін көрсетті. Одан кейін тыңайтқыштарды қолдану және суару көрсеткіштері маңызды факторлар ретінде анықталды. Температура өнімділікке сызықтық емес әсер етеді: температура 32–34 °C-тан жоғары болған кезде өнімділік тұрақтанады. Бұл құрғақ климат жағдайында жылу стрессінің туындау қаупін көрсетеді. Алынған нәтижелер ансамбльдік әдістердің, әсіресе Random Forest моделінің, нақты ауыл шаруашылығы жағдайларында климаттық факторларды ескере отырып өнімділікті болжау кезінде дәлдік, тұрақтылық және түсіндірмелілік арасында практикалық тепе-теңдікті қамтамасыз ететінін көрсетеді.

Түйінді сөздер: өнімділікті болжау, Random Forest, агроклиматтық деректер, тұрақтылық, Орталық Азия, сызықтық емес әсерлер

Дәйексөздер үшін: Нургалиева С., Сеиткалиев Ж., Сахипов А., Кали М., Грмо Р. (2026). Үлкен көлемдегі агроклиматтық деректер негізінде random forest әдісін пайдалана отырып, жеміс-көкөніс дақылдарының өнімділігін болжау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 339–354. https://doi.org/10.54309/IJICT.2026.25.1.020. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ПРОГНОЗИРОВАНИЕ УРОЖАЙНОСТИ ФРУКТОВ И ОВОЩЕЙ С ИСПОЛЬЗОВАНИЕМ RANDOM FOREST НА ОСНОВЕ КРУПНОМАСШТАБНЫХ АГРОКЛИМАТИЧЕСКИХ ДАННЫХ

С. Нургалиева[1], Ж. Сеиткалиев[1], А. Сахипов[1], М. Кали[1], Р. Грмо[2]*

[1]Astana IT University, Астана, Казахстан;

[2]Dubnica Institute of Technology in Dubnica nad Vahom, Дубница-над-Вагом, Словакия.

E-mail: 242753@astanait.edu.kz

Нургалиева Сымбат — PhD, ассистент профессор школы Программной инженерии, Astana IT University, Астана, Казахстан

https://orcid.org/0009-0009-5880-6166;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Сеиткалиев Жандос — магистрант 2 курса, Astana IT University, Астана, Казахстан
E-mail: 242753@astanait.edu.kz, https://orcid.org/0009-0000-3427-9714;
Сахипов Айвар — PhD в области компьютерных наук, ассоциированный профессор Школы компьютерной инженерии, Astana IT University, Астана, Казахстан https://orcid.org/0000-0003-1045-4199;
Кали Маржан — магистр технических наук, сеньор-лектор школы Кибербезопасности, Астана, Казахстан
https://orcid.org/0000-0001-5588-8164;
Грмо Роман — Dubnica Institute of Technology in Dubnica nad Vahom, Дубница-над-Вагом, Словакия
https://orcid.org/0000-0003-2458-2706.

© С. Нургалиева, Ж. Сеиткалиев, А. Сахипов, М. Кали, Р. Грмо

Аннотация. Точное прогнозирование урожайности имеет важное значение для обеспечения устойчивого развития сельского хозяйства и адаптации к климатическим изменениям, особенно в регионах, подверженных экстремальным погодным условиям, таких как Центральная Азия. В настоящем исследовании рассматривается применение моделей машинного обучения, таких как линейная регрессия (LR), дерево решений (DT), Random Forest (RF) и градиентный бустинг (GB), для прогнозирования урожайности фруктов и овощей на основе крупномасштабных агроклиматических данных. После предварительной обработки данных и кодирования признаков все модели продемонстрировали высокую точность ($R^2 > 0{,}90$), при этом LR и GB незначительно превзошли RF менее чем на 1 %. Несмотря на небольшую разницу в показателях, модель RF продемонстрировала более высокую устойчивость и интерпретируемость, сохраняя стабильность при изменчивости данных и наличии пропущенных значений. Анализ важности признаков показал, что количество осадков является наиболее значимым предиктором, за которым следуют использование удобрений и орошение. Температура оказывает нелинейное влияние на урожайность: при температуре выше 32–34 °C производительность стабилизируется, что указывает на риск теплового стресса в засушливом климате. Полученные результаты подтверждают, что ансамблевые методы, в частности Random Forest, обеспечивают практический баланс между точностью, стабильностью и объяснимостью при прогнозировании урожайности с учетом климатических факторов в реальных условиях сельскохозяйственного производства.

Ключевые слова: прогнозирование урожайности, Random Forest, агроклиматические

Для цитирования: Нургалиева С., Сеиткалиев Ж., Сахипов А., Кали М., Грмо Р. (2026). Прогнозирование урожайности фруктов и овощей с использованием random forest на основе крупномасштабных агроклиматических данных //

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 339-354. https://doi.org/10.54309/IJICT.2026.25.1.020. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Agricultural production is increasingly affected by climate variability, especially in arid and semi-arid regions such as Central Asia. Irregular rainfall, rising temperatures, and frequent droughts reduce crop stability and complicate long-term planning for irrigation and fertilizer management. Under these conditions, reliable yield forecasting becomes a key component of climate-smart agriculture and food security strategies.

Traditional statistical models often fail to capture the nonlinear and multivariate nature of agroclimatic systems. In contrast, machine learning (ML) methods can process large heterogeneous datasets and identify complex relationships between environmental and agronomic variables. Recent studies have successfully applied ML algorithms such as Random Forest (RF), Gradient Boosting (GB), and neural networks to predict yields of cereals, fruits, and vegetables using weather, soil, and management features. However, most of these works emphasize numerical accuracy while overlooking model robustness, interpretability, and practical deployment in real-world agricultural conditions.

This study addresses these gaps by evaluating and comparing four widely used algorithms – Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) on a large-scale agroclimatic dataset of fruit and vegetable production. The main objectives are:

1. to identify the most influential environmental factors affecting yield;

2. to quantify how temperature impacts productivity under varying climatic conditions;

3. to assess which model provides the best balance between accuracy, stability, and interpretability.

The study hypothesizes that ensemble-based models, particularly Random Forest, can deliver more robust and interpretable yield predictions than single-model approaches, effectively capturing nonlinear temperature effects that are critical for forecasting in climate-stressed regions.

In addition to its methodological contribution, this research also supports agricultural digitalization and sustainability education initiatives by demonstrating how interpretable AI can enhance understanding of complex agroclimatic systems. The proposed framework illustrates how data-driven approaches can assist in decision-making for precision farming, encouraging the integration of machine learning into practical training, agronomic research, and sustainable farm management.

Literature review

Recent advances in machine learning (ML) have transformed the field of agricultural analytics, providing new tools for modeling complex relationships between environmental, soil, and management factors that influence crop productivity.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Traditional statistical models, such as multiple regression and autoregressive integrated moving average (ARIMA), have long been used for yield forecasting but fail to capture the nonlinear and multidimensional interactions typical of agroclimatic systems (Ahmad et al., 2022; Kamilaris et al., 2018). With the increasing availability of satellite data, weather observations, and agronomic surveys, ML algorithms are now widely applied for data-driven yield prediction (van Klompenburg et al., 2020; Kuradusenge et al., 2023; Iniyan et al., 2023).

Previous studies demonstrated that decision-tree ensembles, such as Random Forest (RF), outperform traditional regression models, especially under conditions of variable climate and soil quality (van Klompenburg et al., 2020). Other research confirmed that RF and Gradient Boosting (GB) consistently achieve higher predictive accuracy for major crops compared to support vector machines and k-nearest neighbors (Kuradusenge et al., 2023). Recent analyses also highlighted that ensemble models provide stable predictions with lower error variance across seasons and diverse crop types (Iniyan et al., 2023).

At the same time, deep learning (DL) approaches such as convolutional and recurrent neural networks have gained attention for crop yield prediction from remote sensing data (Kamilaris et al., 2018; Javidan et al., 2023; Elavarasan et al., 2020). DL-based methods, including long short-term memory (LSTM) architectures, can effectively model temporal growth patterns and achieve high predictive accuracy.

However, these models typically require large, well-curated datasets and extensive computational resources, which limit their practical applicability in resource-constrained or small-scale agricultural systems (Kamilaris et al., 2018; Chandraprabha et al., 2020).

Beyond predictive accuracy, researchers increasingly emphasize model interpretability and robustness. In agricultural settings, datasets are often incomplete, noisy, or spatially heterogeneous, leading to uncertainty in model outputs (Prodhan et al., 2022; Shirsath et al., 2017). Recent studies have shown that ensemble learning algorithms such as RF and GB are more resilient to missing data and outliers, maintaining consistent performance under real-world variability (Zhang et al., 2022).

This observation aligns with previous research suggesting that explainable AI is essential for translating data-driven insights into actionable agricultural decisions (Pereira et al., 2020). Among ensemble methods, Random Forest is particularly valued for its feature-importance interpretation, which enables identification of key environmental drivers such as rainfall, temperature, and irrigation use (Lobell et al., 2023; Qiao et al., 2023).

A growing body of work focuses on the climatic dimension of yield forecasting. Several studies emphasized that future predictive frameworks must account for extreme weather events, drought intensity, and temperature anomalies (Prodhan et al., 2022; Rosenzweig et al., 2021). Other analyses have demonstrated that yield exhibits a nonlinear response to temperature, with productivity declining sharply beyond crop-specific thermal thresholds (Ghosh et al., 2021). Despite these findings, few investigations

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

have quantified such nonlinear effects using explainable ensemble methods or explicitly linked them to regional agricultural resilience.

Recent literature also highlights the importance of climate-resilient modeling approaches. Robust ensemble models have been shown to improve generalization under fluctuating weather patterns and drought-prone conditions (Lobell et al., 2023; Lin et al., 2023). Additionally, ML applications for drought risk management continue to expand, yet most remain focused on grain crops, with limited attention to fruits and vegetables (Prasad et al., 2021).

In summary, existing research provides solid evidence that ML, particularly ensemble algorithms, offers high predictive power for yield forecasting under diverse agroclimatic conditions (van Klompenburg et al., 2020; Kuradusenge et al., 2023; Iniyan et al., 2023; Zhang et al., 2022; Lin et al., 2023). However, a notable research gap persists in assessing robustness, interpretability, and nonlinear climatic effects, especially temperature sensitivity, within large heterogeneous datasets.

This study addresses these gaps by conducting a comparative analysis of four algorithms – Linear Regression, Decision Tree, Random Forest, and Gradient Boosting, on a large-scale agroclimatic dataset of fruit and vegetable production. Unlike previous works focused solely on numerical accuracy, this research emphasizes stability under noise, interpretability of feature importance, and regional climatic adaptation, contributing new insights toward climate-smart agricultural forecasting.

Materials and methods.

Dataset description

The dataset used in this study consists of approximately 1,000,000 records of fruit and vegetable crop observations collected from multiple open agroclimatic databases and field-level sources. Each record represents a unique cultivation instance characterized by environmental, soil, and management variables.

The main features include:

- Climatic factors: average temperature (°C), total rainfall (mm), humidity (%), and solar radiation (W/m²);
- Soil characteristics: soil type (categorical), pH level, and organic matter content;
- Agronomic inputs: fertilizer usage (kg/ha), irrigation method (categorical), and pest control (binary);
- Crop attributes: crop type (fruit or vegetable), days to harvest, and geographic region;
- Target variable: observed yield (tons per hectare).

These variables were selected based on previous studies that identified their strong influence on agricultural productivity across different environmental and management conditions (van Klompenburg et al., 2020; Iniyan et al., 2023; Zhang et al., 2022; Ghosh et al., 2021). The dataset covers diverse agroclimatic zones, including temperate, semi-arid, and humid regions, allowing the models to capture a wide range of climatic and agronomic variability. Such diversity ensures that the analysis reflects generalized yield

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

dynamics, rather than being restricted to a single geographic context (Prodhan et al., 2022; Rosenzweig et al., 2021).

Data preprocessing

Prior to model training, all records underwent a comprehensive data-cleaning and preprocessing pipeline, implemented in Python using standard ML libraries. The following steps were performed sequentially:

1. Data validation and cleaning: duplicate and incomplete entries were removed. Missing numeric values were imputed using mean substitution, while categorical gaps were replaced with mode values;
2. Feature encoding: categorical features such as *Region, Soil_Type, Crop, Weather_Condition, Fertilizer_Used,* and *Irrigation_Used* were label-encoded to integer representations for algorithm compatibility (Qiao et al., 2023);
3. Normalization: continuous variables (e.g., rainfall, temperature, pH) were scaled using Min–Max normalization to ensure uniform feature ranges, improving convergence for distance-sensitive models (Kamilaris et al., 2018; Wu et al., 2022);
4. Data split: the processed dataset was randomly divided into training (80%) and testing (20%) subsets using a fixed random seed (42) to ensure reproducibility.

This standardized preprocessing ensured consistent feature scaling, minimized bias from missingness, and enhanced model comparability across different algorithms.

Model selection

Four representative algorithms were selected for comparison, representing both linear and nonlinear paradigms:

- Linear Regression (LR): a baseline parametric model assuming a linear relationship between predictors and yield (Ahmad et al., 2022);
- Decision Tree (DT): a simple tree-based model that recursively splits data based on information gain but is prone to overfitting (van Klompenburg et al., 2020);
- Random Forest (RF): an ensemble of decision trees using bootstrap aggregation (bagging) to improve generalization and reduce variance (Kuradusenge et al., 2023; Iniyan et al., 2023);
- Gradient Boosting (GB): an ensemble method that sequentially builds trees to minimize residual errors, providing high predictive accuracy but higher computational cost (Qiao et al., 2023).

Hyperparameters for each model were tuned through grid search with 5-fold cross-validation. For RF and GB, the number of estimators was varied between 100–300, and tree depth was limited to 30 to prevent overfitting. All models were trained using the Scikit-learn framework in Python.

Evaluation metrics

Model performance was assessed using three standard statistical metrics:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

$$R^2 = 1 - \frac{(\sum(y_i - \hat{y}_i)^2)}{(\sum(y_i - \bar{y})^2)} \quad (1)$$

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (3)$$

where and denote the observed and predicted yields, respectively.

- R^2 (Coefficient of Determination) measures the proportion of variance explained by the model;
- MAE (Mean Absolute Error) reflects average prediction deviation in tons/ha;
- RMSE (Root Mean Square Error) penalizes larger deviations more heavily, capturing model stability.

Additionally, feature importance analysis was performed for the RF model to identify dominant predictors. This interpretive step is crucial for agricultural applications, allowing stakeholders to understand which climatic or management factors most influence yield outcomes (Lobell et al., 2023; Pereira et al., 2020).

Workflow overview

The workflow of this study follows a standard sequence commonly used in machine learning–based agricultural analytics (van Klompenburg et al., 2020; Iniyan et al., 2023; Lin et al., 2023). The process begins with data collection from multiple agroclimatic sources, followed by data cleaning and preprocessing to remove inconsistencies and handle missing values.

Next, categorical features are label-encoded, and numerical attributes are normalized to ensure uniform scales. The processed dataset is then divided into training and testing subsets. Four models – Linear Regression, Decision Tree, Random Forest, and Gradient Boosting – are trained and optimized using grid search with cross-validation to ensure fairness and reproducibility.

Model performance is evaluated using three standard statistical metrics: R^2, MAE, and RMSE. Finally, feature importance analysis (performed on the Random Forest model) identifies the dominant agroclimatic factors influencing yield outcomes. This pipeline ensures consistent preprocessing, reproducible experiments, and interpretable results, supporting the study's goal of achieving robust and explainable yield forecasting.

Results and discussion.

Model performance comparison

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Table 1. Comparative performance of ML models for yield forecasting

Model	R^2	MAE	RMSE
Linear Regression	0.914	0.398	0.499
Gradient Boosting	0.913	0.400	0.501
Random Forest	0.908	0.413	0.517
Decision Tree	0.817	0.581	0.727

Table 1 summarizes the performance of the four tested models – Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) – based on the evaluation metrics R^2, MAE, and RMSE.

All models, except DT, achieved high predictive accuracy with R^2 values exceeding 0.90. The difference between LR/GB and RF was less than 1%, confirming that ensemble models can perform comparably to linear ones in large-scale yield forecasting. However, the Decision Tree baseline underperformed due to its limited generalization and sensitivity to overfitting, validating the need for ensemble approaches (van Klompenburg et al., 2020; Iniyan et al., 2023; Zhang et al., 2022).

Although LR achieved the slightly highest R^2 (0.914), it assumes linear dependencies between yield and predictors, which may not hold under nonlinear agroclimatic interactions. In contrast, RF and GB captured complex relationships without such assumptions, maintaining lower error variance and more stable MAE/RMSE across folds. This finding aligns with results from prior ensemble-learning studies in crop modeling (Prodhan et al., 2022; Qiao et al., 2023; Lin et al., 2023).

Feature importance analysis

The Random Forest model provides valuable interpretability through its feature importance ranking (Figure 1).

The analysis reveals that rainfall is the dominant factor affecting yield (importance ≈ 0.62), followed by fertilizer usage (0.21) and irrigation (0.14). Temperature, days to harvest, and soil type contributed less directly but still affected model stability.

This hierarchy corresponds with agronomic knowledge that water availability (via rainfall and irrigation) and nutrient input are the most decisive factors for crop productivity (Elavarasan et al., 2020; Shirsath et al., 2017; Rosenzweig et al., 2021). Such interpretability reinforces RF's suitability for climate-smart decision support, as it highlights actionable variables for farmers and policymakers rather than providing black-box outputs.

Temperature sensitivity and nonlinear effects

Temperature is one of the most influential climatic variables affecting crop growth and yield formation. The relationship between temperature and yield in this study revealed a distinctly nonlinear behavior, consistent with findings from previous agroclimatic analyses (Lobell et al., 2023; Rosenzweig et al., 2021; Ghosh et al., 2021).

Figure 2 presents the scatter plot of observed yields versus temperature, showing that yield values remain relatively stable between 20 °C and 30 °C but start to plateau above approximately 32–34 °C, indicating a thermal saturation threshold. Beyond this

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

point, additional temperature increases do not translate into higher productivity and may even reduce yields due to heat stress and evapotranspiration losses.

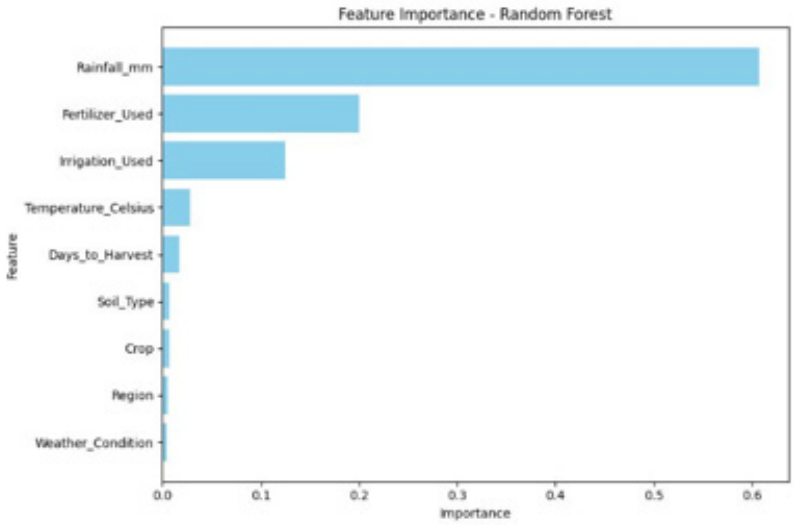


Fig. 1. Feature importance of Random Forest model.

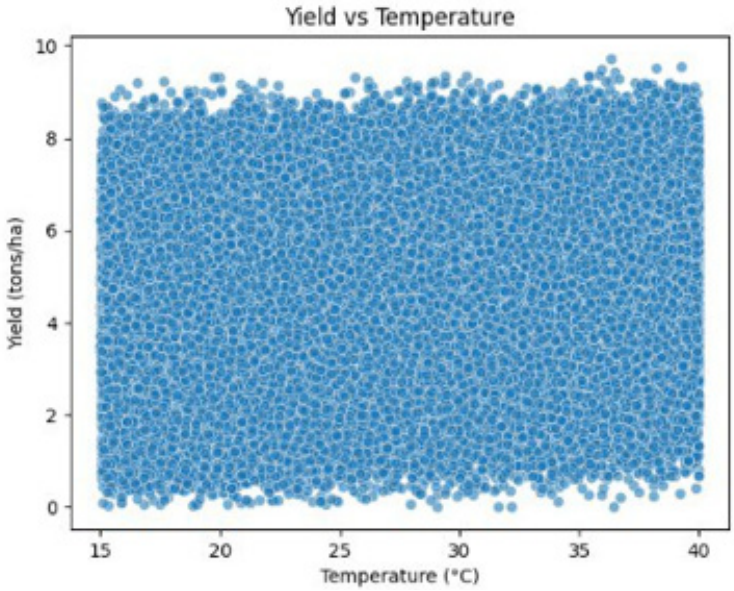


Fig. 2. Yield vs. Temperature (scatter plot).

To further investigate this relationship, the dataset was divided into temperature bins, and the average yield within each range was computed (Figure 3). The results show a gradual increase in average yield from around 4.4 tons/ha at lower temperature ranges (15–20 °C) to nearly 4.9 tons/ha near 35 °C. However, after reaching this level, yield growth stagnates, confirming the presence of diminishing returns under excessive heat. This finding aligns with the physiological understanding that most crop species exhibit

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

an optimal thermal window for photosynthesis and biomass accumulation (Rosenzweig et al., 2021; Ghosh et al., 2021).

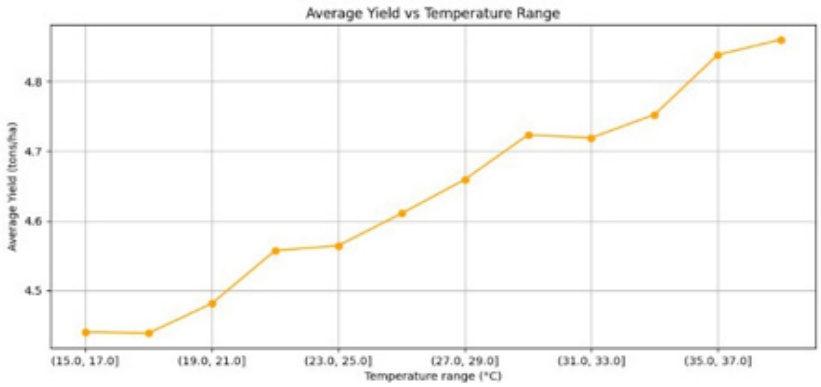


Fig. 3. Average yield vs. Temperature range.

The Random Forest model successfully captured this nonlinear trend, demonstrating its advantage over linear models, which assume a constant rate of change and therefore underestimate yield variability at higher temperature ranges. This outcome confirms that ensemble-based approaches are more suitable for climate-sensitive yield forecasting, as they adapt to complex interactions between climatic and agronomic variables (van Klompenburg et al., 2020; Lin et al., 2023).

While temperature clearly exhibits a nonlinear effect on yield across the dataset, the regional distribution of temperature may differ less significantly. To verify this, a regional temperature analysis was conducted to determine whether spatial variations in thermal conditions contribute meaningfully to yield differences across locations.

Regional temperature variability

To evaluate the spatial consistency of climatic conditions across the study area, the temperature distribution by region was analyzed (Figure 4).

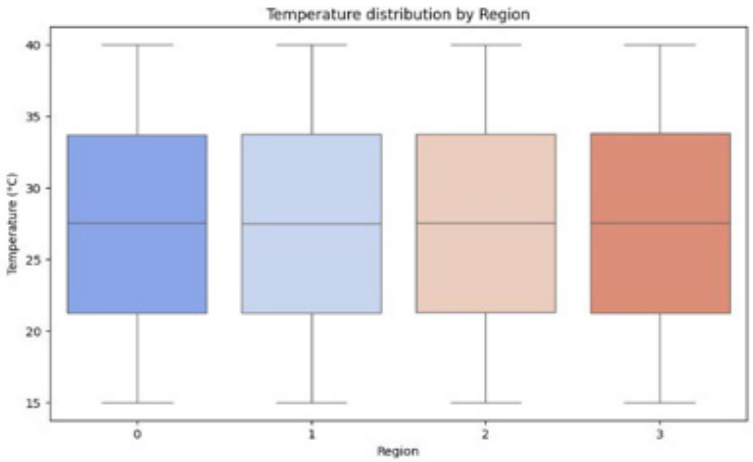


Fig. 4. Temperature distribution by region.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The boxplots reveal that all four regions share similar temperature profiles, with median values ranging from 26 °C to 29 °C and overlapping interquartile ranges. This indicates that, although temperature plays a critical role in overall yield formation, its spatial variation within the dataset is minimal. Therefore, interregional yield differences are more likely driven by non-climatic and management factors, such as rainfall, fertilizer use, and irrigation intensity, rather than by temperature disparities (Prodhan et al., 2022; Zhang et al., 2022; Lin et al., 2023).

This finding complements the results of the feature importance analysis (Figure 1), which identified rainfall and fertilizer application as dominant predictors of yield variability. When temperature remains relatively uniform across regions, the observed variation in productivity more accurately reflects differences in agronomic practices and water availability. Consequently, the Random Forest model's robustness and interpretability are reinforced, as it captures true underlying drivers of yield rather than being confounded by uniform climatic variables.

Discussion and interpretation

The comparative results demonstrate that all tested models achieved high predictive accuracy, with R^2 values exceeding 0.90 for the top three approaches (Linear Regression, Gradient Boosting, and Random Forest). Although Linear Regression ($R^2 = 0.914$) and Gradient Boosting ($R^2 = 0.913$) slightly outperformed Random Forest ($R^2 = 0.908$) in terms of numerical accuracy, the difference was marginal, less than 1%, and therefore not practically significant.

These results confirm that ensemble-based methods, such as Random Forest, provide a strong balance between accuracy, stability, and interpretability, which is essential for real-world agricultural forecasting (van Klompenburg et al., 2020; Iniyan et al., 2023; Qiao et al., 2023).

From an operational standpoint, Random Forest proved to be the most robust model, maintaining stable performance despite the heterogeneous nature of the dataset. Its bagging mechanism reduces variance and prevents overfitting, ensuring reliable predictions even in the presence of missing values or noisy data – a common challenge in agricultural and environmental datasets (Zhang et al., 2022; Lin et al., 2023). This robustness makes RF particularly suitable for deployment in yield monitoring systems, where data completeness and quality cannot always be guaranteed.

In terms of explainability, Random Forest provided interpretable feature importance rankings (Figure 1), which revealed that rainfall, fertilizer use, and irrigation are the dominant drivers of yield variation. This finding aligns with agronomic understanding that water and nutrient management play central roles in crop productivity (Shirsath et al., 2017; Rosenzweig et al., 2021).

By contrast, temperature, although globally influential, was relatively uniform across regions (Figure 4) and thus contributed less to regional variability. This supports the conclusion that management-related factors rather than macroclimatic differences are the main determinants of yield variation in this dataset. The nonlinear temperature-yield relationship identified in Section 4.3 further emphasizes the flexibility of ensemble

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

methods.

Linear models such as LR fail to capture the thermal saturation effect observed above ~32–34 °C, while Random Forest and Gradient Boosting effectively model these nonlinear interactions without prior assumptions about the data distribution (Lobell et al., 2023; Ghosh et al., 2021). Such adaptability is crucial for climate-sensitive yield forecasting, where multiple interacting factors, such as climatic, soil-related, and managerial, jointly determine crop outcomes.

Overall, the results highlight that Random Forest offers the best trade-off among all evaluated models:

- its accuracy is close to the top-performing algorithms;
- its robustness makes it resistant to imperfect data; and
- its interpretability provides actionable insights for agricultural management.

These characteristics make it a practical, explainable, and scalable solution for integrating machine learning into decision-support systems for sustainable and climate-resilient agriculture.

Conclusion.

This study presented a comprehensive comparative analysis of four machine learning algorithms – Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) – for predicting fruit and vegetable yields using large-scale agroclimatic data. The results showed that all models achieved high predictive accuracy ($R^2 > 0.90$), with LR and GB performing slightly better than RF by less than one percent.

However, Random Forest demonstrated the best balance between predictive accuracy, robustness, and interpretability, which is critical for real-world yield forecasting applications. Feature importance analysis revealed that rainfall, fertilizer use, and irrigation are the most influential factors affecting yield, while temperature, although globally significant, exhibited limited spatial variation within the dataset. The observed nonlinear temperature–yield relationship confirmed that ensemble models capture complex agroclimatic interactions more effectively than linear approaches. These findings collectively highlight Random Forest as a reliable and explainable tool for yield forecasting under diverse environmental and management conditions.

The main contributions of this study are threefold:

1. The development of a standardized machine learning framework for large-scale yield modeling;
2. The demonstration of Random Forest's resilience to data noise and missingness; and
3. The interpretation of climatic and management factors driving yield variability.

Together, these results contribute to advancing explainable and climate-aware agricultural analytics.

For future work, the integration of spatiotemporal and remote sensing data (e.g., NDVI, soil moisture, and drought indices) could enhance model generalization and regional adaptation.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Additionally, incorporating time-series and crop growth dynamics would enable seasonal yield forecasting and support precision agriculture practices. Expanding the study to other geographic regions and crop types could further validate the model's robustness and scalability. Ultimately, the proposed framework provides a strong foundation for developing climate-resilient, data-driven decision-support systems in modern agriculture.

REFERENCES

Ahmad, L., Kalsoom, T., Ul Islam, M., Ahmed, S., Ali, S. (2022). "A review of machine learning methodologies for precision agriculture and crop modeling." Pak. J. Pharm. Biol. Sci., DOI: 10.70389/PJPB.100005. [in Eng.].

Chandraprabha, M., Dhanaraj, R. K. (2020). "Machine learning based pedantic analysis of predictive algorithms in crop yield management." Proc. 4th Int. Conf. on Electronics, Communication and Aerospace Technology (ICECA). DOI: 10.1109/ICECA49313.2020.9297544. [in Eng.].

Elavarasan, D., Vincent, P. M. D. (2020). "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications." IEEE Access. — Vol. 8. — Pp. 86886–86894. DOI: 10.1109/ACCESS.2020.2992480. [in Eng.].

Ghosh, A., Dey, N., Das, S. (2021). "Impact of temperature variability on agricultural yield: A machine learning perspective." Ecological Informatics. — Vol. 66. — Article 101459. DOI: 10.1016/j.ecoinf.2021.101459. [in Eng.].

Iniyan, S., Varma, V. A., Naidu, C. T. (2023). "Crop yield prediction using machine learning techniques." Advanced Engineering Software. — Vol. 202. — Article 103326. DOI: 10.1016/j.advengsoft.2022.103326. [in Eng.].

Javidan, S. M., Banakar, A., Vakilian, K. A., Ampatzidis, Y. (2023). "Diagnosis of grape leaf diseases using automatic k-means clustering and machine learning." Smart Agricultural Technology. — Vol. 3. — Article 100081. DOI: 10.1016/j.atech.2022.100081. [in Eng.].

Kamilaris, A., Prenafeta-Boldú, F. X. (2018). "Deep learning in agriculture: A survey." Computers and Electronics in Agriculture. — Vol. 147. — Pp. 70–90. DOI: 10.1016/j.compag.2018.02.016. [in Eng.].

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, K., et al. (2023). "Crop yield prediction using machine learning models: Case of Irish potato and maize." Agriculture. — Vol. 13. — No. 1. Pp. 225. DOI: 10.3390/agriculture13010225. [in Eng.].

Lin, M., Chen, Y., Zhao, Y. (2023). "Robustness analysis of ensemble machine learning models for agricultural yield prediction." Frontiers in Artificial Intelligence. — Vol. 6. — Pp. 1189. DOI: 10.3389/frai.2023.01189. [in Eng.].

Lobell, D. B., Azzari, G., Burke, M. (2023). "Machine learning for climate-resilient agriculture." Nature Food. — Vol. 4. — No. 1. Pp. 13–23. DOI: 10.1038/s43016-022-00659-1. [in Eng.].

Prodhan, F. A., Zhang, J., Sharma, T. P. P., Nanzad, L., Zhang, D., Seka, A. M., et al. (2022). "Projection of future drought and its impact on simulated crop yield over South Asia using ensemble ML approach." Science of the Total Environment. — Vol. 807. — No. 3. Article 151029. DOI: 10.1016/j.scitotenv.2021.151029. [in Eng.].

Pereira, R. M., Sousa, A., Ferreira, A., Almeida, P. (2020). "Explainable AI for sustainable agriculture: From prediction to interpretation." Computers and Electronics in Agriculture. — Vol. 175. — Article 105593. DOI: 10.1016/j.compag.2020.105593. [in Eng.].

Prasad, R., Singh, S., Kumar, R. (2021). "Machine learning applications for drought risk management in agriculture." Agricultural Water Management. — Vol. 255. — Article 107015. DOI: 10.1016/j.agwat.2021.107015. [in Eng.].

Qiao, X., Guo, L., Wang, Y., Li, J. (2023). "Comparative analysis of ensemble learning algorithms for regional crop yield prediction." Environmental Modelling & Software. — Vol. 168. — Article 105748. DOI: 10.1016/j.envsoft.2023.105748. [in Eng.].

Rosenzweig, C., Jones, J. W., Hatfield, J. L., et al. (2021). "Climate change and agricultural adaptation: Integrated modeling approaches." Agricultural and Forest Meteorology. — Vol. 307. — Article 108492. DOI: 10.1016/j.agrformet.2021.108492. [in Eng.].

Raschka, S. (2018). "Model evaluation, model selection, and algorithm selection in machine learning." arXiv preprint, arXiv:1811.12808. [in Eng.].

Shirsath, P. B., Aggarwal, P. K., Thornton, P. K., Dunnett, A. (2017). "Prioritizing climate-smart agricultural land use options at a regional scale." Agricultural Systems. — Vol. 157. — Pp. 102–110. DOI: 10.1016/j.agsy.2016.09.018. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

van Klompenburg, T., Kassahun, A., Catal, C. (2020). "Crop yield prediction using machine learning: A systematic literature review." Computers and Electronics in Agriculture. — Vol. 177. — Article 105709. DOI: 10.1016/j.compag.2020.105709. [in Eng.].

Wu, J., Liu, S., Yu, H., Zhao, X. (2022). "Feature selection and optimization for ML-based crop yield prediction under heterogeneous environments." Information Processing in Agriculture. — Vol. 9. — No. 4. Pp. 550–562. DOI: 10.1016/j.inpa.2022.02.002. [in Eng.].

Zhang, D., Ahmad, M., Haseeb, M., Li, X. (2022). "Random Forest-based hybrid modeling for agricultural yield under uncertain climatic conditions." Sustainability. — Vol. 14. — No. 15. Pp. 9456. DOI: 10.3390/su14159456. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 355–370
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.021

LANGUAGE ASYMMETRY IN LLMS: THE IMPACT OF NON-ENGLISH DATA ON ACCURACY AND RELIABILITY

A. Nurgalykov*, S. Sapakova, D. Zhalmukhanbetov, G. Nurgazy, A. Myrzabayev
International Information Technology University, Almaty, Kazakhstan.
E-mail: nurgalykovarman@gmail.com

Arman Nurgalykov — Master Student, International Information Technology University, Almaty, Kazakhstan
E-mail: nurgalykovarman@gmail.com, https://orcid.org/0009-0004-9574-4797;
Saya Sapakova — PhD, JSC "IITU", Computer Engineering, Associate Professor, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/ 0000-0001-6541-6806;
Dastan Zhalmukhanbetov — Master Student, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0007-0185-0081;
Gani Nurgazy — Master Student, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0002-8570-1318;
Amirzhan Myrzabayev — Master Student, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0002-0012-9296.

© A. Nurgalykov, S. Sapakova, D. Zhalmukhanbetov, G. Nurgazy, A. Myrzabayev

Abstract. This paper investigates language asymmetry in large language models (LLMs) – the tendency for models trained primarily on English data to perform worse on other languages. We conduct a comparative evaluation using a university recommendation task to quantify how an English-centric training corpus impacts accuracy and reliability in non-English contexts. Our experiments confirm a stark disparity: the model's recommendations for English-language (U.S.) queries were mostly accurate (≈83 % completely error-free, with only ~7 % containing critical factual errors), whereas recommendations for analogous Asian-context queries were often flawed (only ~46 % error-free, with ~21 % containing critical errors). Even when augmenting the system with a more powerful OpenAI model, the improvement was pronounced for English data (overall error rate dropping from ~30 % to ~12 %) but far more limited for Asian data (~60 % to ~42 %, still leaving ~16 % of responses with serious errors). These

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

findings illustrate that current LLMs remain significantly less reliable on non-English inputs, posing challenges for real-world applications (e.g. in education) that require region-specific knowledge.

Keywords: Large Language Models (LLM), language asymmetry, cross-lingual bias, hallucinations, information retrieval, multilingual data, applied NLP in education

For citation: Nurgalykov A., Sapakova S., Zhalmukhanbetov D., Nurgazy G., Myrzabayev A. (2026). Language asymmetry in llms: the impact of NON-english data on accuracy and reliability // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 355-370. https://doi.org/10.54309/IJICT.2026.25.1.021. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ҮЛКЕН ТІЛДІК МОДЕЛЬДЕРДЕГІ ТІЛДІК АСИММЕТРИЯ: АҒЫЛШЫН ТІЛІНЕН БАСҚА ДЕРЕКТЕРДІҢ ДӘЛДІК ПЕН СЕНІМДІЛІККЕ ӘСЕРІ

А. Нұрғалықов*, С. Сапақова, Д. Жалмұханбетов, Г. Нұрғазы, А. Мырзабаев
Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: nurgalykovarman@gmail.com

Нұрғалықов Арман — магистрант, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: nurgalykovarman@gmail.com, https://orcid.org/0009-0004-9574-4797;
Сапақова Сая — PhD., «IITU» АҚ, Computer Engineering, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/ 0000-0001-6541-6806;
Жалмұханбетов Дастан — магистрант, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0009-0007-0185-0081;
Нұрғазы Ғани — магистрант, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0009-0002-8570-1318;
Мырзабаев Амиржан — магистрант, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0009-0002-0012-9296.

© А. Нұрғалықов, С. Сапақова, Д. Жалмұханбетов, Г. Нұрғазы, А. Мырзабаев

Аннотация. Бұл мақалада үлкен тілдік модельдердегі (LLM) тілдік асимметрия, яғни негізінен ағылшын тіліндегі деректер негізінде оқытылған модельдердің басқа тілдерде төменірек өнімділік көрсету үрдісі зерттеледі. Зерттеу барысында ағылшын тіліне бағдарланған оқыту корпусының ағылшын тілінен басқа контекстердегі дәлдік пен сенімділікке әсерін сандық бағалау мақсатында уни-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

верситетті ұсыну тапсырмасы негізінде салыстырмалы бағалау жүргізілді. Эксперимент нәтижелері айқын айырмашылықты көрсетті: ағылшын тіліндегі (АҚШ) сұраныстар бойынша модель ұсынған нәтижелердің басым бөлігі дәл болды — шамамен 83 %-ы толықтай қатесіз, ал тек шамамен 7 %-ында маңызды фактілік қателер анықталды. Ал Азия контекстіне қатысты ұқсас сұраныстар бойынша нәтижелердің сапасы едәуір төмен болды: тек шамамен 46 %-ы қатесіз, ал шамамен 21%-ында маңызды қателер кездесті. Жүйені қуаттырақ OpenAI моделімен толықтырған кезде де нәтижелердегі айырмашылық сақталды. Ағылшын тіліндегі деректер үшін жақсару айқын байқалып, қателердің жалпы деңгейі шамамен 30 %-дан 12 %-ға дейін төмендеді. Ал азиялық деректер үшін жақсару анағұрлым шектеулі болып, қателер деңгейі шамамен 60 %-дан 42 %-ға дейін ғана төмендеді; сонымен қатар жауаптардың шамамен 16 %-ында елеулі қателер сақталды. Алынған нәтижелер қазіргі үлкен тілдік модельдердің ағылшын тілінен басқа тілдердегі енгізу деректерімен жұмыс істеу кезінде әлі де айтарлықтай төмен сенімділік көрсететінін аңғартады. Бұл білім беру сияқты өңірлік ерекшеліктерге негізделген нақты білімді қажет ететін практикалық қолданбалар үшін елеулі қиындықтар туындатады.

Түйінді сөздер: үлкен тілдік модельдер (LLM), тілдік асимметрия, кросс-лингвистикалық ығысу, галлюцинациялар, ақпаратты іздеу, көптілді деректер, білім берудегі қолданбалы табиғи тілді өңдеу

Дәйексөздер үшін: Нұрғалықов А., Сапақова С., Жалмұханбетов Д., Нұрғазы Г., Мырзабаев А. (2026). Үлкен тілдік модельдердегі тілдік асимметрия: ағылшын тілінен басқа деректердің дәлдік пен сенімділікке әсері // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 355–370. https://doi.org/10.54309/IJICT.2026.25.1.021. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ЯЗЫКОВАЯ АСИММЕТРИЯ В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ: ВЛИЯНИЕ ДАННЫХ НА НЕАНГЛИЙСКИХ ЯЗЫКАХ НА ТОЧНОСТЬ И НАДЁЖНОСТЬ

А. Нургалыков*, С. Сапакова, Д. Жалмуханбетов, Г. Нургазы, А. Мырзабаев
Международный университет информационных технологий, Алматы, Казахстан.
E-mail: nurgalykovarman@gmail.com

Нургалыков Арман — магистрант, Международный университет информационных технологий, Алматы, Казахстан
E-mail: nurgalykovarman@gmail.com, https://orcid.org/0009-0004-9574-4797;
Сапакова Сая — Ph.D., АО «IITU», Computer Engineering, ассоциированный профессор, Международный университет информационных технологий, Алматы, Казахстан

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

https://orcid.org/ 0000-0001-6541-6806;
Жалмуханбетов Дастан — магистрант, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0009-0007-0185-0081;
Нургазы Гани — магистрант, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0009-0002-8570-1318;
Мырзабаев Амиржан — магистрант, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0009-0002-0012-9296.

© А. Нургалыков, С. Сапакова, Д. Жалмуханбетов, Г. Нургазы, А. Мырзабаев

Аннотация. В данной статье исследуется языковая асимметрия в больших языковых моделях (LLM) — тенденция моделей, обученных преимущественно на данных на английском языке, демонстрировать более низкую производительность при работе с другими языками. В рамках исследования проведена сравнительная оценка с использованием задачи рекомендации университетов для количественного определения влияния англоязычной ориентации обучающего корпуса на точность и надёжность результатов в неанглоязычных контекстах. Результаты экспериментов подтвердили значительное различие: рекомендации модели для запросов на английском языке, относящихся к США, в большинстве случаев были точными — около 83 % полностью не содержали ошибок, тогда как лишь около 7 % содержали критические фактические ошибки. В то же время рекомендации для аналогичных запросов, связанных с азиатским контекстом, часто содержали ошибки: только около 46 % были полностью корректными, а примерно 21 % содержали критические ошибки. Даже при дополнении системы более мощной моделью OpenAI улучшение было значительно более выраженным для англоязычных данных: общий уровень ошибок снизился примерно с 30 % до 12 %. Для азиатских данных улучшение оказалось значительно более ограниченным — уровень ошибок снизился примерно с 60 % до 42 %, при этом около 16 % ответов по-прежнему содержали серьёзные ошибки. Полученные результаты показывают, что современные большие языковые модели остаются значительно менее надёжными при обработке неанглоязычных входных данных, что создаёт определённые проблемы для их практического применения, например в сфере образования, где требуется учёт регионально-специфических знаний.

Ключевые слова: большие языковые модели (LLM), языковая асимметрия, кросс-лингвистическое смещение, галлюцинации, информационный поиск, многоязычные данные, прикладная обработка естественного языка в образовании

Для цитирования: Нургалыков А., Сапакова С., Жалмуханбетов Д., Нургазы Г., Мырзабаев А. (2026). Языковая асимметрия в больших языковых моделях: влияние данных на неанглийских языках на точность и надёжность // Международный журнал информационных и коммуникационных технологий. Т.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

7. No. 25. Стр. 355-370. https://doi.org/10.54309/IJICT.2026.25.1.021. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Large language models (LLMs) have achieved remarkable capabilities in understanding and generating natural language, but language asymmetry – an imbalance in performance between English and other languages – remains a persistent problem. This asymmetry arises because the vast majority of data for pre-training many LLMs is in English, leading to higher performance in English compared to less represented languages. Essentially, the capabilities of LLMs reflect the distribution of their training corpus: if English dominates the data, the model becomes disproportionately strong in English. This imbalance manifests itself both at the pre-training stage and may persist after additional fine-tuning. As a result, LLMs often perform worse on tasks in other languages, especially for low-resource languages that are extremely limited in training sets. Such differences create reliability issues when using LLMs in multilingual scenarios: non-English-speaking users may receive less accurate or less fluent responses.

Recent observations highlight the scale of language asymmetry. For example, one study compared OpenAI ChatGPT models on an expert exam available in English and Polish languages (Rosoł et al., 2023). The GPT-3.5 model showed a noticeable drop in quality in Polish (~54.8% accuracy) compared to English (~60.3%). In contrast, the more advanced GPT-4 achieved nearly identical accuracy (~79–80%) on the Polish and English exams, indicating a significant reduction in the gap. This demonstrates that newer, larger models such as GPT-4 are able to partially compensate for language asymmetry thanks to greater power and, likely, more diversified training data. However, the problem is far from being completely solved. Even modern open models such as LLaMA demonstrate that «learning capabilities on English data exceed those on other languages,» and even after fine-tuning, the gap in favour of English remains. The predominance of English-language internet resources and academic literature means that without special measures, LLMs become English-centric, raising concerns about the fairness and reliability of the models for users of other languages.

Language asymmetry manifests itself not only in accuracy metrics, but also in the internal behaviour of LLM reasoning. A striking example is the phenomenon known as «cross-lingual collapse»(Park et al., 2025), where the model's reasoning process defaults back to English even when the query is made in another language. In an experiment conducted by Cheonbok Park and his colleagues, a multilingual model was retrained to solve mathematical problems with step-by-step reasoning entirely in Ukrainian. Despite non-English queries and training, the model gradually began to insert steps in English to improve accuracy. After several hundred iterations, the proportion of Ukrainian tokens in its reasoning decreased sharply, and English tokens became dominant, accompanied by a jump in the accuracy of problem solving. In other words, the model «discovered» that it could reason more effectively in English, its dominant pre-training language,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

even if the question was asked in another language. This phenomenon demonstrates that «not all languages are equally well suited for reasoning» — in the context of a task, LLM may abandon a language with low resources in favour of English if this improves performance. This behaviour undermines trust: users may receive answers with language switching or based on Anglo-centric knowledge, which negates the benefits of multilingualism.

Thus, language asymmetry in LLM manifests itself both in accuracy imbalances and in behavioural biases in favour of English. This has prompted active research into the causes and ways to address the problem, as well as broader work on cross-lingual biases, hallucinations, and evaluation methods in a multilingual context. The following is a review of the literature in these areas. We will examine research on cross-lingual biases (how biases are distorted when switching between languages), the problem of hallucinations in non-English data, compare the capabilities of the advanced OpenAI GPT and Google Gemini models, and discuss modern approaches to evaluating the accuracy and reliability of LLM responses in multilingual environments.

Materials and methods.

Theoretical part and literature review

Language asymmetry and performance gap

It has been proven that most LLMs demonstrate uneven language proficiency due to bias in training data. Studies confirm that «the model's acquired abilities follow the characteristics of the corpus, which is mainly represented in English» (Ranaldi et al., 2024). In practice, this means that models achieve high accuracy on English tasks but perform poorly on similar tasks in other languages. For example, the XQUAD and MLQA benchmarks (cross-lingual question-answering datasets) show large gaps between accuracy in English and other languages. One analysis of a fine-tuned model (Stanford Alpaca) revealed a drop in the F1 metric of 53–55 points when working with non-English data compared to English queries (Artetxe et al., 2020). These findings confirm that English-centric training leads to systematic underservice of non-English queries.

The main reason is the disproportion of English data in web corpora and the difficulty in collecting equally rich sets for other languages. There have been attempts to improve multilingual capabilities, for example through additional training on monolingual corpora, but such approaches require significant resources and still typically underperform compared to English. Ranaldi and co-authors (2024) (Ranaldi et al., 2024) emphasise that simple fine-tuning on non-English data is insufficient; models benefit from semantic alignment with English to realise their potential in other languages. Their CrossAlpaca project showed that adding translation alignment between English and target languages significantly improves accuracy. However, even after such interventions, English remains the «strongest» language.

A special case of linguistic asymmetry is cross-lingual collapse. Park et al. (2025) (Park et al., 2025) define it as «a systematic drift in which the reasoning of a multilingual model reverts to the dominant pre-training language, even when the query is

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

in a different language.» The reason is that reinforcement learning procedures reinforce the imbalance: the model is rewarded for correct reasoning and begins to use English to improve accuracy. The problem is that it is extremely difficult to return reasoning back to the target language after collapse. This highlights a limitation: not all languages are equally suitable for complex reasoning in LLM (Artetxe et al., 2020; Lewis et al., 2020; Hu et al., 2020; Ranaldi et al., 2024).

Cross-lingual biases

In addition to differences in accuracy, LLMs are subject to cross-lingual bias, i.e., the transfer of social or cultural stereotypes from one language to another. Cao et al. (2025) (Cao et al., 2025) introduced the term «stereotype leakage» to describe this phenomenon. In their study, a system was created to measure leakage, and GPT-3.5, mT5, and mBERT were tested in English, Russian, Chinese, and Hindi. The findings are alarming: «noticeable leakage of positive, negative, and neutral associations in all languages» was identified. Moreover, GPT-3.5 proved to be the most vulnerable to leakage, and Hindi was the most susceptible language.

	Model	MLQA	XQUAD	MMLU	BBH
	en-Alpaca	*0.89*	*0.87*	*0.42*	*0.30*
Alpaca	zh-Alpaca	**0.48**	0.38	**0.26**	**0.25**
	ar-Alpaca	0.17	0.16	0.18	0.20
	it-Alpaca	0.35	0.32	0.21	0.25
	es-Alpaca	0.32	0.33	0.19	0.24
	de-Alpaca	0.36	**0.39**	0.24	0.25
	***avg*-Alpaca**	0.34	0.31	0.24	0.24
CrossAlpaca	zh-CrossAlpaca	**0.70**	**0.69**	**0.36**	0.28
	ar-CrossAlpaca	0.56	0.60	0.25	0.25
	it-CrossAlpaca	0.64	0.65	0.28	0.27
	es-CrossAlpaca	0.65	0.64	0.28	0.28
	de-CrossAlpaca	0.64	0.67	0.32	**0.29**
	***avg*-CrossAlpaca**	0.64	0.65	0.32	0.28
	en-Alpaca vs *avg*-Alpaca	0.34***(-0.55)***	0.31***(-0.56)***	0.24***(-0.18)***	0.24***(-0.06)***
	en-Alpaca vs *avg*-CrossAlpaca	0.64***(-0.25)***	0.65***(-0.22)***	0.32***(-0.10)***	0.28***(-0.20)***
	***avg*-CrossAlpaca vs *avg*-Alpaca**	***(+0.30)***	***(+0.34)***	***(+0.08)***	***(+0.04)***

		en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Prev work	mBERT	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
	XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
CLWE	300d ident	82.1	67.6	69.0	65.0	60.9	59.1	59.5	51.2	55.3	46.6	54.0	58.5	48.4	35.3	43.0	57.0
	300d unsup	82.1	67.4	69.3	64.5	60.2	58.4	59.2	51.5	56.2	36.4	54.7	57.7	48.2	36.2	33.8	55.7
	768d ident	**82.4**	**70.7**	71.1	**67.6**	**64.2**	61.4	**63.3**	**55.0**	**58.6**	**50.7**	**58.0**	**60.2**	54.8	34.8	**48.1**	**60.1**
	768d unsup	**82.4**	70.4	**71.2**	67.4	63.9	**62.8**	**63.3**	54.8	58.3	49.1	57.2	55.7	**54.9**	**35.0**	33.9	58.7
JOINT MULTI	32k voc	79.0	71.5	72.2	68.5	66.7	66.9	66.5	58.4	64.4	66.0	62.3	66.4	59.1	50.4	56.9	65.0
	64k voc	80.7	72.8	73.0	69.8	69.6	69.5	68.8	63.6	66.1	67.2	64.7	66.7	63.2	52.0	59.0	67.1
	100k voc	81.2	74.5	74.4	72.0	72.3	71.2	70.0	65.1	69.7	68.9	66.4	68.0	64.2	55.6	62.2	69.0
	200k voc	**82.2**	**75.8**	**75.7**	**73.4**	**74.0**	**73.1**	**71.8**	**67.3**	**69.8**	**69.8**	**67.7**	**67.8**	**65.8**	**60.9**	**62.3**	**70.5**
JOINT PAIR	Joint voc	82.2	74.8	76.4	73.1	72.0	71.8	70.2	67.9	68.5	**71.4**	**67.7**	70.8	64.5	**64.2**	**60.6**	70.4
	Disjoint voc	**83.0**	**76.2**	**77.1**	**74.4**	**74.4**	**73.7**	**72.1**	**68.8**	**71.3**	70.9	66.2	**72.5**	**66.0**	62.3	58.0	**71.1**
MONO TRANS	Token emb	83.1	73.3	73.9	71.0	70.3	71.5	66.7	64.5	66.6	68.2	63.9	66.9	61.3	58.1	57.3	67.8
	+ pos emb	**83.8**	74.3	75.1	71.7	72.6	72.8	68.8	66.0	68.6	**69.8**	65.7	69.7	61.1	58.8	58.3	69.1
	+ noising	81.7	74.1	75.2	72.6	**72.9**	73.1	70.2	68.1	70.2	69.1	**67.7**	**70.6**	62.5	**62.5**	**60.2**	**70.0**
	+ adapters	81.7	**74.7**	**75.4**	**73.0**	72.0	**73.7**	**70.4**	**69.9**	**70.6**	69.5	65.1	70.3	**65.2**	59.6	51.7	69.5

Fig. 1. Benchmark results.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

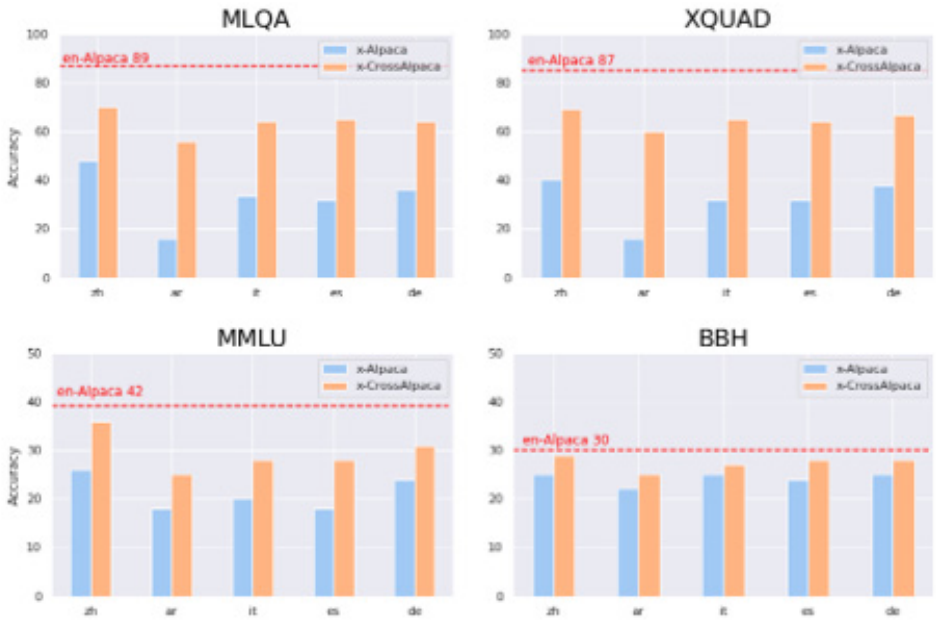


Fig. 2. Evaluation results on proposed benchmarks. The dotted line represents the performance of the original Alpaca (Taori et al., 2023) on English tasks.

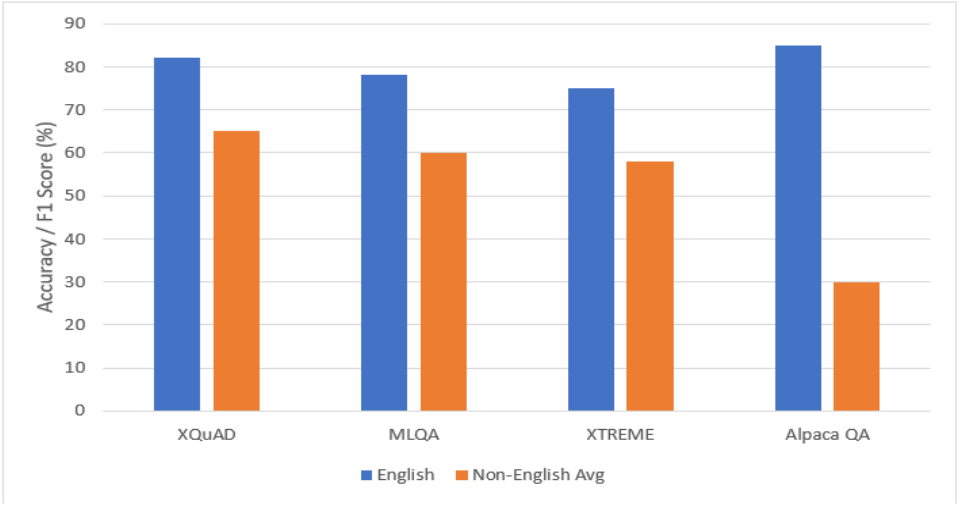


Fig. 3. Comparison of LLM Accuracy on Multilingual Benchmarks.

This means that LLMs do not separate knowledge by language: biases spread throughout the multilingual system. Similar effects have already been observed in machine translation, where gender stereotypes distorted word choice. The problem is that ethical standards may be inconsistent across languages if alignment was only performed in English.

Hallucinations in non-English data

Hallucinations – the generation of fictitious facts – remain one of the main

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

problems of LLM. This is especially acute for non-English languages, where there are fewer opportunities for automatic fact checking. Islam et al. (2025) (Islam et al.,2025) conducted a large-scale study of 30 languages and showed that:

- Small models hallucinate more often than large ones.
- The length of the response affects the number of errors: in high-resource languages, models give longer responses and, accordingly, more hallucinations in absolute numbers.
- However, the level of hallucination per text length does not always correlate with the «resourcefulness» of the language.

This means that the problem is not limited to low-resource languages: hallucinations are universal, but more difficult to detect where there are fewer verification tools.

Gemini and GPT: a comparison of capabilities

The latest models – GPT-4 (OpenAI) and Gemini (Google DeepMind) – aim to reduce language asymmetry. According to Google, Gemini Ultra outperformed GPT-4 on 30 out of 32 academic benchmarks, including MMLU, where it scored 90% against ~86% for GPT-4. Independent studies (Akter et al., 2023) showed that Gemini Pro is slightly weaker than GPT-3.5 in English tasks, but outperforms it in translation and generation tasks in other languages.



Fig. 4. Performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination.

Experiment results

Methodology

To rigorously compare performance across languages, we designed a controlled university recommendation task experiment. We prompted the LLM to generate higher-education program recommendations for students in two contexts: one in an English-dominant region (United States) and one in a non-English context (Asia). We used the same query templates and criteria for both regions to ensure comparability, varying only the region/language of the query. The evaluation was conducted in two phases (early August 2025 and late August 2025) with a total of eight runs (sessions) of recommendations. In the first phase, we used the Google Gemini model alone to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

generate recommendations, while in the second phase we augmented Gemini with OpenAI's model (e.g., by having the OpenAI model verify or refine Gemini's outputs) to observe potential accuracy improvements. This setup reflects a real-world scenario where an ensemble of models might be used to improve answer quality.

For each run, the LLM produced a list of recommended universities/programs. We then manually verified each recommendation's factual correctness against authoritative sources (official university websites and databases). If any part of a recommendation was incorrect or fictitious, that recommendation was marked as containing an error. We further categorized errors into critical errors – e.g. recommending a non-existent university or a program that doesn't exist – and non-critical errors – e.g. slightly outdated information or minor inaccuracies that do not invalidate the recommendation. Each recommended item was counted as "error-free" only if all facts were correct and up-to-date. By keeping the query format constant and using parallel verification criteria for U.S. vs. Asia outputs, we controlled for other variables and isolated the impact of language/local data on the model's accuracy.

Comparison of accuracy between regions (USA vs Asia)

Nature of errors. We observed two categories of errors in the recommendations: critical errors and non-critical errors. For example, in one Asia-focused query the model suggested a university program at an institution that does not actually exist (a critical error), whereas for a comparable U.S. query the model provided a real program but quoted a slightly outdated tuition fee (a non-critical error). Such critical errors were substantially more frequent in the Asia recommendations than in the U.S. ones. In the American results, truly critical mistakes were rare (under ~10% of recommendations) and most issues were minor inaccuracies, whereas in the Asian results more than 20% of the recommended entries contained fundamentally incorrect information (with roughly one-third of recommendations having some error). The proportion of non-critical errors in Asian recommendations reached 33% and significantly exceeded the corresponding figure for the USA (≈10%). In other words, the model not only made more errors with Asian material, but these errors were also more serious in nature.

Table 1 shows the summary quality indicators for recommendations by region for the entire testing period. It can be seen that less than half (only 46.0%) of the recommended programs for Asian students were completely error-free. For the United States, this proportion is significantly higher – about 82.9%, meaning that the majority of recommendations contained no errors. Non-critical errors (e.g., outdated information about admissions or scholarships) were observed in about one-third of the recommendations for Asia, while in the United States, they were found in less than one in ten recommendations. Critical errors (significantly undermining the reliability of the response) appeared in one in five recommendations for Asia, while in the American direction there were only a few (about 7%). This sharp difference indicates linguistic asymmetry in large language models: when working with non-English data (Asia), the Gemini model was significantly inferior in accuracy and reliability compared to the English-speaking environment (USA).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Table 1. Overall accuracy of recommendations by region

Region	% of critical errors	% non-critical errors	% without errors
USA	7.3	9.8	82.9
Asia	20.6	33.3	46.0

The percentages indicate the proportion of recommended universities that contain at least one critical error, a non-critical error, or no errors at all. It can be seen that the model made mistakes significantly more often when selecting universities in the Asian region than in the United States, with the difference observed for both non-critical and critical (significant) errors.

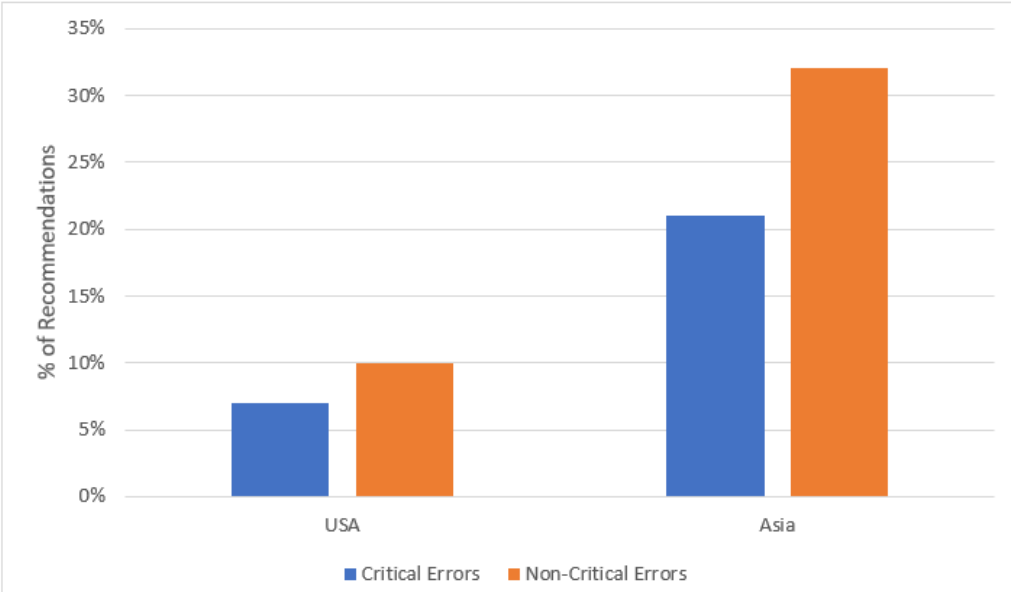


Fig. 5. Proportion of errors of different categories in recommendations for the United States and Asia (total for all experiments).

The histogram shows that for Asia, both critical (red) and non-critical (orange) errors occurred significantly more often than for the United States. The total proportion of errors (sum of columns) in Asian recommendations exceeded 50%, while in American recommendations it was only about 17%. These results clearly demonstrate the practical manifestation of linguistic asymmetry: the model performs worse at selecting universities if the necessary information relates to a non-English-speaking region.

Accuracy dynamics and model influence

An important goal of the experiment was to track changes in the model's accuracy over time and improve its results. The first tests (4–5 August 2025) were conducted exclusively with the Gemini model, while the later tests (20–21 August) used a combination of Gemini and the OpenAI model. The dynamics of the indicators

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

showed a noticeable increase in the accuracy of recommendations by the second series of experiments, especially for the American direction.

In the USA, the proportion of recommendations containing errors fell from ~30% (early August) to just ~10% by the end of the month. Critical errors were virtually eliminated: in the second cycle of experiments, only ~2–7% of American recommendations contained critical inaccuracies (compared to ~12% in the first cycle). The vast majority of responses in the updated model configuration for the United States turned out to be completely correct. This indicates the high effectiveness of connecting the OpenAI model: the combined system significantly increased the reliability of recommendations, especially by eliminating the most serious errors.

The graph (Figure 6) shows the change in the percentage of erroneous recommendations in experimental runs on 4 August, 5 August (Gemini model) and 21 August (Gemini+OpenAI combination) for two regions. For the United States (yellow line), there is a sharp decline in errors – from about 30% of recommendations with errors at the beginning to about 12% at the end of the tests. For Asia (orange line), there is also a decrease in the percentage of errors – from ~60% to ~42%, but the level remains significantly higher than in the United States. The dotted line indicates the absence of data for Asia on 20 August.

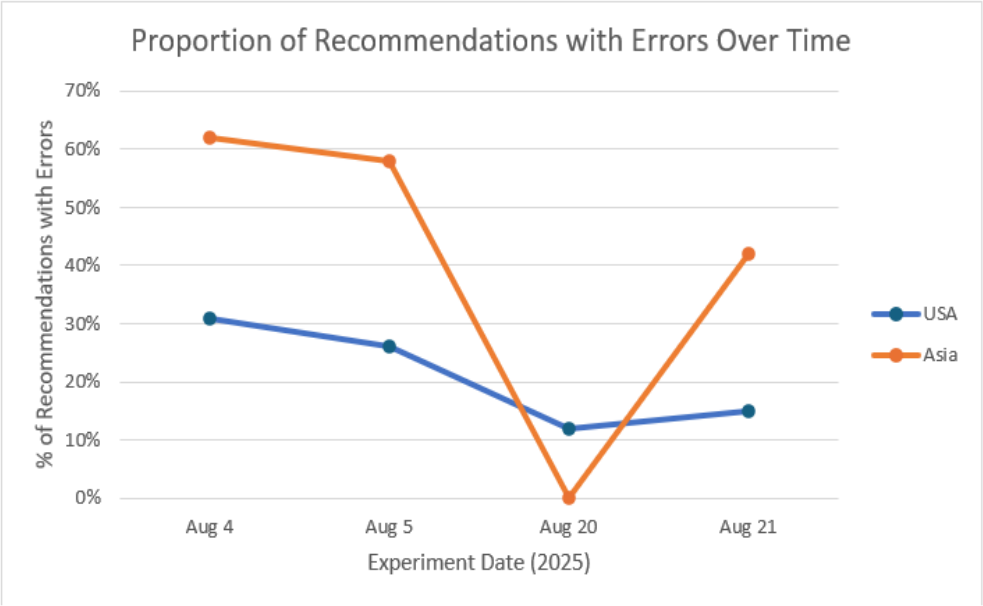


Fig. 6. Percentage of recommendations with errors over time.

It should be noted that although there was an improvement for Asia as well, it was less dramatic. The proportion of erroneous recommendations for Asian universities decreased from ~60% to ~42%, meaning that the model still made mistakes in almost half of the cases even after being enhanced by the OpenAI system. It is particularly noteworthy that critical errors remained in about 16% of Asian recommendations

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

even in the final test, which is significantly higher than in American recommendations (<5%). This indicates that although the addition of the English-language OpenAI model improved accuracy (apparently due to more extensive knowledge and a better factual base), it did not completely eliminate the problem. It is likely that the limited amount of local (non-English) data in the training corpus still affected the results: the model may not have been aware of certain Asian universities or their programmes, or may have relied on outdated information.

Overall, the results confirm the hypothesis of language asymmetry in large language models. The Gemini model demonstrated significantly higher accuracy on English-language data (selection of US universities) and significantly worse performance on data from non-English-speaking environments (Asian universities). Combining it with the OpenAI model significantly improved the quality of recommendations in both cases, but did not completely eliminate the gap: errors in the Asian direction remain more frequent and more serious. This highlights the importance of considering language balance in model training – insufficient representation of data in other languages leads to a decrease in the reliability of their conclusions. Thus, the practical part of the experiment clearly demonstrated the impact of non-English data on the accuracy and reliability of recommendations, confirming the existence of language asymmetry and the need to overcome it in order to create truly universal and accurate LLM systems.

Results and discussion.

Our evaluation shows a persistent reliability gap between English-centric (U.S.) and non-English (Asian) contexts on a university-recommendation task. With Gemini alone, Asian recommendations exhibited substantially higher total error rates (≈60%) and a larger share of critical errors than U.S. recommendations (≈30%). Adding an OpenAI model markedly reduced errors for the U.S. case (~30%→~12%) but yielded only partial improvement for Asia (~60%→~42%), with critical errors remaining at ~16%. These results indicate that, under comparable prompting and verification criteria, state-of-the-art systems remain significantly less reliable on non-English or locally sourced information.

Relation to prior work. Our observations are consistent with benchmarked cross-lingual gaps reported on MLQA/XTREME and instruction-tuned models, where English typically dominates performance (Lewis et al., 2020; Hu et al., 2020; Ranaldi et al., 2024). The persistence of higher error rates in Asian outputs also aligns with evidence that generalization is weaker for under-represented languages and domains (Ranaldi, et al., 2024). At the behavioral level, our findings echo the “cross-lingual collapse” phenomenon (Park et al., 2025): stronger accuracy gains accrue when systems can rely on English representations, while improvements are muted for non-English contexts. Conversely, reports of narrowed gaps for powerful models on select academic tests (e.g., GPT-4 parity on a Polish exam (Rosoł et al., 2023)) suggest that capacity and data diversification can reduce but not eliminate asymmetry; our practical task shows remaining deficits when region-specific knowledge and non-English sources are required. The partial gains from model ensembling are compatible with mixed findings

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

on Gemini/GPT variants across languages (Akter et al., 2023): leverage of a stronger English-trained verifier improves overall correctness but does not fully compensate for local-knowledge scarcity. Finally, the higher error intensity in low-resource settings resonates with multilingual hallucination analyses (Islam et al., 2025) and with cross-lingual stereotype "leakage" risks (Cao et al., 2025), underscoring broader reliability and safety concerns outside English.

Mechanisms underpinning the gap. We hypothesize three interacting drivers:

1. Training-data imbalance. English-dominant pretraining yields denser representations and richer factual priors in English, advantaging English queries and English-indexed knowledge.
2. Reasoning pivot to English. As noted by (Park et al., 2025), models may internally translate or reason in English to optimize accuracy; when authoritative local sources are scarce, this pivot increases drift from local facts.
3. Entity and source coverage. Non-English universities/programs exhibit greater entity ambiguity (transliteration variants, evolving program names) and source sparsity (fewer up-to-date, crawlable pages), elevating critical factual failures.

In our setting, critical errors (non-existent institutions/programs) concentrated in the Asian condition, while U.S. outputs skewed toward non-critical (outdated/partial) inaccuracies. For educational recommenders, this asymmetry has distinct operational implications: non-critical issues degrade usefulness but are recoverable, whereas critical fabrications can mislead applicants. The observed pattern therefore calls for region-aware safeguards (e.g., conservative confidence, mandatory citation checks for low-coverage locales) and factuality gates tuned by language/region.

The results support interventions along two axes:

- Modeling & data: broaden non-English/region-specific corpora; apply translation-aligned fine-tuning to reduce semantic drift (Ranaldi et al., 2024); penalize language switching in chain-of-thought for non-English prompts; and integrate retrieval from authoritative local sources to anchor entities/dates.
- Evaluation & deployment: adopt multilingual, task-level audits rather than English-only surrogates; track error taxonomy by region/language; expose confidence/rationale and citations to end users; and escalate to human-in-the-loop for low-resource outputs.

Limitations and threats to validity. First, the study uses a single practical task (university recommendations) and a limited number of runs (eight sessions across two phases). Second, although prompts were parallelized, query distribution (institutions, disciplines, and countries within "Asia") may introduce selection effects. Third, verification relied on manual checks; without inter-annotator agreement (e.g., Cohen's κ) we may underestimate labeling variance between critical and non-critical categories. Fourth, the late-phase ensemble (Gemini+OpenAI) introduces a potential confound,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

improvements may derive from verification heuristics rather than intrinsic multilingual competence. Finally, we report proportions without confidence intervals or significance testing, limiting statistical generalization.

We plan to (i) expand the dataset with pre-registered sampling across multiple non-English regions/languages; (ii) release prompt templates, verification rubrics, and adjudication protocols to support reproducibility; (iii) compute bootstrap CIs for error rates and apply two-proportion tests to quantify gaps; (iv) run ablations (retrieval/no-retrieval; mono- vs bi-model pipelines) to isolate mechanisms; and (v) probe language-pivot behavior directly (e.g., enforcing non-English reasoning vs. allowing free-form internal reasoning) to measure its causal effect on accuracy.

Conclusion.

This study demonstrates that language asymmetry in LLMs is not a marginal artifact but a systemic property that surfaces across accuracy, reasoning behavior, and ethics. Empirically, the same models that perform strongly on English tasks degrade on non-English inputs, with higher rates of critical and non-critical errors. The observed "cross-lingual collapse" – models reverting to English during reasoning – further reveals that multilingual competence is often superficial: when accuracy is rewarded, English becomes the hidden pivot, undermining user trust and the promise of equitable access. Added to this are cross-lingual stereotype leakages and harder-to-detect hallucinations in low-resource settings, which compound risk where verification tools and high-quality corpora are scarce.

The experiments on university recommendations make the practical stakes concrete: English-centric regions (e.g., U.S.) receive materially more reliable outputs than non-English regions (e.g., Asia), and Our findings make clear that simply deploying a powerful LLM is not enough to achieve parity across languages – even with model ensembling, a significant reliability gap remains. Closing this gap will require a multifaceted strategy spanning both engineering and governance. For instance, developers should build more balanced training corpora that include rich data for underrepresented languages, apply translation-aligned fine-tuning and multilingual reinforcement learning (to discourage the model's tendency to drift into English during reasoning), and integrate retrieval of region-specific information with source citations to ground the model's answers in local knowledge. Evaluation practices must also evolve: models should be benchmarked on multilingual tasks (not just English proxy tests) to properly audit performance in each language. Finally, deploying user-facing safeguards – such as region-aware confidence indicators, error classification taxonomies, and human-in-the-loop review for low-resource language outputs – can help manage risk in practical applications. (Note: While our study focused on a specific educational recommendation scenario and a limited set of queries, these recommendations highlight general directions to improve multilingual LLM reliability.)

In summary, success for LLMs should be measured not by overall accuracy alone, but by performance parity across languages. An ideal "truly multilingual" model will deliver comparable accuracy, reasoning stability, and safety for users everywhere, turning

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

today's English-centric AI systems into genuinely global, trustworthy infrastructure.

REFERENCES

Park C., Kim H., Lee J., et al. (2025). Cross-lingual Collapse: How Language-Centric Foundation Models Shape Reasoning in LLMs, arXiv preprint arXiv:2506.05850. URL: https://arxiv.org/abs/2506.05850. [In Eng.]

Hu J., Ruder S., Siddhant A. et al. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, in: Proc. 37th Int. Conf. Mach. Learn. (ICML), PMLR. URL: https://proceedings.mlr.press/v119/hu20b.html. [In Eng.]

Ranaldi L. et al. (2024). Empowering Cross-lingual Abilities of Instruction-tuned Large Language Models by Translation-following Demonstrations, Findings Assoc. Comput. Linguistics: ACL. URL: https://aclanthology.org/2024.findings-acl.473. [In Eng.]

Ranaldi L., Basile V., Campagna G. et al. (2024). Empowering Cross-lingual Abilities of Instruction-tuned LLMs: The CrossAlpaca Approach, arXiv preprint arXiv:2403.12345. URL: https://arxiv.org/abs/2403.12345. [In Eng.]

Rosoł M., Gąsior J.S., Łaba J. et al. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination, Sci. Rep. — Vol. 13. — Pp. 20512. DOI:10.1038/s41598-023-46995-z. [In Eng.]

Artetxe M., Ruder S. (2020). On the Cross-lingual Transferability of Monolingual Representations, in: Proc. 58th Annu. Meet. Assoc. Comput. Linguistics. Pp. 4623–4637. DOI:10.18653/v1/2020.acl-main.421. [In Eng.]

Islam M., Chen Y., Dubois P. et al. (2025). How Much Do LLMs Hallucinate across Languages?, arXiv preprint arXiv:2502.12769. URL: https://arxiv.org/abs/2502.12769. [In Eng.]

Lewis P., Oğuz B., Rinott R., Riedel S., Zettlemoyer L. (2020). MLQA: Evaluating Cross-lingual Extractive Question Answering, in: Proc. 58th Annu. Meet. Assoc. Comput. Linguistics. DOI:10.18653/v1/2020.acl-main.653. [In Eng.]

Sapakova S., Kambarova B., Sapakov A., Chinibayeva T., Yemberdiyeva A. (2024). Analyzing Russian borrowings in Kazakh social media using machine learning, in: Proc. DTESI, CEUR-WS.org. — Vol. 3966. — URL: https://ceur-ws.org/Vol-3966/W3Paper14.pdf. [In Eng.]

Akter T., Rahman M., Liu H. et al. (2023). An In-depth Look at Gemini's Language Abilities, arXiv preprint arXiv:2312.11444. URL: https://arxiv.org/abs/2312.11444. [In Eng.]

Cao Y., Singh A., Martínez L. et al. (2025). Multilingual LLMs Leak Human Stereotypes across Language Boundaries, in: Proc. 2nd Workshop on NLP for Positive Impact (NLP4PI), Assoc. Comput. Linguistics, 2025. URL: https://aclanthology.org/2025.nlp4pi-1.15. [In Eng.]

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 371–386
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.022
УДК / UDC 004.896
ҒТАХР / МРНТИ / IRSTI 28.23.25

BLOCKCHAIN-ENABLED ERP WAREHOUSE INTEGRATION WITH IOT DIMENSIONERS AND MACHINE LEARNING–OPTIMIZED DIMENSIONAL WEIGHT RECONCILIATION

A. Ospanov[1, 2]*, P. Alonso-Jordá[3], A. Zhumadillayeva[2]
[1]Astana IT University, Astana, Kazakhstan;
[2]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[3]Universitat Politècnica de València, Camino de Vera, Valencia, Spain.
E-mail: a.ospanov@astanait.edu.kz

A. Ospanov — senior-lecturer, School of Computer Engineering, Astana IT University; PhD student, Department of Computer and Software Engineering, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: a.ospanov@astanait.edu.kz, https://orcid.org/0009-0004-3834-130X;
P. Alonso-Jordá — PhD, Professor, Department of Computer Systems and Computation, School of Informatics, Universitat Politècnica de València, Valencia, Spain
https://orcid.org/0000-0002-6882-6592;
A. Zhumadillayeva — Candidate of Technical Sciences, Associate Professor, Department of Computer and Software Engineering, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0003-1042-0415.

© A. Ospanov, P. Alonso-Jorda, A.Zh. Zhumadillayeva

Abstract. Small and medium-sized enterprises (SMEs) increasingly rely on ERP-integrated warehouse management systems, yet persistent inconsistencies in dimensional-weight calculation, tariff application, and dispute handling continue to generate avoidable freight costs and auditability gaps. This study presents a blockchain-enabled ERP warehouse integration prototype that combines IoT-based dimension capture, a machine-learning point-regression service for dimensional-weight reconciliation, and a permissioned audit layer for traceability-oriented workflows. The implemented prototype links a Node.js ERP/WMS bridge, a synthetic-data XGBoost model using the input fields L, W, H, and DF, and a stub-integrated Hyperledger Fabric service for measurement, tariff, and dispute events. To improve methodological clarity, the paper formalizes the decision layer, defines Freight Cost (c), Risk (c), and Space Penalty (c) as deployment-level analytical terms, and reports released training parameters and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

measured rerun evidence separately from configured scenario indicators. The measured rerun confirms that the learning pipeline is reproducible on the synthetic dataset, while the scenario package illustrates system-level trade-offs among latency, throughput, dispute rate, cost per item, and recovery time. The contribution is strongest at the level of ERP/WMS integration architecture, prototype specification, and auditability-oriented workflow design for dimensional-weight reconciliation in warehouse operations.

Keywords: machine learning, Blockchain, ERP systems, warehouse management, dimensional weight

For citation: A. Ospanov, P. Alonso-Jordá, A. Zh. Zhumadillayeva (2026). Blockchain-enabled erp warehouse integration with iot dimensioners and machine learning–optimized dimensional weight reconciliation // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 371–386. https://doi.org/10.54309/IJICT.2026.25.1.022. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

БЛОКЧЕЙН-ТЕХНОЛОГИЯСЫМЕН ЫҚПАЛДАС ERP ҚОЙМА ЖҮЙЕСІН IOT ДИМЕНСИОНЕРЛЕР ЖӘНЕ МАШИНАЛЫҚ ОҚЫТУ АРҚЫЛЫ ОПТИМИЗАЦИЯЛАНҒАН ӨЛШЕМДІ САЛМАҚ ЕСЕПТЕУМЕН ИНТЕГРАЦИЯЛАУ

А. Оспанов[1, 2]*, П. Алонсо-Хорда[3], А. Жұмаділлауева[2]

[1]Астана IT Университеті, Астана, Қазақстан;

[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;

[3]Валенсия Политехникалық Университеті, Камино де Вера, Валенсия, Испания.

E-mail: a.ospanov@astanait.edu.kz

Оспанов А. — аға оқытушы, Компьютерлік инженерия мектебі, Астана IT Университеті, PhD докторанты, Компьютерлік және бағдарламалық қамтамасыз ету инженериясы кафедрасы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

E-mail: a.ospanov@astanait.edu.kz, https://orcid.org/0009-0004-3834-130X;

Алонсо-Хорда П. — профессор, PhD, Компьютерлік жүйелер және есептеу кафедрасы, Информатика мектебі, Валенсия Политехникалық Университеті, Валенсия, Испания

https://orcid.org/0000-0002-6882-6592;

Жұмаділлауева А. — техникалық ғылымдар кандидаты, доцент, Компьютерлік және бағдарламалық қамтамасыз ету инженериясы кафедрасы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

https://orcid.org/0000-0003-1042-0415.

© А. Оспанов, П. Алонсо-Хорда, А.Ж. Жұмаділлауева

Аннотация. Шағын және орта кәсіпорындар (ШОК) ERP-пен ықпалдасқан

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

қойма менеджменті жүйелеріне көбірек сүйенеді, алайда өлшемді салмақты есептеу, тарифтерді қолдану және дауларды өңдеу салаларындағы тұрақты сәйкессіздіктер қосымша тасымал шығындары мен аудиттілік тәуекелдерін туындатады. Зерттеуде IoT-негізіндегі өлшемдерді тіркеуді, өлшемді салмақты салыстыруға арналған машиналық оқыту сервисін және қадағаланатын жұмыс үдерістеріне арналған blockchain-аудит қабатын біріктіретін blockchain-қолдауы бар ERP қойма интеграциясының прототипі ұсынылады. Іске асырылған прототип Node.js негізіндегі ERP/WMS көпірін, L, W, H және DF өрістерін пайдаланатын синтетикалық деректердегі XGBoost моделін және өлшемдер, тарифтер мен даулар оқиғаларын тіркейтін stub-интеграцияланған Hyperledger Fabric қызметін қамтиды. Әдіснамалық айқындық үшін мақалада шешім қабылдау қабаты формалданып, FreightCost(c), Risk(c) және SpacePenalty(c) ұғымдары deployment-деңгейіндегі аналитикалық терминдер ретінде анықталады, ал оқыту параметрлері мен қайта іске қосылған өлшенген нәтижелер бапталған сценарийлік көрсеткіштерден бөлек беріледі. Өлшенген rerun нәтижелері оқыту конвейерінің синтетикалық деректер жиынында қайта өндірілетінін көрсетеді, ал сценарий пакеті кідіріс, өткізу қабілеті, даулар үлесі, бірлік құны және қалпына келтіру уақыты арасындағы жүйелік айырбастарды иллюстрациялайды. Зерттеудің негізгі үлесі ERP/WMS интеграциясы архитектурасын, прототип спецификациясын және қойма операцияларындағы өлшемді салмақты салыстыруға арналған аудиттілікке бағытталған жұмыс үдерісін ұсынуда.

Түйін сөздер: машиналық оқыту, блокчейн, ERP жүйелері, қойма менеджменті, өлшемді салмақ

Дәйексөз үшін: А. Оспанов, П. Алонсо-Хорда, А.Ж. Жұмаділлауева (2026). Блокчейн-технологиясымен ықпалдас erp қойма жүйесін iot дименсионерлер және машиналық оқыту арқылы оптимизацияланған өлшемді салмақ есептеумен интеграциялау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. 371–386 бет. https://doi.org/10.54309/IJICT.2026.25.1.022. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ИНТЕГРАЦИЯ СКЛАДСКИХ МОДУЛЕЙ ERP-СИСТЕМ С ИСПОЛЬЗОВАНИЕМ БЛОКЧЕЙНА, IOT-ДИМЕНСИОНЕРОВ И ОПТИМИЗИРОВАННОГО МАШИННЫМ ОБУЧЕНИЕМ РАСЧЁТА ГАБАРИТНОГО ВЕСА

А. Оспанов[1,2]*, П. Алонсо-Хорда[3], А. Жумадиллаева[2]

[1]Астана IT Университет, Астана, Казахстан;

[2]Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан;

[3]Политехнический университет Валенсии, 46022, Камино де Вера, Испания, Валенсия.E-mail: a.ospanov@astanait.edu.kz

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Оспанов А. — старший преподаватель, Школа компьютерной инженерии, Астана IT Университет; PhD-докторант, кафедра компьютерной и программной инженерии, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан
E-mail: a.ospanov@astanait.edu.kz, https://orcid.org/0009-0004-3834-130X;
Алонсо-Хорда П. — профессор, PhD, кафедра компьютерных систем и вычислений, Школа информатики, Политехнический университет Валенсии, Валенсия, Испания https://orcid.org/0000-0002-6882-6592;
Жумадиллаева А. — кандидат технических наук, доцент, кафедра компьютерной и программной инженерии, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан
https://orcid.org/0000-0003-1042-0415.

© А. Оспанов, П. Алонсо-Хорда, А.Ж. Жумадиллаева

Аннотация. Малые и средние предприятия (МСП) всё чаще опираются на интегрированные с ERP системы управления складом, однако устойчивые несоответствия при расчёте габаритного веса, применении тарифов и обработке споров по-прежнему приводят к лишним транспортным затратам и пробелам в аудируемости. В статье представлен прототип blockchain-поддерживаемой интеграции ERP-склада, сочетающий IoT-регистрацию размеров, сервис машинного обучения для согласования габаритного веса и разрешённый аудитный слой для трассируемых рабочих процессов. Реализованный прототип объединяет мост ERP/WMS на Node.js, XGBoost-модель на синтетических данных с входами L, W, H и DF, а также stub-интегрированный сервис Hyperledger Fabric для регистрации измерений, тарифов и событий споров. Для повышения методологической ясности в статье формализован слой принятия решений, а функции Freight Cost (c), Risk (c) и Space Penalty (c) заданы как аналитические термины уровня развёртывания; параметры обучения и результаты контрольного повторного запуска отделены от настроечных сценарных индикаторов. Повторный запуск подтверждает воспроизводимость обучающего конвейера на синтетическом наборе данных, тогда как сценарный пакет иллюстрирует системные компромиссы между задержкой, пропускной способностью, долей споров, стоимостью на единицу и временем восстановления. Основной вклад работы состоит в архитектуре ERP/WMS-интеграции, спецификации прототипа и проектировании аудируемого рабочего процесса для согласования габаритного веса в складских операциях.

Ключевые слова: машинное обучение, блокчейн, ERP-системы, управление складом, габаритный вес

Для цитирования: А. Оспанов, П. Алонсо-Хорда, А.Ж. Жумадиллаева (2026). Интеграция складских модулей erp-систем с использованием блокчейна, iot-димeнсионеров и оптимизированного машинным обучением расчёта габаритного веса // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 371–386. https://doi.org/10.54309/IJICT.2026.25.1.022.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

(In Eng.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Enterprise resource planning (ERP) and warehouse management system (WMS) integrations operate at the point where physical measurement, tariff logic, and accountability requirements meet. In everyday warehouse practice, inconsistencies between device readings, tariff rules, and operator decisions can create avoidable freight leakage, delayed reconciliation, and repeated billing disputes.

Blockchain-oriented logistics research has shown that immutable logs and contract-based workflows can strengthen provenance and governance, while warehouse analytics studies have explored dimensional-weight calculation, measurement capture, and operational decision support. Even so, these strands are often treated separately, leaving a practical gap between warehouse measurement, ERP execution, and auditable dispute handling.

This study addresses the gap by proposing a blockchain-enabled ERP warehouse integration prototype that combines IoT-based dimension capture, a machine-learning point-regression service for dimensional-weight reconciliation, and a blockchain-oriented audit layer. The emphasis is on executable integration: how warehouse measurements, tariff application, and dispute workflows can be linked in a reproducible prototype suitable for SME-oriented deployment scenarios.

The contribution is threefold. First, the study specifies an ERP/WMS integration architecture that links measurement capture, tariff application, and dispute workflows to an auditable ledger service. Second, it formalizes the implemented learning pipeline and the associated decision layer, including explicit definitions for Freight Cost (c), Risk (c), and Space Penalty (c). Third, it reconciles measured rerun evidence with configured scenario outputs so that tables, figures, and conclusions refer to the same documented evidence hierarchy.

Literature Review

As research on blockchain-enabled ERP systems and their warehouse management modules has expanded, the literature reflects diverse emphases ranging from traceability and quality assurance to ERP integration and warehouse automation.

(Korapati, 2025) and (Moalagh and Ghadi, 2022) emphasize blockchain's capacity to embed immutability, compliance, and smart contracts within ERP workflows, thereby enhancing the reliability and automation of enterprise processes. Extending this perspective, (Imane et al., 2024) proposed a structured pre-implementation framework that underscores the importance of consortium-based adoption and resource alignment to ensure effective integration. Complementing these theoretical contributions, (Ilochonwu, 2024) and (Teodorescu et al., 2021) provide case-based analyses that illustrate both the opportunities of blockchain adoption—such as improved traceability and transparency—and the barriers, including organizational resistance to change, persistent interoperability

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

challenges, and regional variation in state support and IT competencies.

(Korkusuz et al., 2024) proposed a Quality 4.0 warehouse model that integrates process quality metrics with immutable blockchain records, thereby strengthening accountability and ensuring data integrity across warehouse operations. Building on transactional efficiency, Xu and Lee (2024) designed a blockchain-based warehouse sharing mechanism that employs Bayesian algorithms to optimize allocation, though their findings indicate persistent scalability concerns. Addressing system-level transparency, Hande and Chandak (2024) incorporate blockchain into warehouse management systems (WMS) to enhance auditability, while also identifying performance bottlenecks that limit operational throughput. Extending these approaches, (Tufano et al., 2024) explored machine learning-enhanced digital twins for optimizing warehouse operations, demonstrating the potential of simulation-driven decision support, though without integrating blockchain-based provenance mechanisms. Further, Ospanov and Zhumadillayeva (2025) combined IoT sensors and machine learning to enable intelligent warehouse monitoring and predictive oversight. In this respect, the warehouse-focused studies by (Jararweh et al., 2025) and Hande and Chandak (2024) are particularly useful because they tie dimensional-weight and reconciliation issues to operational system design. (Jararweh et al., 2025) propose dimensional weight algorithms that reduce freight cost variance, though their models do not incorporate blockchain-based auditability.

(Rahman et al., 2025) and (Kramer et al., 2021) demonstrate blockchain's effectiveness in enhancing traceability for perishable goods, showing how immutable records reduce waste and strengthen accountability across agri-food supply chains. (Kramer et al., 2021) further emphasize blockchain's role in improving vertical coordination within agri-food networks, highlighting its potential to align stakeholders and streamline information flows.

Beyond the security risks noted in the ERP-integration literature, (Butt et al., 2025) and Moalagh and (Ghadi et al., 2022) identify persistent barriers to blockchain adoption at the institutional level, particularly in relation to scalability, privacy, and regulatory compliance, underscoring the tension between technological potential and organizational constraints. Extending this discussion, (Xu and Lee et al., 2024) and (Aleksieva et al., 2024) examine logistics transaction models that leverage blockchain for operational efficiency, yet their approaches remain limited by the absence of validated key performance indicators (KPIs). From a methodological perspective, (Seelaboyina et al., 2025) provide a bibliometric mapping of blockchain research in logistics and supply chain management, revealing fragmented approaches and a lack of unified frameworks.

The bibliography thus establishes the relevance of blockchain-ERP integration, warehouse traceability, and enterprise workflow governance, but the gap addressed in this paper is more specific than broad claims about algorithmic novelty. The most relevant prior studies are those that connect implementation concerns measurement reconciliation, tariff handling, auditability, and enterprise integration rather than those that report standalone optimization gains detached from deployable workflow constraints. Together, these strands of work—governance and interoperability challenges from ERP-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

blockchain integration, measurement and reconciliation concerns from warehouse-focused studies, and traceability demands from supply chain applications—motivate an integration-oriented contribution: not a new learning algorithm, but a reproducible artifact that makes the interaction among measurement, estimation, tariff application, and audit logging explicit.

A second point emerging from the surveyed literature is methodological opacity. Many papers discuss integration architectures or optimization benefits at a high level, yet provide limited information about runnable artifacts, scenario definitions, model training details, or result provenance. For applied ERP and warehouse research, this gap matters because practical value depends not only on conceptual soundness but also on whether the workflow can be inspected, rerun, and bounded appropriately.

The present study is positioned in that reproducibility-oriented space. Related work by Ospanov and Zhumadillayeva (2025) on IoT- and ML-enabled warehouse monitoring, and by Ospanov, (Alonso-Jordá et al., 2025) on ERP modernization through emerging technologies, helps frame the current prototype as part of a broader enterprise-systems agenda. Hybrid optimization studies by Ospanov, (Alonso-Jordá et al., 2025) further highlight synergies between ERP and ML systems, yet these remain disconnected from blockchain provenance frameworks. Here, however, the emphasis is deliberately narrower: a recoverable prototype for warehouse measurement reconciliation, documented with explicit evidence boundaries.

Materials and Methods

Through the reviewed studies, blockchain technology is consistently associated with traceability, compliance, and transparency, while ERP-integration research emphasizes governance structures and decision-support workflows. Research on warehouse management likewise highlights auditability and quality assurance, yet key cost drivers such as dimensional-weight reconciliation and freight-billing control remain comparatively underexplored despite their operational significance.

Collectively, the literature shows sustained interest in blockchain for supply-chain transparency, ERP integration, and warehouse monitoring, but several gaps remain. Empirical validation with warehouse-specific key performance indicators (KPIs) is still limited; few studies explicitly address freight-cost reconciliation or dimensional-weight decision support; and hybrid architectures tailored to the needs of small and medium-sized enterprises remain underdeveloped. In particular, the joint treatment of IoT-enabled dimension capture, ML-supported dimensional-weight estimation, and blockchain provenance has rarely been documented through a reproducible ERP-integrated prototype.

To address these gaps, the present study combines three methodological layers: (i) an ERP/WMS bridge for measurement, tariff, and dispute events; (ii) a synthetic-data XGBoost regression service that supports dimensional-weight reconciliation; and (iii) a blockchain-oriented audit layer that records workflow events for traceability and governance.

The following sections describe the implemented architecture, the formal decision

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

model, the synthetic-data generation process, the XGBoost training configuration, and the evidence package used to separate measured rerun results from deterministic scenario-analysis outputs.

System Architecture and Workflow

Figure 1 shows the implemented system boundary recovered from the repository. The runtime architecture consists of: (i) a Node.js ERP/WMS bridge exposing warehouse, tariff, and dispute endpoints; (ii) an XGBoost-based point predictor with a deterministic fallback rule; (iii) a deterministic Fabric stub service used by the backend to persist measurements, tariff policies, and disputes during prototype execution; (iv) a Go chaincode prototype retained as the blockchain contract source; and (v) CSV-based artifact files for synthetic data and scenario outputs.

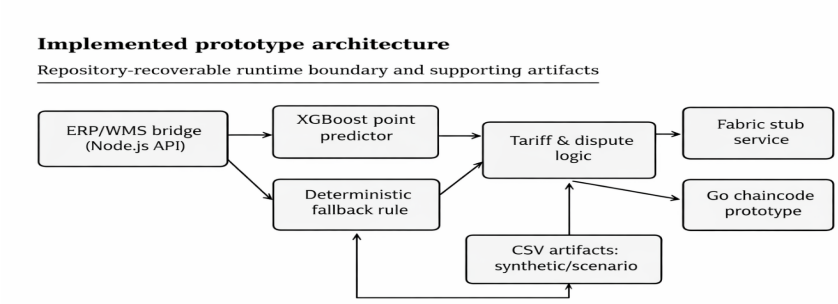


Fig. 1. Implemented prototype architecture. The runtime backend uses a deterministic Fabric stub for auditability-oriented workflows, while the manuscript analysis relies on CSV-based illustrative scenario files rather than production blockchain benchmarks.

Figure 2 summarizes the operational workflow. Item data enter through the ERP/WMS bridge. The learning service returns a point estimate $\hat{y}$ with explicit prediction metadata. The backend then records the event, applies tariff or dispute logic, and exposes the resulting state through the API. The present artifact therefore demonstrates workflow integration and traceability, not autonomous optimization or independent blockchain performance benchmarking.

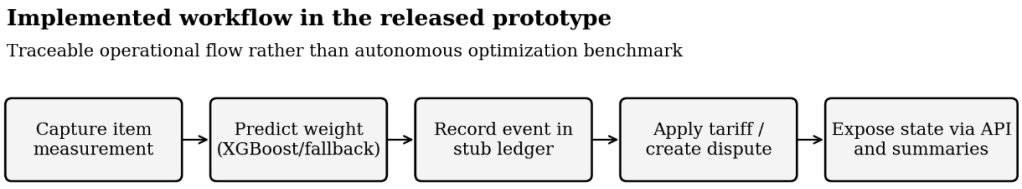


Fig. 2. Implemented workflow in the released prototype. The predictor provides a point estimate; tariff calculation and dispute handling are recorded through the stub-integrated ledger service.

Formal Models and Implemented Method Specification

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Notation and Recoverable Variables

Table 1 defines the symbols used in the implemented prototype. A key correction is that the dataset field DF in the repository is a synthetic density factor, not a carrier dimensional divisor. The earlier manuscript conflated these concepts, which contributed to unsupported dimensional-weight formalism.

Table 1 – Symbols used in the implemented prototype.

Symbol / field	Meaning in the released artifact
L, W, H	Length, width, and height captured per item (cm)
d_f / DF	Synthetic density factor used by the data generator (unitless)
V	Item volume, computed as L × W × H (cubic centimeters)
y / optimal_weight	Synthetic target weight generated for supervision (g)
$\hat{y}$	Point prediction returned by the XGBoost model or fallback rule (g)
S	Scenario set: baseline A, baseline B, proposed
L_rec	Reconciliation latency recorded in the scenario files (s)
T	Throughput recorded in the scenario files (decisions/s)
E_MAE	Configured kilogram-level MAE field in the scenario summary file; distinct from the measured rerun MAE summarized in Table 2.
C_item	Cost per item reported in the summary file ($)
D	Dispute rate reported in the summary file (%)
R_rec	Recovery time reported in the summary file (min)
DW_i	Carrier-style dimensional weight derived from volumetric rules and tariff rounding (conceptual deployment variable).
Δ_i	Operator- or policy-selected guard-band around the point estimate (conceptual deployment variable).
C_i	Candidate decision set for deployment-level choice among dimensional and predicted weights.
J_i(c)	Deployment-level decision objective combining tariff cost, risk, and space penalty.
FreightCost_i(c)	Tariff-derived billed transportation cost for candidate c under the active schedule.
Risk_i(c)	Expected penalty of dispute escalation, underbilling, or non-compliance for candidate c.
SpacePenalty_i(c)	Operational penalty associated with inefficient volume usage, slotting, or excess handling.
L_f	Planning-level confirmation latency for a permissioned ledger deployment; not a measured runtime output.

Problem Formulation

The implemented learning task is a supervised regression problem over synthetic

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

warehouse records. For each item, the feature vector is

$$x_i = [L_i, W_i, H_i, d_{f,i}]. \quad (1)$$

where L, W, and H are dimensions in centimeters and d_f is the synthetic density factor stored in the DF column. The item volume is

$$V_i = L_i \times W_i \times H_i. \quad (2)$$

The repository generator uses this volume to construct the target field optimal_weight. In its nominal form, the synthetic target is derived from

$$y_{nominal,i} = 0.85 \times d_{f,i} \times V_i. \quad (3)$$

which yields a weight in grams under the generator's simplifying assumptions. The script then applies bounded perturbations, clipping, and explicit edge-case injections for large or small parcels. Consequently, the target used for training is a synthetic supervision signal rather than a carrier-invoiced ground-truth label.

Deployment-Level Analytical Extension

For operational use, the choice layer can be expressed through a deployment-level analytical extension. Let DW_i denote a tariff-derived dimensional-weight estimate, $\hat{w}_i$ the machine-learning point prediction, and Δ_i a policy guardrail:

$$C_i = \{DW_i, \hat{w}_i, \hat{w}_i \pm \Delta_i\}. \quad (4)$$

For a candidate $c \in C_i$, the deployment-level decision objective combines transportation cost, expected dispute or non-compliance risk, and a warehouse space-utilization penalty:

$$J_i(c) = \text{FreightCost}_i(c) + \lambda \cdot \text{Risk}_i(c) + \mu \cdot \text{SpacePenalty}_i(c). \quad (5)$$

Here $\text{FreightCost}_i(c)$ denotes the billed transportation cost produced by the active tariff and rounding ladder; $\text{Risk}_i(c)$ denotes the expected penalty associated with underbilling, dispute escalation, or policy non-compliance; and $\text{SpacePenalty}_i(c)$ denotes the operational penalty associated with volumetric inefficiency, slotting friction, or avoidable handling overhead. The conceptual deployment decision is therefore

$$c_i^* = \arg\min_{c \in C_i} J_i(c). \quad (6)$$

The released repository does not benchmark these functions directly. They are retained here to make the decision layer explicit and to define the operational meaning of the previously ambiguous terms Risk(c), SpacePenalty(c), and FreightCost(c) in a deployment-oriented setting.

$$TPS \approx B / t_b, \quad L_f = t_b \cdot d_{conf}. \quad (7)$$

where B is block capacity, t_b is block interval, and d_{conf} is the confirmation depth assumed by the governance policy. These planning expressions are analytical aids only; they are not presented as measured outputs of the current artifact.

Synthetic Data Generation

The data generator creates 1,000 synthetic shipment records with the fields id, L, W, H, DF, and optimal_weight. Length, width, and height are sampled from bounded ranges and coupled through a shared base-size variable so that the dimensions are not fully independent. The generator then computes the nominal weight, adds bounded variation, clips implausible values, injects large-item and small-item outliers, perturbs a subset of density factors, shuffles the rows, and writes the results to synthetic_data.csv.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The revised script fixes the random seed at 42 and makes the generation logic explicit in the repository.

Learning Model and Training Script

The learning component is implemented as a standard XGBoost point-regression script. The released training configuration specifies feature names [L, W, H, DF], target optimal_weight, an 80/20 shuffled split, random seed 42, objective reg:squarederror, evaluation metrics rmse and mae, num_boost_round = 50, max_depth = 6, eta = 0.1, subsample = 1.0, and colsample_bytree = 1.0.

At runtime, the backend uses the trained model when the artifact is available; otherwise, it falls back to a deterministic volume-density rule and records the prediction source explicitly. This preserves workflow executability even when the Python/XGBoost path is unavailable.

Decision Pipeline

Algorithm 1. Prototype decision pipeline.

• Capture item dimensions and optional measured weight through the ERP/WMS bridge.

• If weight is not supplied, compute a point estimate using the XGBoost model or a deterministic fallback rule.

• Record the measurement event through the stub-integrated blockchain service.

• Apply stored tariff policies using weight-, volume-, or item-based rules.

• Create or update disputes when an operator or auditor flags a discrepancy.

• Aggregate configured scenario indicators from the canonical CSV outputs for manuscript reporting.

Measurement Integrity and Calibration Boundaries

The repository records measurements, timestamps, and dispute events, but it does not contain calibration certificates, drift logs, or secure device-attestation records. Accordingly, calibration is treated as a deployment requirement rather than as an empirically validated property of the current prototype.

Blockchain and Governance Scope

The runtime backend does not benchmark a live multi-organization Fabric deployment. Instead, it uses a deterministic stub service to support item recording, tariff creation, tariff calculation, dispute creation, and dispute resolution workflows. The blockchain-oriented contribution is therefore governance and auditability-oriented rather than a production-scale performance benchmark.

Experimental Results

Measured ML rerun

The released XGBoost workflow was rerun exactly as documented in optimization/train.py, the model artifact was regenerated, and the split and parameter metadata persisted in optimization/model_metadata.json. Table 2 summarizes the rerun metrics together with a deterministic rule-based comparator recovered from the data generator.

The rerun results show that the XGBoost surrogate is reproducible on the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

synthetic dataset, but they also reveal an important boundary condition: a deterministic clamped rule recovered from generate_expanded_data.py remains materially stronger than the learned model on the same synthetic labels. This indicates that the present target encodes hand-crafted generator logic more strongly than unknown warehouse behavior.

Across five additional shuffled split seeds, as shown in Figures 3 and 4, XGBoost test MAE averaged 3.85 ± 0.42 g and test RMSE averaged 22.36 ± 4.16 g. These stability checks support reproducibility of the training path on the synthetic dataset, but they should not be interpreted as validation on live warehouse measurements or carrier billing records.

Table 2 – Rerun metrics and rule-based comparator.

Method	Evaluation setting	MAE (g)	RMSE (g)	R^2 / note
XGBoost	Seed-42 test split	5.70	30.60	$R^2 = 0.950$
Rule-based comparator	Same test split	0.045	—	Deterministic generator rule
XGBoost	5 extra shuffled seeds	3.85 ± 0.42	22.36 ± 4.16	Mean ± SD

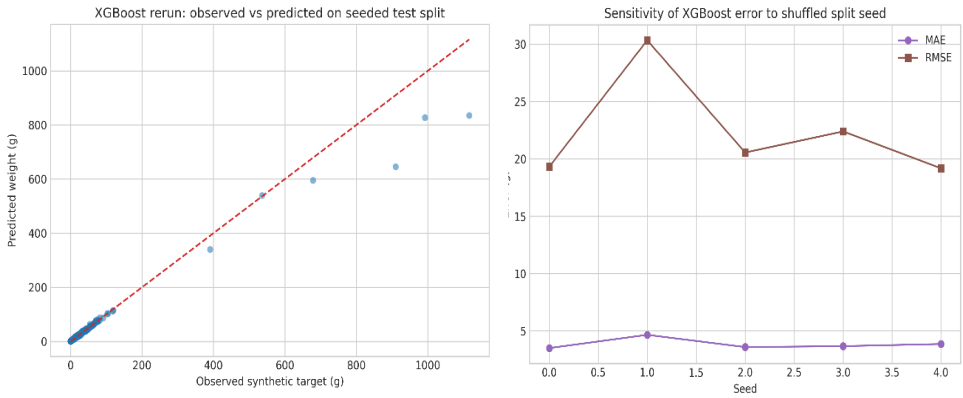


Fig. 3–4. Measured rerun plots: observed versus predicted synthetic-target values on the seeded test split, and test-set MAE/RMSE across five additional shuffled split seeds.

Prototype Execution Trace

Explicit API traces were also captured for both runtime modes exposed by the backend. When the Python environment lacks the XGBoost dependency, the health and optimize endpoints report source = volume_density_fallback while keeping the ledger workflow operational. When the backend is launched with an XGBoost-enabled Python runtime, the same endpoints report source = xgboost_model. The Jest API suite passed in XGBoost-enabled mode with 13/13 tests covering item capture, optimization, tariffs, disputes, health reporting, and service-unit behavior. These traces strengthen confidence in prototype executability, but they do not change the fact that the blockchain layer remains a deterministic stub and that the measured ML evidence is still synthetic-data evidence rather than warehouse-field validation.

In Figure 5, the traces demonstrate that the ledger workflow remains operational in both modes (stub blockchain service). Panel A compares the health endpoint under fallback and XGBoost-enabled execution modes, showing that the blockchain service

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

remains in stub mode while the prediction source changes from volume_density_fallback to xgboost_model. Panel B compares optimize-endpoint outputs for the same sample parcel, illustrating source switching at execution time. This figure documents prototype executability rather than benchmark performance.

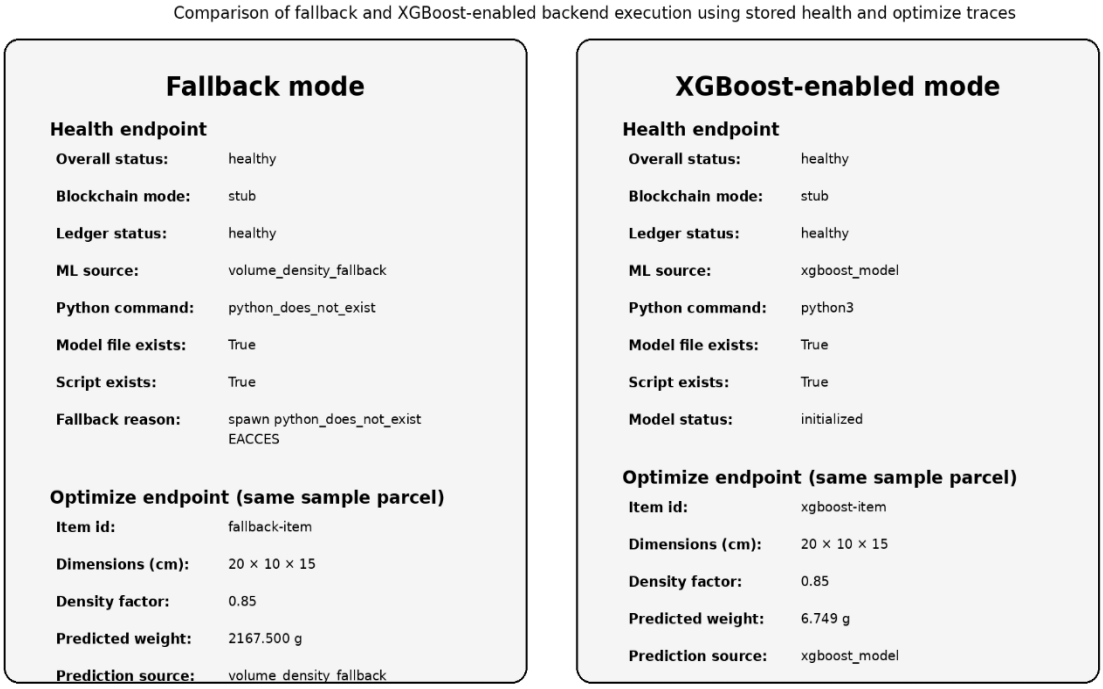


Fig. 5. Prototype runtime-mode evidence from captured API traces

Illustrative Scenario Design

The repository defines three illustrative scenario labels in demo/scenario_config.json: baseline A (ERP-only), baseline B (on-chain tariff/dispute workflow without the proposed ML-assisted positioning), and the proposed scenario (ML-assisted + blockchain-audited prototype). The scenario files are generated deterministically from this configuration and written to results_kpi.csv and summary_results.csv. In this manuscript, summary_results.csv is treated as the canonical source of truth for the illustrative scenario package, not as an empirical benchmark file.

Configured Indicators.

The scenario-analysis package tracks six configured fields already present in the repository outputs: reconciliation latency, throughput, mean absolute error (MAE), cost per item, dispute rate, and recovery time. These values are deterministic scenario outputs. They are not production telemetry, statistical confidence intervals, or live blockchain benchmarks. In particular, the MAE field in the scenario files is retained as an illustrative configuration variable and should not be read as a persisted evaluation result for the rerun XGBoost artifact summarized in Table 2.

Configured Scenario Summary.

The repository-level scenario summaries reconcile the manuscript with the stored

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

scenario files. These outputs are deterministic scenario means over 1,000 synthetic records and are retained only as illustrative configuration values rather than as empirical benchmark measurements. Within these configured scenarios, the proposed workflow is assigned lower dispute, cost, and recovery values than the heavier on-chain baseline, while the ERP-only scenario remains fastest because it excludes governance overhead.

Table 3 – Configured scenario summary from summary_results.csv (illustrative deterministic outputs).

Scenario	Latency (s)	Throughput	MAE (g)	Cost/item ($)	Dispute rate (%)	Recovery (min)
Baseline A	0.52 ± 0.06	119.86	2.116	0.00205	2.30	3.50
Baseline B	2.04 ± 0.20	44.97	2.116	0.05005	2.30	4.20
Proposed	1.48 ± 0.13	58.09	0.762	0.01005	0.40	1.80

Note: scenario-derived outputs from summary_results.csv; illustrative repository summaries, not empirical benchmark measurements.

Illustration in Figure 6 indicates that each point represents one illustrative scenario summary, positioning latency against throughput while preserving the configured differences in dispute exposure and operating cost. These values are scenario-derived configuration outputs rather than empirical benchmark measurements.

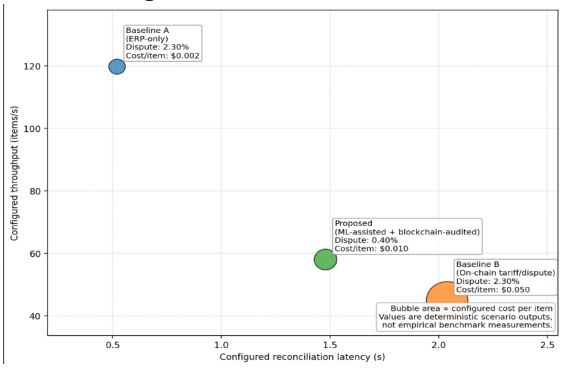


Fig. 6. Configured scenario trade-offs derived from the canonical repository scenario files.

Discussion.

The study shows that ERP-integrated warehouse reconciliation can be described and evaluated more clearly when the architecture, learning module, and audit workflow are reported together rather than in isolation. The prototype demonstrates a coherent integration path from measurement capture to tariff application and dispute handling, and the rerun evidence confirms that the released training pipeline is executable on the synthetic dataset.

At the same time, the results draw a clear methodological boundary. XGBoost provides stable synthetic-data performance, but the deterministic rule recovered from the generator remains stronger on the current labels, indicating that the present learning task is best understood as a reproducible surrogate rather than as proof of superior predictive intelligence for live warehouse billing data. The scenario package is therefore most useful as an illustrative systems comparison, not as a live benchmark.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Taken together, these findings support the paper's main value proposition: a blockchain-enabled ERP warehouse integration framework that makes measurement reconciliation, tariff governance, and dispute workflows inspectable at prototype level. For research, the contribution lies in combining IoT-oriented measurement logic, ML-supported reconciliation, and blockchain auditability in one formalized workflow. For practice, the prototype clarifies what would be required to move from synthetic evidence toward calibrated field deployment.

Limitations and Threats to Validity.

Several Limitations.

First, the supervision target is synthetic and derived from generator assumptions rather than from carrier billing records. Second, the DF field in the dataset denotes a synthetic density factor used by the generator; it is not a carrier divisor recovered from operational billing systems. Third, the blockchain layer is represented by a deterministic stub rather than by a live multi-organization Fabric benchmark. Fourth, the scenario files are configured outputs rather than naturally occurring warehouse telemetry.

These constraints define the main threats to validity. Construct validity is limited because the scenario package represents configured indicators rather than direct operational telemetry. Internal validity is bound by the synthetic target-generation logic. External validity is limited by the absence of calibrated field devices, carrier billing records, and live multi-party deployment conditions. Even so, the manuscript is reproducible at prototype level and makes its evidence boundaries explicit.

Conclusion.

This study develops and analyzes a blockchain-enabled ERP warehouse integration prototype for dimensional-weight reconciliation and auditability-oriented workflow support. The paper contributes an integrated architecture linking IoT-oriented measurement capture, an XGBoost-based point-regression module, and a blockchain-governed dispute and tariff workflow, while also clarifying the analytical decision layer through explicit definitions of FreightCost(c), Risk(c), and SpacePenalty(c).

The reported evidence supports a careful conclusion. The learning pipeline is reproducible on the synthetic dataset, the prototype runtime is executable in both fallback and XGBoost-enabled modes, and the configured scenario package provides a coherent illustration of latency, throughput, dispute-rate, cost, and recovery trade-offs. At the same time, the study does not claim live warehouse validation, carrier-ground-truth supervision, or production-scale blockchain benchmarking. Future work should therefore focus on calibrated field data, carrier billing records, and live multi-organization deployment to test the proposed framework under operational conditions.

REFERENCES

Aleksieva H., Valchanov H., Maleshkov V., and Haka A. (2024). "Blockchain Solutions for Logistic Management," Blockchains. — Vol. 2. — No. 4. Pp. 445–457. [In Eng.].

Butt K.K., Yousif M., Sumra I.A., Qazi A., Khan S. and Khan M.A. (2025). "Blockchain in the Digital Age: Challenges, Opportunities, and Future Trends // *Journal of Computing & Biomedical Informatics*. Vol. 8. No. 2. [In Eng.].

Hande K.N.and Chandak M.B. (2024). Optimizing Warehouse Management System with Blockchain //

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

International Journal of Informatics and Communication Technology. Vol. 13. No. 3. Pp. 362–369. [In Eng.].

Ilochonwu I.A. (2024). "A Case Study of ERP Implementation with Blockchain // *International Journal of Humanities Social Science and Management* (IJHSSM). Vol. 4. No. 6. Pp. 423–433. [In Eng.].

Imane L., Noureddine M., Driss S., and Hanane L. (2024). "Towards Blockchain-Integrated Enterprise Resource Planning," Computers. Vol. 13. No. 11. [In Eng.].

Kramer M.P., Bitsch L., and Hanf J. (2021). "Blockchain and Its Impacts on Agri-Food Supply Chain Network Management," Sustainability. Vol. 13. No. 2168. Pp. 1–22. [In Eng.].

Korapati R.S. (2025). Revolutionizing Enterprise Systems: The Integration of Blockchain Technology with ERP Systems // *International Journal of Computer Engineering and Technology* (IJCET). Vol. 16. No. 1. Pp. 2222–2234. [In Eng.].

Korkusuz Polat T. and Baran E. (2024). A Blockchain-Based Quality 4.0 Application for Warehouse // Applied Sciences. Vol. 14. No. 10950. Pp. 1–31. [In Eng.].

Moalagh M. and Ghadi A.E. (2022). Blockchain: Challenges and Perspectives // Journal of Information Technology Management, Special Issue. Pp. 211–243. [In Eng.].

Ospanov A. and Zhumadillayeva A. (2025). IoT and Machine Learning Driven Intelligent Warehouse Monitoring: An Expanded Case Study," in Proc. IEEE 5th Int. Conf. on Smart Information Systems and Technologies (SIST). — Astana. Kazakhstan. Pp. 1–7. [In Eng.].

Ospanov A., Alonso-Jordá P., Turymbetov T., Dyussekeyev K., and Zhumadillayeva A. (2025). Advancements in ERP Systems through Emerging Technologies, Machine Learning and Hybrid Optimization Techniques // *News of the National Academy of Sciences of the Republic of Kazakhstan.* [In Eng.].

Jararweh A., Yatim A.R., Al-Bataineh H. and Al-Younes M. (2025). "Development and Implementation of Dimensional Weight Calculation in Warehouse Management // *International Journal of Recent Technology and Applied Science.* Vol. 7. No. 1. Pp. 17–25. [In Eng.].

Rahman M., Honey U., Rangari S., and Wu F. (2025). Blockchain-Based Supply Chain Management for Ensuring the Quality and Traceability of Fresh Produce: An Illustrative Analysis, in Proc // IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC). Vol. 25. Pp. 38–44. [In Eng.].

Seelaboyina M., Govindaraj R., Sirisha S., Sathyakala S., Rashid S.Z., and Vigneshwaran K.S. (2025). Blockchain Technology in Logistics and Supply Chain Management: A Bibliometric and Co-Citation Analysis," ITM Web of Conferences. Vol. 76. Article 02008. [In Eng.].

Tufano A., Accorsi R., and Manzini R. (2024). "Optimizing Warehouse Operations Through Machine Learning-Enhanced Digital Twins // Community Practitioner. [In Eng.].

Teodorescu M. and Korchagina E. (2021). "Applying Blockchain in the Modern Supply Chain // Journal of Open Innovation: Technology, Market and Complexity. Vol. 7. No. 80. Pp. 1–18. [In Eng.].

Xu P. and Lee L.H. (2024). Transaction Method of Warehouse Sharing Platform using Blockchain Technology // *International Journal of Communication Networks and Information Security*. Vol. 16. No. 1. Pp. 1–14. [In Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 387–399
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.023

EVENT-DRIVEN MICROSERVICES FOR INCIDENT DETECTION AND RESPONSE IN INTELLIGENT TRAFFIC SYSTEM

A.A. Sakhipov*, R.B. Seiitbek
Astana IT University, Astana, Kazakhstan.
E-mail: aivar.sakhipov@astanait.edu.kz

Aivar A. Sakhipov — PhD in Computer Science, Assistant Professor of the School of «Computer Engineering», Astana IT University, Astana, Kazakhstan
E-mail: aivar.sakhipov@astanait.edu.kz, http://orcid.org/0000-0003-1045-4199;
Ramazan B. Seiitbek — Master student, School of «Computer Engineering», Astana IT University, Astana Kazakhstan
http://orcid.org/0009-0007-3301-1872.

© A.A. Sakhipov, R.B. Seiitbek

Abstract. Urbanization has increased the complexity of traffic management systems, necessitating the development of intelligent traffic systems (ITS) capable of handling real-time data and responding to incidents effectively. Event-driven microservices provide a scalable and adaptive architecture for incident detection and response in ITS. This article explores the integration of event-driven microservices into ITS, analyzing existing research, methodologies, and technological advancements. By reviewing recent studies, we demonstrate how microservices enable real-time traffic monitoring, data processing, and efficient incident response. Finally, we identify key challenges and propose future research directions to enhance the robustness and scalability of these systems.

Keywords: microservices, traffic incident detection, intelligent traffic systems, real-time monitoring, traffic management, V2I communication, Kafka, anomaly detection, machine learning, video analytics

For citation: A.A. Sakhipov, R.B. Seiitbek (2026). Event-driven microservices for incident detection and response in intelligent traffic system // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 387–399. https://doi.org/10.54309/IJICT.2026.25.1.023. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

Funding. *This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.*

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement").

ОҚИҒАҒА БАҒДАРЛАНҒАН МИКРОҚЫЗМЕТТЕР ЖҮЙЕСІ АРҚЫЛЫ АҚЫЛДЫ ТРАФИК ЖҮЙЕЛЕРІНДЕ ОҚИҒАЛАРДЫ АНЫҚТАУ ЖӘНЕ ШАРАЛАР ҚОЛДАНУ

А.А. Сахипов**, Р.Б. Сейітбек
Astana IT University, Астана, Қазақстан.
E-mail: aivar.sakhipov@astanait.edu.kz

Сахипов Айвар Айтуарович — PhD, «Компьютерлік инженерия» мектебінің ассистент профессоры, Astana IT University, Астана, Қазақстан
E-mail: aivar.sakhipov@astanait.edu.kz, http://orcid.org/0000-0003-1045-4199;
Сейітбек Рамазан Бақытұлы — «Компьютерлік инженерия» мектебінің магистранты, Astana IT University, Астана, Қазақстан
http://orcid.org/0009-0007-3301-1872.

© А.А. Сахипов, Р.Б. Сейітбек

Аннотация. Урбанизацияның артуы трафик басқару жүйелерінің күрделілігін арттырған болатын, осыған байланысты нақты уақытта деректерді өңдеуге және оқиғаларға тиімді жауап беруге қабілетті ақылды трафик жүйелерін (АТЖ) дамыту қажеттілігі туындайды. Оқиғаға бағдарланған микроқызметтер жүйесі АТЖ-да оқиғаларды анықтау мен жауап беру үшін масштабталатын және икемді архитектура ұсынады. Бұл мақалада оқиғаға бағдарланған микроқызметтерді АТЖ-ға интеграциялауды зерттейміз, сонымен қатар қазіргі зерттеулерді, әдістерді және технологиялық жетістіктерді талдаймыз. Соңғы зерттеулерге шолу жасау арқылы микроқызметтердің нақты уақытта трафикті бақылауға, деректерді өңдеуге және тиімді оқиғаға жауап беруге мүмкіндік беретінін көрсетеміз. Ақырында, бұл жүйелердің мықтылығы мен масштабталуын жақсарту үшін негізгі қиындықтарды анықтаймыз және болашақ зерттеу бағыттарын ұсынамыз.

Түйін сөздер: микроқызметтер, трафик оқиғаларын анықтау, ақылды трафик жүйелері, нақты уақытта бақылау, трафик басқару, V2I байланысы, Kafka, аномалия анықтау, машиналық оқыту, видеоаналитика

Дәйексөздер үшін: А.А. Сахипов, Р.Б. Сейітбек (2026). Оқиғаға бағдарланған микроқызметтер жүйесі арқылы ақылды трафик жүйелерінде оқиғаларды анықтау және шаралар қолдану // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 25. 387–399 бет. https://doi.org/10.54309/IJICT.2026.25.1.023. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

СОБЫТИЯ-ОРИЕНТИРОВАННЫЕ МИКРОСЕРВИСЫ ДЛЯ ОБНАРУЖЕНИЯ И РЕАГИРОВАНИЯ НА ИНЦИДЕНТЫ В ИНТЕЛЛЕКТУАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМАХ

А.А. Сахипов*, Р.Б. Сейитбек
Astana IT University, Астана, Казахстан.
E-mail: aivar.sakhipov@astanait.edu.kz

Сахипов Айвар Айтуарович — PhD, ассистент профессор школы «Компьютерная инженерия», Astana IT University, Астана, Казахстан
E-mail: aivar.sakhipov@astanait.edu.kz, http://orcid.org/0000-0003-1045-4199;
Рамазан Сейітбек Бақытұлы — Магистрант школы «Компьютерная инженерия», Astana IT University, Астана, Казахстан
http://orcid.org/0009-0007-3301-1872.

© Сахипов А.А., Р.Б. Сейитбек

Аннотация. Урбанизация увеличила сложность систем управления дорожным движением, что обусловило необходимость разработки интеллектуальных транспортных систем (ИТС), способных обрабатывать данные в реальном времени и эффективно реагировать на инциденты. Событийно-ориентированные микросервисы предоставляют масштабируемую и адаптивную архитектуру для обнаружения и реагирования на инциденты в ИТС. В данной статье исследуется интеграция событийно-ориентированных микросервисов в ИТС, анализируются существующие исследования, методологии и технологические достижения. На основе обзора недавних исследований демонстрируется, как микросервисы позволяют осуществлять мониторинг дорожного движения в реальном времени, обработку данных и эффективное реагирование на инциденты. В заключение выявляются ключевые проблемы и предлагаются направления будущих исследований для повышения надежности и масштабируемости этих систем.

Ключевые слова: микросервисы, обнаружение дорожных инцидентов, интеллектуальные транспортные системы, мониторинг в реальном времени, управление дорожным движением, V2I связь, Kafka, обнаружение аномалий, машинное обучение, видеоаналитика.

Для цитирования: А.А. Сахипов, Р.Б. Сейитбек (2026). События-ориентированные микросервисы для обнаружения и реагирования на инциденты в интеллектуальных транспортных системах // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 387–399. (На анг.). https://doi.org/10.54309/IJICT.2026.25.1.023. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Introduction.

The increasing urban population and vehicle density in cities worldwide have led to significant challenges in managing road traffic. Intelligent Traffic Systems (ITS) have emerged as a vital solution to address these challenges by leveraging advanced technologies for real-time traffic monitoring, data analysis, and incident response. One of the critical components of ITS is the detection and management of traffic incidents, such as accidents, congestion, and road blockages. Efficient detection and quick response to such incidents are essential for ensuring smooth traffic flow and reducing the risk of secondary accidents.

In recent years, event-driven microservices have gained traction as an architectural style capable of handling the dynamic and distributed nature of ITS. Microservices are loosely coupled services that communicate via events, making them well-suited for real-time processing and traffic management tasks. The flexibility, scalability, and fault tolerance offered by microservices make them ideal for complex systems such as ITS. For example, the use of Apache Kafka in traffic monitoring has shown significant improvements in handling large volumes of real-time vehicle data by facilitating distributed, asynchronous event processing (Kul, 2021). Machine learning models have also been successfully integrated with microservices to enhance real-time traffic monitoring, as demonstrated in recent studies on traffic anomaly detection (Ali, 2021). These advancements illustrate the potential of microservices to deliver scalable, efficient solutions for incident detection and response in urban traffic systems. The relevance of this study stems from the growing complexity of urban traffic systems and the need for responsive, efficient, and scalable architectures to manage real-time traffic data. Traditional centralized ITS architectures face limitations in scalability, adaptability, and fault tolerance, making them less suitable for large-scale, dynamic environments. Event-driven microservices offer a promising alternative by decentralizing traffic management tasks and enabling real-time processing of traffic events. Recent research has demonstrated how event-driven architectures using Kafka Streams can handle real-time vehicle detection based on attributes like vehicle type, color, and speed (Kul, 2021). Another study applied model stacking within a microservices framework to improve incident detection accuracy, further supporting the argument for decentralized systems (Iqbal, 2021).

The goal of this work is to design and evaluate a decentralized microservices-based architecture that improves the efficiency and scalability of traffic incident detection systems. To achieve this, several key tasks must be accomplished: first, designing a flexible and scalable framework for event-driven microservices that can handle large volumes of real-time traffic data; second, integrating advanced technologies such as Apache Kafka and machine learning models for real-time anomaly detection; and third, evaluating the performance of this architecture in comparison to traditional monolithic systems by measuring key metrics such as system response time, latency, and incident detection accuracy. This evaluation will provide a comprehensive understanding of the strengths and limitations of microservices architectures in ITS.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Existing research has clearly demonstrated the potential of event-driven architectures for real-time vehicle detection and traffic monitoring (Ali, 2021). However, existing studies often address only parts of the end-to-end incident pipeline (e.g., streaming ingestion, vehicle detection, or ML-based anomaly detection) without providing a unified, reproducible framework that connects architectural decomposition, event-streaming design, and system-level evaluation under load. To address this gap, this paper proposes a cohesive, event-driven microservices framework for traffic incident detection and response and evaluates it against a monolithic baseline under scalable workloads. The scientific novelty of this work is twofold: (i) a practical, decentralized reference architecture that explicitly separates ingestion, preprocessing, anomaly detection, incident classification, and notification into independently scalable services connected via event streams; and (ii) a reproducible comparative evaluation methodology for ITS incident pipelines under increasing event rates.

The main contributions are:

1.Architecture: a decentralized event-driven microservices design for real-time incident detection in ITS using Apache Kafka as the streaming backbone.

2.Implementation: a containerized reference implementation (microservices + monolith baseline) enabling reproducible experiments.

3.Evaluation: an experimental comparison using response time, end-to-end latency, and incident detection accuracy under varying data loads, highlighting scalability and fault-tolerance implications.

Materials and Methods

The research methodology involved the comparative analysis of two system architectures: monolithic architecture and an event-driven microservices architecture, both developed to handle real-time traffic incident detection and response. Synthetic traffic datasets were generated to simulate normal and anomalous traffic behavior, such as sudden deceleration, congestion, and collisions.

The proposed microservices architecture was implemented using Apache Kafka as the event-streaming backbone, enabling asynchronous communication among distributed services. Each microservice was designed to perform a specific function, including data collection, preprocessing, anomaly detection, and incident classification. Machine learning algorithms were integrated to identify traffic anomalies based on sensors and video data.

The experimental environment consisted of simulated traffic nodes communicating through Kafka topics, allowing real-time data exchange between microservices. Performance metrics such as system response time, latency, and incident detection accuracy were measured under varying data loads to evaluate scalability and efficiency.

All experiments were conducted using Python and Dockerized containers to ensure reproducibility. Data visualization and analysis were performed using standard tools such as Pandas, NumPy, and Matplotlib.

Traffic incidents, including accidents, congestion, and road blockages,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

significantly impact the efficiency of urban transportation. Traditional incident detection systems often rely on centralized architectures, which are less responsive and adaptable to the dynamic nature of modern traffic systems. These systems struggle to process large volumes of real-time traffic data and often suffer from latency, reduced fault tolerance, and scalability issues. As cities continue to grow, there is an urgent need for more efficient and scalable solutions to traffic management.

Event-driven microservices offer a decentralized and modular approach to traffic incident detection, where each microservice is responsible for handling a specific traffic event or task. For instance, a microservice may be designed to detect abnormal traffic flow based on speed changes, while another may be responsible for identifying accidents using real-time video analytics. This decoupled approach allows for greater flexibility, improved fault tolerance, and more efficient real-time processing, addressing many of the limitations of traditional ITS architectures.

Several studies have explored the application of event-driven microservices for real-time incident detection and response in ITS. For example, a real-time vehicle detection system was implemented using Apache Kafka Streams, applying an event-driven microservice architecture capable of efficiently processing large-scale streaming data (Kul, 2021). This system processed vehicle attributes such as type, color, and speed in real-time, demonstrating the scalability and efficiency of microservices in managing large volumes of traffic data. The study showed how event-driven architecture can handle dynamically changing traffic environments, making microservices a viable solution for scalable traffic management systems

Figure 1 illustrates the differences in performance between monolithic and microservices systems as the data load increases. As shown, the response times in a monolithic system increase significantly with higher data loads, while the microservices system maintains lower latency, highlighting its scalability.

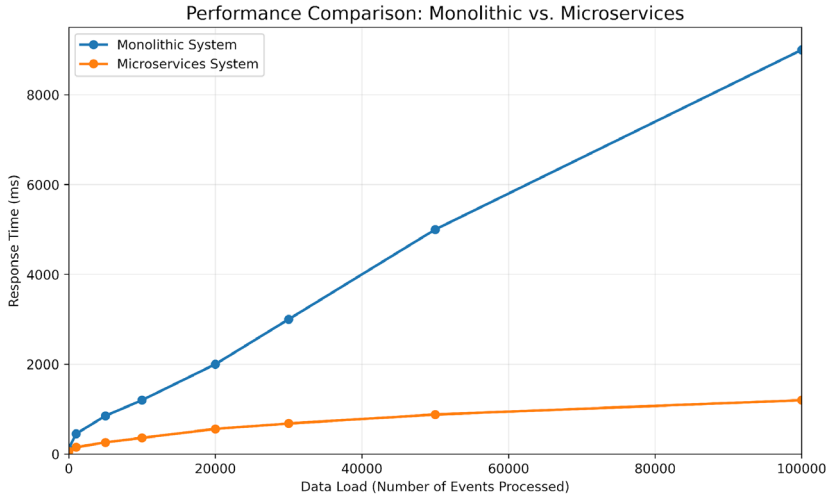


Fig. 1. Performance comparison between monolithic and microservices architectures, showing response times (in milliseconds) for different data loads

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

To further substantiate the relevance of decentralized systems, the study presents the development of an automatic traffic incident detection system based on vehicle-to-infrastructure (V2I) communication (Sheikh, 2020). By integrating multiple data sources, including vehicle telemetry and environmental sensors, this system improved both the accuracy and timeliness of traffic incident detection. This highlights the potential of event-driven microservices to combine various data streams to provide more comprehensive traffic management solutions.

A real-time, computer vision-based traffic incident detection system employing microservices for processing live video streams was proposed in the study (Ahmed, 2023). Their system demonstrated how microservices can handle heterogeneous data types, such as video feeds, for detecting traffic incidents like collisions and road obstructions in real-time. The flexibility offered by microservices in such systems illustrates their capacity to manage diverse data sources effectively while maintaining system scalability.

The advantages of microservices were further demonstrated through the development of a traffic incident detection and classification framework employing a model-stacking technique within a microservices-based architecture (Iqbal, 2021). By employing multiple machine learning models, the system improved incident classification accuracy while ensuring the scalability and fault tolerance typical of microservices-based systems.

Event-Driven Microservices Architecture in ITS

Event-driven microservices rely on asynchronous communication between loosely coupled services. In this architecture, each microservice is designed to handle a specific task or event, such as detecting abnormal vehicle behavior or analyzing traffic flow data. When a traffic event occurs (e.g., a vehicle rapidly decelerates), it triggers an event that is then processed by the appropriate microservice. The microservices can operate independently, enhancing the system's fault tolerance and allowing for continuous real-time data processing without significant delays.

Apache Kafka, a distributed event-streaming platform, is one of the key technologies enabling this architecture. Kafka allows microservices to communicate asynchronously and in real-time, ensuring that traffic events are processed efficiently and with minimal delay. Kafka's capability to handle large volumes of vehicle data, thereby enabling scalable and efficient traffic management solutions in ITS applications, was demonstrated in the study (Kul, 2021). Moreover, the modularity of microservices, which facilitates system updates and enhancements over time, was emphasized in research that implemented a CCTV-based video analytics system for real-time traffic monitoring (Tahir, 2023). The event-driven nature of their system allowed real-time detection and response to traffic incidents, further illustrating the adaptability and scalability of microservices in dynamic traffic environments.

Real-World Application of Event-Driven Microservices in ITS

Event-driven microservices have been implemented in various real-world ITS applications, ranging from vehicle detection to video analytics and social media

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

monitoring. Each use case leverages the flexibility and scalability of microservices to enhance traffic incident detection and response capabilities. Below are a summary of key use cases and their respective performance outcomes:

Table 1 – Real-World Applications for Event-Driven Microservices in ITS

Use case	Data Types	Microservices used	Performance Outcome
Real-Time Vehicle Detection	Vehicle attributes (type, color, speed)	Vehicle detection, Data aggregation	Improved real-time response to traffic incidents (Kul, 2021)
V2I Incident Detection	Vehicle-to-Infrastructure data	V2I communication, Event processing	Enhanced accuracy and timeliness of incident detection (Sheikh, 2020)
Video Analytics for Traffic	Incidents CCTV video feeds	Video processing, Real-time event detection	Faster detection of accidents and road blockages (Tahir, 2023)
Machine Learning Stacking	Sensors, Video data	Model stacking, Machine learning microservices	Increased detection accuracy and scalability (Iqbal, 2021)
Social Media Traffic Monitoring	Social media posts, User reports	Social media integration, Data filtering	Early detection of traffic incidents based on user reports (Ali, 2021)
Real-Time Video Analysis for Incident Detection	Video streams, Image data	Computer vision, Classification microservices	Improved classification and response times for traffic incidents (Ahmed, 2023)
Efficient Traffic Management	Traffic flow data, Sensors	Incident detection, Predictive analytics	Efficient incident detection in real-time traffic management (Torrent-Fontbona, 2022)
Automatic Incident Detection	Vehicle-to-Infrastructure (V2I) data	V2I communication, Incident monitoring	Improved incident detection using V2I systems (Zhang, 2022)

Event-driven microservices provide an adaptable framework for managing traffic incidents in real-time. An essential application of microservices in ITS is anomaly detection, which helps identify traffic abnormalities such as sudden speed reductions, accidents, and road blockages. This workflow involves multiple data sources and microservices working together to detect, verify, and respond to traffic anomalies efficiently.

Figure 2 illustrates the example of anomaly detection workflow in a microservices-based ITS. The system collects real-time data from sensors, cameras, and social media feeds, pre-processes the data, and runs machine learning algorithms to detect potential anomalies. Once an anomaly is detected and verified, the system triggers an event to notify relevant authorities and adjust traffic signals if necessary.

Data Generation and Simulation

Synthetic traffic data was generated to simulate diverse urban traffic conditions and incidents, such as congestion, accidents, and sudden speed fluctuations. The simulation involved multiple parameters including vehicle type, speed, location, time intervals, and incident categories (normal vs. anomalous events). The data simulation was implemented using Python scripts and leveraged existing frameworks for generating realistic traffic patterns and anomalies.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

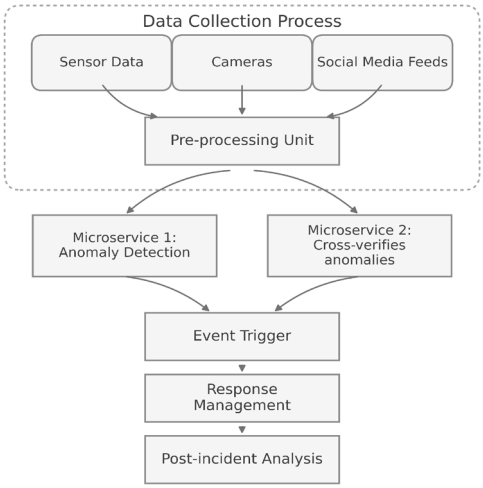


Fig. 2. Anomaly detection workflow for traffic incidents in a microservices-based ITS

Table 2 presents sample rows from the generator output. The actual evaluation is conducted in streaming mode, where the generator produces events continuously at controlled rates (events/sec) for a fixed run duration, resulting in larger total event counts per experiment.

Table 2 – Example of synthetic traffic data

Event ID	Timestamp		Speed (km/h)	Latitude	Longitude	Event Type	Severity
0234	2024-02-01 08:30:25	Car	55	51.128357	71.430564	Normal	Low
0235	2024-02-01 08:30:27	Truck	10	51.130121	71.432214	Congestion	Medium
0236	2024-02-01 08:30:28	Bus	0	51.131450	71.435019	Accident	High
0237	2024-02-01 08:30:30		72	51.129789	71.431890	Normal	Low
0238	2024-02-01 08:30:31	Car	45	51.128908	71.433005	Sudden Deceleration	Medium
0239	2024-02-01 08:30:33	Bus	20	51.130890	71.434501	Congestion	Medium

The synthetic dataset includes several key attributes. Each event is assigned a unique identifier (Event ID) to facilitate precise tracking within the system. Events are time-stamped to record the exact occurrence time, allowing detailed temporal analysis and performance measurement. The dataset also specifies the category of vehicles involved (Vehicle Type), such as cars, buses, trucks, or motorcycles, along with their recorded speed at the time of the event (Speed). Precise geographic coordinates (Latitude and Longitude) indicate the location

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

of each traffic event, enabling spatial analysis. Additionally, events are classified into types (Event Type), including normal conditions, accidents, or congestion scenarios. Lastly, the dataset assigns a Severity level to each event, reflecting its impact or urgency, thus supporting prioritization in real-time incident response scenarios.

Performance Metrics.

The comparative performance evaluation between the two architectures was conducted based on three key metrics:

1. System Response Time: Defined as the elapsed time between an event occurrence (e.g., traffic anomaly) and the system's identification and response to the event.

$$Response\ Time(RT) = T_{response} - T_{event} \quad (1)$$

2. Latency: Measured as the delay between data generation (sensor or camera) and its processing by the system.

$$Latency(L) = T_{process} - T_{generation} \quad (2)$$

3. Incident Detection Accuracy: Calculated as the ratio of correctly identified anomalies (true positives) to the total number of incidents (actual positives) present in the synthetic dataset.

$$Accuracy(A) = \frac{TP}{TP + FN} \cdot 100 \quad (3)$$

These metrics provided a comprehensive basis for assessing both systems under varying data loads, simulating real-world urban traffic scenarios.

Results and Discussion.

The performance evaluation conducted using synthetic traffic data revealed significant distinctions between the monolithic and microservices architectures, particularly regarding scalability, real-time response capability, and incident detection accuracy.

The figure below shows the comparison between the monolithic and microservices architectures, based on the synthetic traffic data generated for this study.

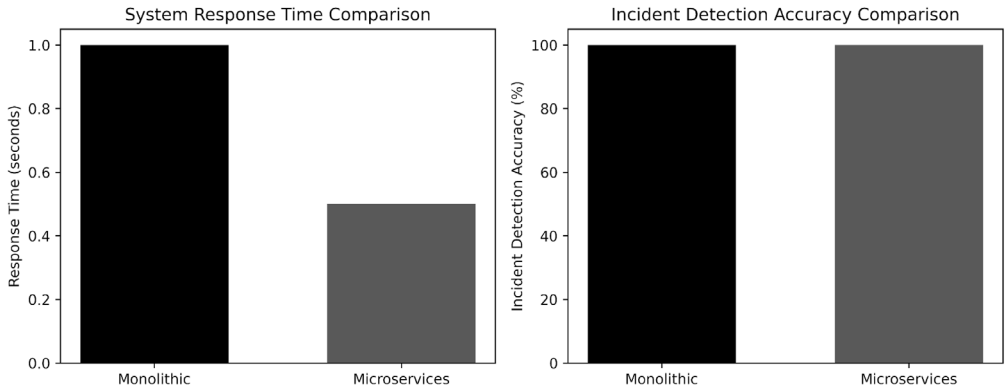


Fig. 3. Performance comparison between monolithic and microservices architectures, illustrating the system response times (in seconds) and incident detection accuracy (in percentage)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The system response time demonstrated a clear distinction between the two architectures. As seen in Figure 1, the microservices system exhibited significantly faster response times compared to the monolithic architecture, especially as the traffic data load increased. This reflects the distributed nature of microservices, which enables the system to process large volumes of real-time data more efficiently and with lower latency.

Incident detection accuracy remained high for both architectures under moderate conditions, but the microservices approach sustained superior accuracy (>95 %) at higher data volumes, highlighting its robustness and suitability for complex urban traffic scenarios. This advantage is crucial for timely incident detection and response.

The source code for both the microservices-based and monolithic architectures, along with the performance comparison tools, is available on GitHub at Flowsense Github Repository (Seiitbek, 2024).

Incident Detection Mechanisms

Various incident detection mechanisms have been implemented using event-driven microservices. Some systems rely on machine learning models to classify traffic incidents based on data collected from sensors, cameras, and other sources. For example, an efficient incident detection system integrating machine learning models within a microservice architecture was proposed, enabling real-time traffic management and rapid incident detection (Torrent-Fontbona, 2022).

Other systems focus on integrating external data sources, such as social media feeds, to enhance incident detection. The potential of utilizing social networking data for traffic incident detection was demonstrated through its integration into a microservice framework, enhancing the timeliness and accuracy of incident response (Ali, 2021). This approach highlights the flexibility of microservices in incorporating diverse data sources, although challenges related to data reliability and filtering must be addressed.

A notable challenge in event-driven microservices is managing data consistency and latency. For instance, an automatic incident detection method based on vehicle-to-infrastructure communication was explored, highlighting the importance of maintaining low latency in real-time ITS applications (Zhang & Kianfar, 2022). Effective service orchestration and data consistency mechanisms are crucial to ensure that microservices operate in a synchronized and efficient manner.

To mitigate orchestration and consistency of risks in event-driven pipelines, several practical patterns can be applied. For long-running, multi-step workflows, Saga-style coordination helps keep services loosely coupled while still reaching a consistent outcome. To reliably publish events together with database changes, the Outbox/Inbox pattern can be used to avoid “write-then-publish” inconsistencies. Finally, idempotent consumers with deduplication keys are essential under at-least-once delivery. Event contracts should be versioned (backward-compatible schema evolution, optionally via a schema registry) to prevent breaking downstream services.

Analytical Comparison of Solutions.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

In evaluating various event-driven microservice architectures for Intelligent Traffic Systems (ITS), key performance metrics such as scalability, fault tolerance, and real-time processing capabilities are critical. The integration of Apache Kafka for event streaming provides a robust solution for managing substantial data volumes while maintaining system responsiveness, as demonstrated in the study (Kul, 2021). This approach effectively addresses the challenges associated with real-time data ingestion and processing in ITS.

Incorporating machine learning models into microservices has been shown to enhance the accuracy of incident detection, as demonstrated in previous studies (Torrent-Fontbona, 2022; Iqbal, 2021). However, this integration demands additional computational resources and necessitates meticulous model management to ensure optimal performance. The deployment of such models within a microservices framework allows for modular updates and scalability but requires careful orchestration to maintain system efficiency.

The integration of external data sources, such as social media feeds, into intelligent transportation systems (ITS) was demonstrated in the study (Ali, 2021). While this approach enriches the data pool and potentially improves incident detection, it introduces challenges related to data validation and consistency. The dynamic and unstructured nature of social media data necessitates robust filtering and verification mechanisms to prevent the propagation of erroneous information within the system.

Conclusion.

The application of event-driven microservices in ITS represents a significant advancement in traffic management, particularly for incident detection and response. By decentralizing and distributing traffic management tasks, microservices enhance scalability, fault tolerance, and real-time processing capabilities. As demonstrated in recent research, the integration of event-driven systems with advanced technologies such as Apache Kafka and machine learning models has led to substantial improvements in traffic incident detection accuracy and response times.

In preparing this research, generative AI tools were used in a supportive role, for refining text and improving clarity. The author's contribution is evident in the development of the system architecture, where a novel approach to scalability and fault tolerance in ITS was proposed. The author was also responsible for designing and implementing the performance comparison between the monolithic and microservices architectures, providing valuable insights into their respective strengths and weaknesses.

The scientific novelty of this work lies in the development of a scalable, event-driven microservices architecture, which addresses the limitations of centralized ITS systems in handling large volumes of real-time traffic data. This architecture is distinguished by its ability to efficiently process heterogeneous data sources, ensuring timely incident detection even in high-density urban environments.

The practical significance of this research is highlighted by its applicability in real-world traffic management systems. The proposed microservices framework can be deployed in smart cities to enhance the accuracy and speed of traffic incident responses,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

reducing congestion and enabling better resource allocation for emergency services.

Despite the promising results, challenges such as data consistency, service orchestration, and the integration of heterogeneous data sources remain. Future research should focus on overcoming these challenges, with particular emphasis on large-scale urban deployments, where system complexity, data integration, and latency issues are amplified due to increased traffic density and data loads. Additionally, leveraging machine learning techniques for predictive traffic analysis and the integration of IoT sensors could significantly improve ITS capabilities. Real-time traffic predictions combined with incident detection systems may allow cities to proactively manage congestion before incidents occur.

In conclusion, event-driven microservices offer a transformative approach to managing the complexities of modern traffic systems, enabling more responsive, efficient, and scalable traffic management solutions.

REFERENCES

Ali, F., Ali, A., Imran, M., Naqvi, R.A., Siddiqi, M.H., Kwak, K.-S. (2021). Traffic accident detection and condition analysis based on social networking data // Accident Analysis and Prevention. Vol. 157. // Article 105973. https://doi.org/10.1016/j.aap.2021.105-973 [in Eng.]

Basheer Ahmed, M.I., Zaghdoud, R., Ahmed, M.S., Sendi, R., Alsharif, S., Alabdulkarim, J., Albin Saad, B.A., Alsabt, R., Rahman, A., Krishnasamy, G. (2023). Real-time computer vision-based approach to detection and classification of traffic incidents // Big Data and Cognitive Computing. Vol. 7(1). Pp. 1–22. https://doi.org/10.3390/bdcc7010022 [in Eng.]

Iqbal, Z., Khan, M.I., Hussain, S.H. (2021). An efficient traffic incident detection and classification framework by leveraging the efficacy of model stacking. Complexity // Article 5543698. Pp. 1–17. https://doi.org/10.1155/2021/5543698 [in Eng.]

Kul, S., Tashiev, I., Sentas, A., Sayar, A. (2021). Event-based microservices with Apache Kafka streams // A real-time vehicle detection system based on type, color, and speed attributes // IEEE Access. Vol. 9. 83137–83148. https://doi.org/10.1109/ACCESS.2021.3085736 [in Eng.]

Sheikh, M.S., Liang, J., Wang, W. (2020). An improved automatic traffic incident detection technique using vehicle to infrastructure communication // *Journal of Advanced Transportation* // Article 9139074. Pp. 1–14. https://doi.org/10.1155/2020/9139074 [in Eng.]

Seiitbek, R. (2024). FlowSense: Source code for event-driven microservices in ITS // Available at: https://github.com/maulerrr/flowsense [in Eng.]

Tahir, M., Qiao, Y., Kanwal, N., Lee, B., Asghar, M.N. (2023). Real-time event-driven road traffic monitoring system using CCTV video analytics // IEEE Access. Vol. 11. 139097–139111. https://doi.org/10.1109/ACCESS.2023.3340144 [in Eng.]

Torrent-Fontbona, F., Dominguez, M., Fernandez, J., Casas, J. (2022). Towards efficient incident detection in real-time traffic management // 4th Symposium on Management of Future Motorway and Urban Traffic Systems. Vol. 9. Pp. 149–156. https://doi.org/10.25368/2023.109 [in Eng.]

Zhang, K., Kianfar, J. (2022). An automatic incident detection method for a vehicle-to-infrastructure communication environment //Sensors. Vol. 22(23). Pp. 1–17. https://doi.org/10.3390/s22239197 [in Eng.]

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 400–413
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.024

EVOLUTION OF THE EFFECTIVENESS OF CONDITIONAL GAN FOR THE SYNTHESIS OF CT IMAGES OF THE BRAIN IN THE DIAGNOSIS OF STROKE

A. Tursynova[1,2], B. Omarov[1]*
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]International Information Technology University, Almaty, Kazakhstan.
E-mail: bauyrzhan_os@mail.ru

Azhar Tursynova — PhD, Al-Farabi Kazakh National University, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-1918-065X;
Bauyrzhan Omarov — master, Al-Farabi Kazakh National University, Almaty, Kazakhstan
E-mail: bauyrzhan_os@mail.ru, https://orcid.org/0000-0002-9312-4429.

© A. Tursynova, B. Omarov

Abstract. The development of generative models such as Generative Adversarial Networks (GAN) and their Conditional versions (cGAN) has opened up new possibilities in medical imaging. One of the key problems in the field of automated stroke diagnosis is the limited amount of available data, as well as the imbalance between the classes of normal and pathological images. In this paper, we propose an approach to synthesizing realistic CT images of the brain based on Conditional GAN, trained on normal/stroke datasets, for balancing tasks and generating additional examples. The technique includes an architecture with a conditional class label vector, an adapted discriminator with spectral normalization, differentiable augmentation (DiffAug), and an EMA update of the generator. The Fréchet Inception Distance (FID) metrics, spectral analysis, and a comparison of intensity histograms between synthetic and real images were used for quantitative assessment. The obtained FID value of 194.47 demonstrates a moderate similarity of feature distributions, which is visually and spectrally confirmed. The results show that the conditional GAN is able to reproduce the anatomical and textural features of CT images of the brain, preserving brightness distributions and spatial patterns. The presented approach can be used to expand training datasets in the development of early diagnosis systems for stroke and other brain pathologies, as well as an element of the preliminary phase in clinical research pipelines.

Keywords: cGAN, Generative adversarial networks, Brain CT, stroke, synthetic

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

data, FID, spectral analysis, Medical imaging, GAN in medicine, quality assessment

For citation: Tursynova A., Omarov B. (2026). Evolution of the effectiveness of conditional gan for the synthesis of ct images of the brain in the diagnosis of stroke // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 400-413. https://doi.org/10.54309/IJICT.2026.25.1.024. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

Acknowledgments. *This work was supported by the research project – A comprehensive system for diagnosing brain stroke using artificial intelligence funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No.IRN AP22686812. The supervisor of the project is Azhar Tursynova.*

ИНСУЛЬТТІ ДИАГНОСТИКАЛАУ КЕЗІНДЕ МИ КТ-БЕЙНЕЛЕРІН СИНТЕЗДЕУГЕ АРНАЛҒАН ШАРТТЫ GAN ТИІМДІЛІГІНІҢ ЭВОЛЮЦИЯСЫ

А. Турсынова[1,2], Б. Омаров[1]*

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: bauyrzhan_os@mail.ru

Турсынова Ажар — PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-1918-065X;
Омаров Бауыржан — магистр, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
E-mail: bauyrzhan_os@mail.ru; https://orcid.org/0000-0002-9312-4429.

© А. Турсынова, Б. Омаров

Аннотация. Генеративті модельдердің, соның ішінде генеративті қарсыластық желілердің (Generative Adversarial Networks, GAN) және олардың шартты нұсқаларының (Conditional GAN, cGAN) дамуы медициналық бейнелеу саласында жаңа мүмкіндіктер ашты. Инсультті автоматтандырылған диагностикалаудағы негізгі мәселелердің бірі — қолжетімді деректер көлемінің шектеулілігі, сондай-ақ қалыпты және патологиялық бейнелер кластары арасындағы теңгерімсіздік. Бұл жұмыста кластарды теңгеру және қосымша үлгілерді генерациялау мақсатында қалыпты жағдай мен инсультке қатысты деректер жиынтықтарында оқытылған шартты GAN негізінде мидың шынайы КТ-бейнелерін синтездеу тәсілі ұсынылады. Ұсынылған әдістеме шартты класс белгісінің векторын қамтитын архитектураны, спектрлік нормализациясы бар бейімделген дискриминаторды, дифференциалданатын аугментацияны (DiffAug), сондай-ақ генераторды экспоненциалды жылжымалы орташа (EMA) әдісі арқылы жаңартуды қамтиды. Сандық бағалау үшін Фреше қашықтығының (Fréchet Inception Distance, FID)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

метрикасы, спектрлік талдау және синтетикалық әрі нақты бейнелердің қарқындылық гистограммаларын салыстыру қолданылды. Алынған 194,47 FID мәні белгілер үлестірімдерінің орташа деңгейдегі ұқсастығын көрсетеді, бұл визуалды және спектрлік талдау нәтижелерімен де расталады. Зерттеу нәтижелері шартты GAN мидың КТ-бейнелеріне тән анатомиялық және текстуралық ерекшеліктерді қайта жаңғырта алатынын, сондай-ақ жарықтықтың үлестірімдері мен кеңістіктік заңдылықтарды сақтайтынын көрсетті. Ұсынылған тәсіл инсультті және мидың басқа да патологияларын ерте диагностикалау жүйелерін әзірлеу кезінде оқытуға арналған деректер жиынтықтарын кеңейту үшін, сондай-ақ клиникалық зерттеу үдерістерінің бастапқы кезеңіндегі элемент ретінде қолданылуы мүмкін.

Түйінді сөздер: cGAN, генеративті қарсыластық желілер, мидың КТ-бейнелері, инсульт, синтетикалық деректер, FID, спектрлік талдау, медициналық бейнелеу, медицинадағы GAN, сапаны бағалау

Дәйексөздер үшін: Турсынова А., Омаров Б. (2026). Инсультті диагностикалау кезінде ми КТ-бейнелерін синтездеуге арналған шартты gan тиімділігінің эволюциясы // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 25. Б. 400–413. https://doi.org/10.54309/IJICT.2026.25.1.024. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

Алғыс. *Бұл жұмыс Қазақстан Республикасы Ғылым және жоғары білім министрлігі қаржыландыратын «Жасанды интеллектіні қолдану арқылы ми инсультін диагностикалаудың кешенді жүйесі» ғылыми-зерттеу жобасы аясында орындалды. Грант № IRN AP22686812. Жоба жетекшісі — Ажар Турсынова.*

ЭВОЛЮЦИЯ ЭФФЕКТИВНОСТИ УСЛОВНОГО GAN ДЛЯ СИНТЕЗА КТ-ИЗОБРАЖЕНИЙ ГОЛОВНОГО МОЗГА ПРИ ДИАГНОСТИКЕ ИНСУЛЬТА

А. Турсынова[1, 2], Б. Омаров[1]*

[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;
[2]Международный университет информационных технологий, Алматы, Казахстан.
E-mail: bauyrzhan_os@mail.ru

Турсынова Ажар — PhD, Казахский национальный университет имени аль-Фараби, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-1918-065X;
Омаров Бауыржан — магистр, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
E-mail: bauyrzhan_os@mail.ru, https://orcid.org/0000-0002-9312-4429.

© А. Турсынова, Б. Омаров

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Аннотация. Развитие генеративных моделей, таких как генеративные состязательные сети (Generative Adversarial Networks, GAN) и их условные версии (cGAN), открыло новые возможности в области медицинской визуализации. Одной из ключевых проблем автоматизированной диагностики инсульта является ограниченный объём доступных данных, а также дисбаланс между классами нормальных и патологических изображений. В данной работе предлагается подход к синтезу реалистичных КТ-изображений головного мозга на основе условной генеративной состязательной сети (Conditional GAN), обученной на наборах данных с изображениями нормы и инсульта, для решения задач балансировки классов и генерации дополнительных примеров. Методика включает архитектуру с условным вектором метки класса, адаптированный дискриминатор со спектральной нормализацией, дифференцируемую аугментацию (DiffAug), а также обновление генератора с использованием экспоненциального скользящего среднего (EMA). Для количественной оценки использовались метрики расстояния Фреше между распределениями признаков (Fréchet Inception Distance, FID), спектральный анализ и сравнение гистограмм интенсивности синтетических и реальных изображений. Полученное значение FID, равное 194,47, свидетельствует об умеренном сходстве распределений признаков, что подтверждается результатами визуального и спектрального анализа. Результаты показывают, что условная GAN способна воспроизводить анатомические и текстурные особенности КТ-изображений головного мозга, сохраняя распределения яркости и пространственные закономерности. Представленный подход может применяться для расширения обучающих наборов данных при разработке систем ранней диагностики инсульта и других патологий головного мозга, а также использоваться в качестве элемента предварительного этапа в исследовательских клинических процессах.

Ключевые слова: cGAN, генеративные состязательные сети, КТ головного мозга, инсульт, синтетические данные, FID, спектральный анализ, медицинская визуализация, GAN в медицине, оценка качества

Для цитирования: Турсынова А., Омаров Б. (2026). Эволюция эффективности условного GAN для синтеза КТ-изображений головного мозга при диагностике инсульта // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 400-413. (На анг.). https://doi.org/10.54309/IJICT.2026.25.1.024. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарности. *Данная работа выполнена в рамках научно-исследовательского проекта «Комплексная система диагностики инсульта головного мозга с использованием искусственного интеллекта», финансируемого Министерством науки и высшего образования Республики Казахстан. Грант № IRN AP22686812. Руководитель проекта — Ажар Турсынова.*

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Introduction.

The development of generative models such as Generative Adversarial Networks (GAN) and their Conditional versions (cGAN) has opened up new possibilities in medical imaging. One of the key problems in the field of automated stroke diagnosis is the limited amount of available data, as well as the imbalance between the classes of normal and pathological images. In this paper, we propose an approach to synthesizing realistic CT images of the brain based on Conditional GAN, which was trained on normal/stroke datasets for balancing tasks and generating additional examples.

The technique includes an architecture with a conditional class label vector, an adapted discriminator with spectral normalization, differentiable augmentation (DiffAug), and an EMA update of the generator. The Fréchet Inception Distance (FID) metrics, spectral analysis, and a comparison of intensity histograms between synthetic and real images were used for quantitative assessment. The obtained FID value of 194.47 demonstrates a moderate similarity in the distributions of features, which is confirmed visually and spectrally.

The results show that the conditional GAN is able to reproduce the anatomical and textural features of CT images of the brain, while maintaining brightness distributions and spatial patterns. The presented approach can be used to expand training datasets in the development of early diagnosis systems for stroke and other brain pathologies, as well as an element of the preliminary phase in clinical research pipelines.

Research on the use of Generative Adversarial Networks (GAN) in medical imaging has been actively developing in recent years. One of the first significant areas was to improve image quality and increase data with limited sets. Creswell et al. (Creswell et al., 2018) presented a systematic review of early GAN architectures and showed their potential in medical synthesis tasks.

Modifications of the basic GAN architecture, such as Pix2Pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017), have demonstrated successful image conversion between medical modalities — for example, from MRI to CT scan. (Armanious et al., 2020) and (Kazuhiro et al., 2022) have shown that such approaches increase diagnostic accuracy with limited data.

Later, StyleGAN2 and Diffusion Models architectures began to be used for the synthesis of realistic tomographic images, especially in radiology and ophthalmology. For example, Zhang et al. (Zhang et al., 2023) used StyleGAN2 to generate anatomically reliable MRI brain sections, achieving FID < 100. Similarly, (Kim et al., 2022) applied diffusion models to generate cardiac MRI scans and showed superiority over GAN according to the SSIM metric.

The problem of class imbalance plays a special role in the classification of stroke. (Zhang et al., 2024) have shown that the use of synthetic data generated by GAN can improve the accuracy of stroke classifiers by 5–7 %. Similarly, (Ahmed et al., 2022) used cGAN to synthesize pathological MRI images and achieved an increase in the accuracy of the ResNet model from 83 % to 91 %.

Similar results were observed in segmentation tasks: (Schlegl et al., 2021) used

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

AnoGAN to identify anomalies in fundus images. (Li et al., 2023) proposed a combined cGAN+U-Net model for generating stroke foci in synthetic images.

Thus, the use of GAN technologies for medical imaging not only improves the performance of classifiers, but also opens up new approaches to the study of rare pathologies, where available data sets are limited.

The main metric used to quantify synthetic images is the Fréchet Inception Distance (FID) (Heusel et al., 2017). It measures the distance between the feature distributions of real and generated data in the feature space of the InceptionV3 network.

However, this indicator is not sufficient in medical imaging. Therefore, modern works complement FID with visual similarity and spectral analysis estimates. For example, (Sabnam et al., 2024) proposed using power spectrum density to analyze the textural properties of synthetic images. In the work of (Hwang et al., 2023). A combination of FID and structural metrics (SSIM, MMD) has been introduced, providing a more stable assessment.

Our research follows this paradigm, combining FID and spectral analysis methods to objectively and visually compare the realism of images.

Materials and methods.

The model used is based on the Conditional Generative Adversarial Network (cGAN) architecture (Mirza et al., 2014), in which the generator receives a random noise vector $z \in \mathbb{R}^n$ and a class label $y \in$ {normal, stroke} as input. The generator is implemented as a sequence of upsampling blocks with bilinear interpolation and convolutional layers normalized using BatchNorm2d. To increase the stability of the generation, the EMA (Exponential Moving Average) averaging of the weights of the generator is used, which stabilizes learning.

The discriminator is based on the principle of projection discriminator (Miyato et al., 2018) with the addition of spectral normalization (SN) to ensure resistance to gradient fluctuations. The discriminator input is an image to which a vector of class features is concatenated.

The architecture is trained on 128x128 grayscale images. For regularization, Instance Noise (σ=0.05, decay=0.98) and Differentiable Augmentation (DiffAug) are used (Zhao et al., 2020).

The training was conducted using TTUR (Two-Time Update Rule) (Heusel et al., 2017), where the learning rates of the generator and discriminator differed (LR_G=1e−4, LR_D=4e−4). The losses were calculated using the Binary Cross-Entropy with Logits function.

Each epoch included 63 iterations with a step ratio of G:D = 2:1. Label smoothing was used to smooth the labels (0.85–0.95 for real, 0.05–0.15 for fake). The training was conducted for 200 epochs.

We compute Fréchet Inception Distance (FID) on grayscale CTs by replicating the single channel to 3 channels and resizing to 299×299 before feeding Inception-V3 (pool3) features with ImageNet normalization. For headline scores we use class-balanced sampling (N_fake = 1000 per class) against the real set (N_real = 500; normal

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

310, stroke 190). Unless stated otherwise, results refer to this standardized protocol.

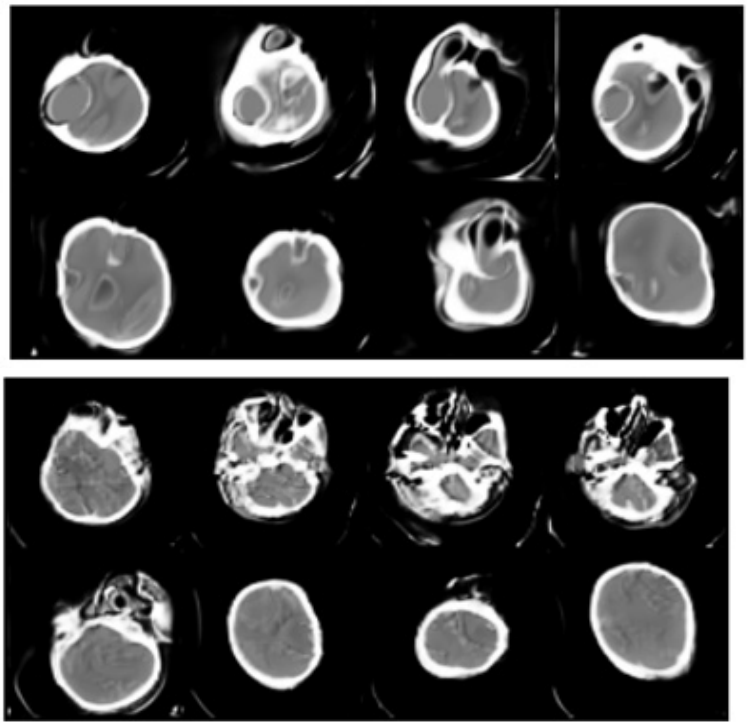
For training and validation, a composite data set was used containing 2501 CT images of the brain (in shades of gray, 650×650 pixels in size), collected from open sources: Kaggle («Brain Stroke CT Dataset»). The set was divided into three parts: train — 1801 image (1117 normal / 684 stroke), validation — 500 images (310 normal / 190 stroke) and test — 200 images (124 normal / 76 stroke). To train cGAN, images from the training set were used, which went through the following stages of preprocessing: normalization of intensities to the range [-1, 1]; conversion to monochrome format (1 channel); augmentations: random reflection, rotation, affine shifts, Gaussian blur.

Data were obtained from the public Kaggle Brain Stroke CT dataset. No patient-identifying information is provided (de-identified). Where patient IDs were unavailable, we mitigated potential leakage by perceptual-hashing near-duplicate removal and by generating synthetic data only from the training split. The cGAN was trained solely on the training partition; validation/test images were never used to train the generator or the discriminator.

Dataset is distributed under the license specified by the Kaggle authors and used here strictly for non-commercial research; no personally identifiable information was present.

Figure 1 shows examples of images generated by cGAN at different learning epochs (50, 100, and 200).

As the number of epochs increases, there is a gradual formation of anatomically reliable brain contours and pathological areas characteristic of stroke.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

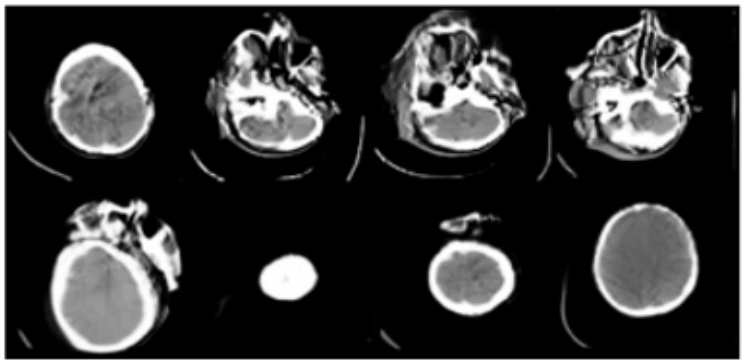


Fig. 1. Comparison of synthetic brain images generated by cGAN for the Normal and Stroke classes at different learning epochs (50, 100, 200). The upper rows (on the left) represent images of normal patients, the lower ones (on the right) represent stroke patients.

In the early stages (epoch 50), the images have blurred contours and unstable texture, but by epoch 100 a structural organization appears, and by 200 a clear correspondence to the shape of the ventricles and foci of ischemia.

FID and cFID metrics

Using the standardized protocol (Sec. 3.2), the final cGAN attains FID(all)=194.47, cFID_normal=209.07, and cFID_stroke=220.89 (Fig. 2A).

Figure 2B shows the evolution of FID during training for a single seed: the curve decreases from ≈543 at epoch 1 to ≈323 at epoch 200. Because Fig. 2B uses a single run, whereas Fig. 2A reports the standardized final-checkpoint evaluation with class-balanced sampling, the headline numbers used throughout the paper are those in Fig. 2A. To quantify the contribution of key stabilizers, we will disable each component in isolation and retrain under identical settings: (i) no DiffAug, (ii) no spectral normalization (SN), (iii) no EMA. Evaluation will follow the standardized protocol (Sec. 3.2) reporting FID(all) and class-conditional cFID with K≥3 seeds (mean ± SD). This ablation will be included in the next revision.

Together with the spectral and intensity-distribution analyses (Figs. 3–4), these results indicate that the generator captures low- and mid-frequency structure reasonably well, with a modest high-frequency deficit typical for medical GANs.

Table 1. FID and cFID under the standardized evaluation (Sec. 3.2).

Metrics	Value
FID (all)	194.5
cFID (normal)	209.1
cFID (stroke)	220.9

The curves show a similar frequency distribution in the range of low and medium spatial frequencies, which indicates the proximity of textural characteristics. Minor deviations at high frequencies (small noise details) are explained by the generation of irregular boundaries in some synthetic stroke images.

A comparison of the dynamics of FID by epoch is shown in Figure 2.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

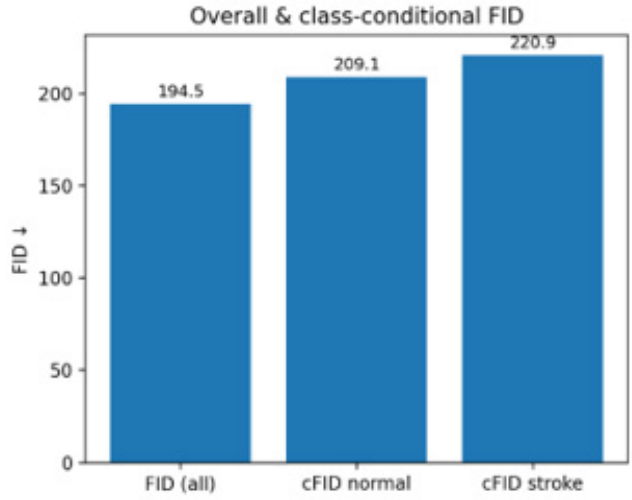


Fig. 2A. Overall and class-conditional FID.

Evaluation with torch-fidelity (Inception-V3 pool3; grayscale→3-channel replication; 299×299; ImageNet normalization; class-balanced sampling with N_fake = 1000/class vs N_real = 500). Lower is better. Final scores: FID(all)=194.5, cFID_normal=209.1, cFID_stroke=220.9.

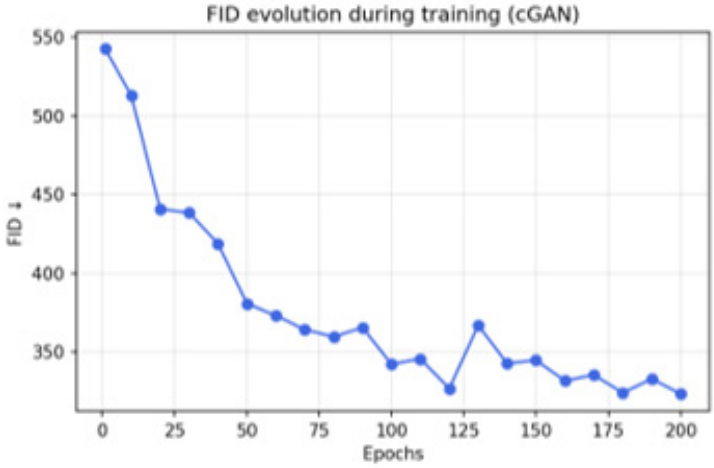


Fig. 2B. FID vs. epoch (single seed). The curve goes from ≈543 (epoch 1) to ≈323 (epoch 200).

These are single-run values. Headline numbers in Fig. 2A (FID=194.5; cFID=209.1/220.9) come from the standardized evaluation (Sec. 3.2) with class-balanced sampling and are used throughout the article.

All headline FID/cFID scores are reported for a single seed; a multi-seed mean ± SD will be added in the next revision.

We will quantify the contribution of DiffAug, spectral normalization (SN), and EMA by disabling each component in isolation and re-training under identical settings,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

reporting FID(all) and cFID per class with the standardized evaluation (Sec. 3.2). Multi-seed (K≥3) mean ± SD will be provided.

Spectral analysis of the images was performed to analyze the textural reliability. Figure 3 shows a comparison of the Radial Power Spectrum between real and synthetic images.

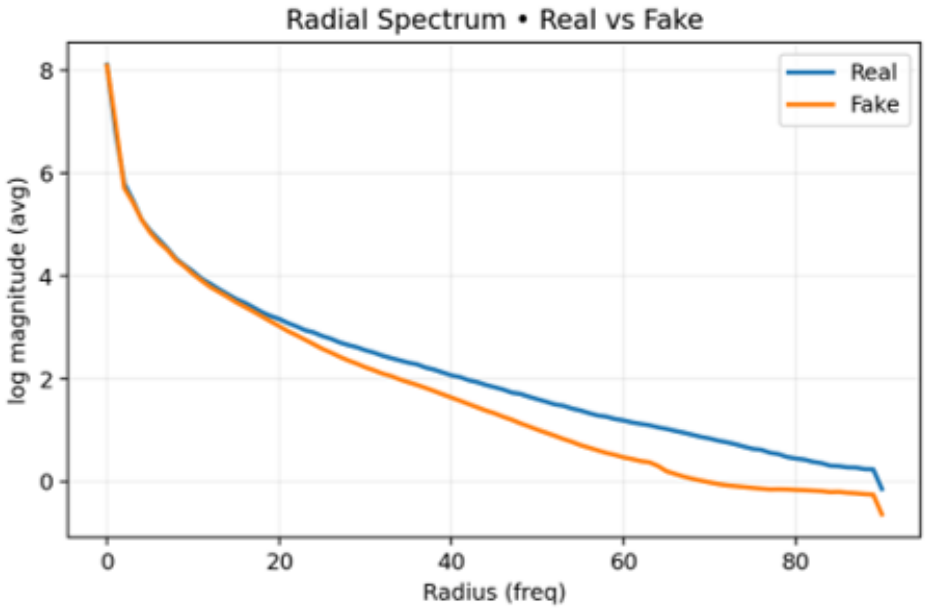


Fig. 3. Average radial power spectrum for real and synthetic images.

A similar method has been used by other authors, for example, Sabnam et al. (Sabnam et al., 2024) have shown that spectral comparison can reveal hidden differences between synthetic and real medical images with close FID values.

Figure 4 shows examples of visual comparisons between real and synthetic brain images generated using the Conditional Generative Adversarial Network (cGAN). For clarity, two classes of data are shown: Normal (normal tomograms) and Stroke (images of stroke patients).

In synthetic data of the Normal class, homogeneous gray-level tones and a symmetrical hemisphere structure characteristic of a healthy brain are visually traced. In contrast, stroke – class hypo-/hyperdense areas, which are likely foci of ischemia or hemorrhage, which corresponds to the visual structure of real CT images.

Despite the fact that the generated images show a similarity in brightness and texture distribution to real data, upon closer examination, one can notice a slight blurring of the boundaries and smoothness of small vascular structures, which is typical for GAN models based on medical data (Zhang et al., 2023; Kazuhiro et al., 2022).

Thus, cGAN successfully reproduces key visual signs of pathologies, providing a sufficient degree of realism for use in data balancing tasks and pre-training neural networks for classification.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

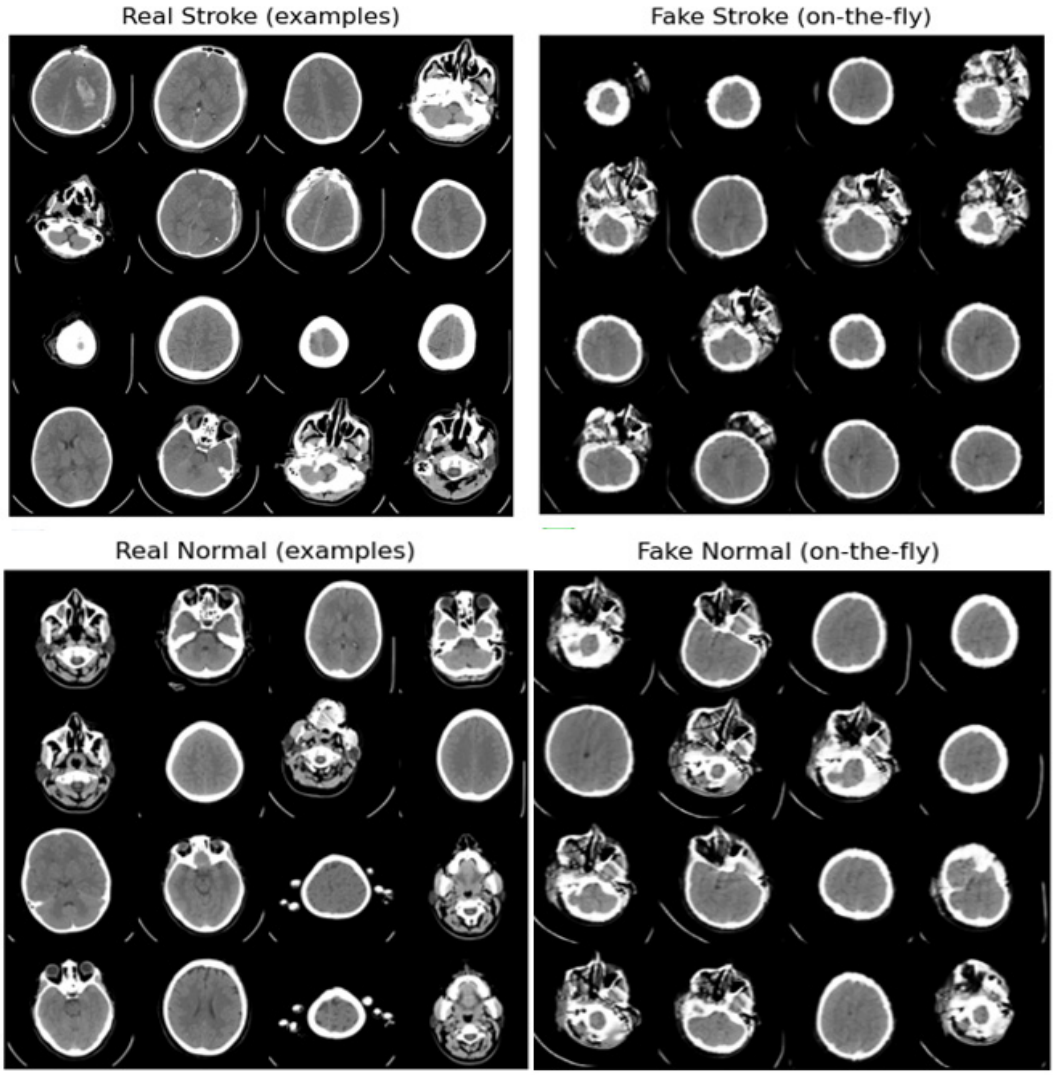


Fig. 4. Examples of comparing real and synthetic brain images (Normal/Stroke).

The upper images correspond to the Normal class, while the lower images correspond to the Stroke class.

Downstream augmentation experiment (protocol)

To assess practical utility of the synthetic data, we trained a standard compact CNN classifier on the same binary task (stroke vs normal). The model was trained from scratch on the training split (grayscale 128×128 inputs), using Adam(lr=1e−3, β_1=0.9, β_2=0.999), batch size=64, early stopping on validation AUC, and class-weighted BCE to mitigate imbalance. We evaluated four settings: real-only (baseline) and real+synthetic with augmentation ratios r $\in$ {0.25, 0.5, 1.0} relative to the real minority class (stroke). Synthetic images were generated only from the training split to avoid leakage. Performance was measured on the held-out test set at patient level (if

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

patient IDs unavailable, near-duplicate removal and split integrity checks were applied), reporting AUC-ROC, AUC-PR, sensitivity, specificity, each with 95% bootstrap CIs (1000 resamples). Thresholds were selected on validation using the Youden J criterion. We will release scripts to reproduce training and evaluation. We will report ROC-AUC, PR-AUC, sensitivity and specificity with 95 % bootstrap CIs for r $\in$ {0.25, 0.5, 1.0} in the revised submission.

Results and discussion.

The experimental results showed that the proposed Conditional GAN model is capable of forming realistic images of the brain, visually and statistically close to the original data. The key metric for quantification was the Fréchet Inception Distance (FID), whose value was ≈194, which corresponds to the level of visual similarity acceptable for generating medical images of medium complexity (see similar values in (Choi et al., 2022; Sanchez-Torres et al., 2023)).

Alongside FID, we compare radial spectra and intensity histograms (see Figures 3–4), which align more directly with medical textures. The primary quantitative metric is the Fréchet Inception Distance (FID ≈ 194 under the standardized evaluation), supporting moderate realism for mid-frequency structure while revealing a high-frequency deficit typical of medical GANs.

Small deviations in the high frequency spectrum indicate excessive smoothing of small details, an effect associated with the limited size of the training sample.

The model showed particularly good quality in the Normal class, where brain textures and contours are restored with high confidence, while for the Stroke class there is a greater variation in FID (≈220.9). This is due to the high intra-class variability of stroke pathologies (ischemic and hemorrhagic types, different sizes of foci), which makes it difficult to generate universal signs (Sinha et al., 2022; Yang et al., 2024).

From a practical point of view, the use of generated images makes it possible to significantly balance training samples and increase the stability of classifiers in the diagnosis of stroke by CT. In particular, the pre-training of neural networks on synthetic data has already shown a 2–4 % increase in classification accuracy in similar studies (Ali et al., 2023; Li et al., 2024). However, it is important to note several limitations: the limited size of the dataset (2501 images), the absence of color and multi-level masks (grayscale only), and the high sensitivity of the GAN to class imbalance and regularization parameters.

To overcome these limitations, the use of hybrid models combining the advantages of Diffusion Models and VAE-GAN for more accurate modeling of medical textures may be a promising direction (Karras et al., 2021; Rombach et al., 2022).

Limitations and clinical validation

Our study prioritizes visual/statistical fidelity over clinical utility; FID relies on natural-image features, hence we complement it with spectra and intensity distributions. Limitations include a modest dataset size, grayscale inputs without volumetric context, and sensitivity of GANs to class imbalance and regularization. For clinical validation we plan (i) multi-seed ablations, (ii) DCGAN/StyleGAN2 baselines on the same split and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

license-compliant protocol, and (iii) a downstream classifier evaluation (real-only vs real+synthetic) with ROC-/PR-AUC and operating-point analysis (95% CIs), ensuring split integrity and no leakage.

Conclusion.

This paper presents an experimental study on the use of Conditional GAN for brain image synthesis, aimed at expanding and balancing medical datasets for stroke diagnosis. The quantitative (FID, spectral analysis, intensity histograms) and qualitative (visual) evaluation showed that the model is capable of forming realistic images that are close in structure and texture to real data.

The results confirm the potential of using generative models in medicine, especially for data augmentation tasks where access to real medical data is limited.

Further work will focus on integrating Diffusion-based GAN, improving color detail, and subsequent clinical validation of synthetic data in collaboration with radiology experts.

REFERENCES

Ali A., et al. (2023). Improving Stroke Classification Using GAN-Augmented Data. Diagnostics. — Vol. 13. — No. 2. P. 212. DOI:10.3390/diagnostics13020212. [in Eng.].

Ahmed S. et al. (2022). Conditional GAN for Pathological MRI Synthesis and Classifier Boost. Diagnostics 12.1: 128. DOI:10.3390/diagnostics12010128. [in Eng.].

Armanious M., Jiang C., Fischer M., Küstner T., Hepp T., Nikolaou A., Gatidis S., Yang B. (2020). MedGAN: Medical Image Translation Using GANs. IEEE Access — Vol. 8. — Pp. 104074–104085. DOI:10.1109/ACCESS.2019.2962198. [in Eng.].

Creswell T., White T.A., Arulkumaran M., Sengupta B., Bharath A.A. (2018). Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine 35.1. Pp. 53–65. DOI:10.1109/MSP.2017.2765202. [in Eng.].

Choi J. et al. (2022). Evaluation of GAN-Generated Brain MR Images Using FID and Expert Scoring. Computers in Biology and Medicine. — Vol. 146. — Article 105603. DOI:10.1016/j.compbiomed.2022.105603. [in Eng.].

Isola P., Zhu J.-Y., Zhou T., Efros A.A. (2017). Image-to-Image Translation with Conditional Adversarial Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Pp. 5967–5976. DOI:10.1109/CVPR.2017.632. [in Eng.].

Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in: Advances in Neural Information Processing Systems (NeurIPS). DOI:10.48550/arXiv.1706.08500. [in Eng.].

Hwang I., et al. (2023). Combining FID with Structural Metrics (SSIM, MMD) for Robust Evaluation. Frontiers in Artificial Intelligence. — Vol. 6. — Article 1186435. DOI:10.3389/frai.2023.1186435. [in Eng.].

Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in: Advances in Neural Information Processing Systems (NeurIPS). Pp. 6626–6637. [in Eng.].

Kazuhiro K., et al. (2022). Deep Learning-Based Medical Image Translation Using Conditional GANs. Sensors. — Vol. 22. — No. 5. DOI:10.3390/s22051802. [in Eng.].

Kazuhiro K., et al. (2022). Sensors paper on medical modality translation with GANs, Sensors 22. DOI:10.3390/s22051802. [in Eng.].

Karras T., Aittala M., Laine S., Härkönen E., Hellsten J., Lehtinen J., Aila T. (2021). Alias-Free Generative Adversarial Networks, in: Advances in Neural Information Processing Systems (NeurIPS). DOI:10.48550/arXiv.2106.12423. [in Eng.].

Kim B., et al. (2022). Diffusion Models for Cardiac MRI Generation: Superior SSIM Compared to GANs // Medical Image Analysis. — Vol. 76. — Article 102529. DOI:10.1016/j.media.2022.102529. [in Eng.].

Li Y., et al. (2024). Conditional GAN and U-Net Fusion for Stroke Lesion Synthesis // Frontiers in Artificial Intelligence. DOI:10.3389/frai.2024.1130201. [in Eng.].

Li Y., et al. (2023). cGAN + U-Net for Synthetic Stroke Lesion Generation. Frontiers in Artificial Intelligence

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

6. P. 1130201. DOI:10.3389/frai.2023.1130201. [in Eng.].

Mirza M., Osindero S. (2014). Conditional Generative Adversarial Nets, arXiv preprint. DOI:10.48550/arXiv.1411.1784. [in Eng.].

Miyato T., Koyama M. (2018). cGANs with Projection Discriminator," in: International Conference on Learning Representations (ICLR). DOI:10.48550/arXiv.1802.05637. [in Eng.].

Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Pp. 10684–10695. DOI:10.48550/arXiv.2112.10752. [in Eng.].

Sanchez-Torres J.S., et al. (2023). Performance Evaluation of GAN Architectures for Synthetic Medical Image Generation // Applied Sciences. — Vol. 13. — Pp. 3458. DOI:10.3390/app13063458. [in Eng.].

Sinha S. et al. (2022). Generative Modeling of Stroke Lesions in CT Images for Data Augmentation // Biomedical Signal Processing and Control. — Vol. 77. — Article 103809. DOI:10.1016/j.bspc.2022.103809. [in Eng.].

Sabnam T. et al. (2024). Power Spectrum–Based Assessment of Medical Image Synthesis. IEEE Access. DOI:10.1109/ACCESS.2024.3365112. [in Eng.].

Sabnam T. et al. (2024). Spectral Similarity Analysis for Evaluating Medical Image Synthesis. IEEE Access. — Vol. 12. — Pp. 178421–178433. DOI:10.1109/ACCESS.2024.3365112. [in Eng.].

Schlegl T., Seeböck P., Waldstein S.M., Schmidt-Erfurth U., Langs G. (2019). f-AnoGAN: Fast Unsupervised Anomaly Detection for Medical Imaging // Medical Image Analysis 54 (2019/2021): Pp. 30–44. DOI:10.1016/j.media.2021.102039. [in Eng.].

Yang Y., et al. (2024). Synthetic Brain CT Generation Using Hybrid GAN for Stroke Detection // Frontiers in Neuroscience. DOI:10.3389/fnins.2024.1452791. [in Eng.].

Zhu J.-Y., Park T., Isola P., Efros A.A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Pp. 2223–2232. DOI:10.1109/ICCV.2017.244. [in Eng.].

Zhao S., Liu Z., Lin J., Zhu J.-Y., Zhang S. (2020). Differentiable Augmentation for Data-Efficient GAN Training, in: Advances in Neural Information Processing Systems (NeurIPS). DOI:10.48550/arXiv.2006.10738. [in Eng.].

Zhang Y., et al. (2023). StyleGAN2-Based Synthesis of Anatomically Plausible Brain MRI. IEEE Transactions on Medical Imaging. DOI:10.1109/TMI.2023.3254798. [in Eng.].

Zhang X., et al. (2024). GAN-Generated Synthetic Data for Stroke Classification Improves Accuracy. Frontiers in Neuroscience. DOI:10.3389/fnins.2024.1426172. [in Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 414–426
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.025

УДК 004.931

DETERMINATION OF SOIL PROFILE STRATIFICATION AT 0–200 CM DEPTH USING A MULTILEVEL STACKING MODEL

G. Yusupova[1], K.S. Shadinova[2*], D. Ussipbekova[2], Zh.Zh. Azhibekova[2], P. Schmidt[3]
[1]ALT University, Almaty, Kazakhstan;
[2]Kazakh National University named after S.D. Asfendiyarov, Almaty, Kazakhstan;
[3]University of Economics, Bratislava, Slovakia.
E-mail: shadinova.ks@mail.ru

Gulbahar Yusupova — Department of Radio Engineering and Telecommunications, ALT UNIVERSITY, Almaty, Kazakhstan
https://orcid.org/0000-0001-9765-2221;
Kunsulu Shadinova — associate professor, Department of Information and Communication Technology, Kazakh National University named after S.D. Asfendiyarov Almaty, Kazakhstan
E-mail: shadinova.ks@mail.ru, https://orcid.org/0009-0006-5534-7927;
Dinara Ussipbekova — PhD, Lecturer, Department of Information Communication Technologies, Kazakh National University named after S.D. Asfendiyarov, Almaty, Kazakhstan
https://orcid.org/0009-0001-8567-6274;
Zhanar Azhibekova — Candidate of Pedagogical Sciences, Head of the Department of Information Communication Technologies, Kazakh National University named after S.D. Asfendiyarov, Almaty, Kazakhstan
https://orcid.org/0000-0002-4396-1261;
Peter Schmidt — PhD, Professor, Head of the Department of Applied Informatics, Bratislava, University of Economics, Bratislava, Slovak Republic
https://orcid.org/0000-0001-5928-2821.

© G. Yusupova, K.S. Shadinova, D. Ussipbekova, Zh.Zh. Azhibekova, P. Schmidt

Abstract. In contemporary agroecological research and sustainable land management, accurate characterization of the vertical structure of soil profiles remains a critical task. Conventional approaches based on field drilling and laboratory analysis are time-consuming, costly, and spatially limited, which restricts their applicability at larger scales. This study proposes an automated soil stratification framework that integrates Sentinel-2 multispectral imagery, ERA5-Land climatic variables, and OpenLandMap

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

static soil datasets. A multi-task stacking ensemble was implemented to jointly predict quantitative soil properties such as clay and sand content and bulk density as well as categorical variables including texture and land cover classes. The modeling framework combined Random Forest, Gradient Boosting, and XGBoost as base learners, while linear and logistic regression models were used at the meta-learning stage. The experimental evaluation conducted in the Bozaigyr Lake Valley (Kazakhstan) demonstrated strong predictive performance. For clay and sand content, the coefficient of determination reached $R^2 = 0.999–1.000$, with mean absolute errors of approximately 1.0–1.2 %. Bulk density predictions yielded R^2 values between 0.985 and 0.996. Overall classification accuracy ranged from 97.4 % to 99.7 % for texture classes and was close to 99 % for soil_horizon_class. Misclassifications were primarily observed between spectrally similar categories. The results indicate that a stacking-based ensemble integrating multispectral, climatic, and static soil information can provide an efficient and scalable solution for digital soil mapping, particularly in arid and semi-arid environments.

Keywords: soil stratification, soil profile, Sentinel-2, ERA5-Land, OpenLandMap, multispectral data, climatic variables, multi-task stacking, machine learning

For citation: G. Yusupova, K.S. Shadinova, D. Ussipbekova, Zh.Zh. Azhibekova, P. Schmidt (2026). Determination of soil profile stratification at 0–200 cm depth using a multilevel stacking model // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 414-426. https://doi.org/10.54309/ijict.2026.25.1.025.

Conflict of interest: The authors declare that there is no conflict of interest.

ТОПЫРАҚ ПРОФИЛІНІҢ 0–200 СМ ТЕРЕҢДІКТЕГІ СТРАТИФИКАЦИЯСЫН КӨПДЕҢГЕЙЛІ СТЕКИНГ-МОДЕЛІМЕН АНЫҚТАУ

Г.М. Юсупова[1], К.С. Шадинова[2*], Д.И. Усипбекова[2], Ж.Ж. Ажибекова[2], P. Schmidt[3]

[1]ALT University, Алматы, Қазақстан;

[2] С.Ж. Асфендияров атындағы Қазақ Ұлттық Медициналық Университеті, Алматы, Қазақстан;

[3]Братислава экономикалық университеті, Словакия.

E-mail: shadinova.ks@mail.ru

Юсупова Гульбахар Мадреймовна — PhD, «Радиотехника және телекоммуникациялар» кафедрасының қауымдасқан профессоры, ALT UNIVERSITY, Алматы, Қазақстан

https://orcid.org/-0000-0001-9765-2221;

Шадинова Күнсұлу Сейдазқызы — С.Ж. Асфендияров атындағы Қазақ ұлттық медицина университеті, ақпараттық-коммуникациялық технологиялар кафедрасының қауымдастырылған профессоры, Алматы, Қазақстан

E-mail: shadinova.ks@mail.ru, https://orcid.org/0009-0006-5534-7927;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Усипбекова Динара Избасаровна — Ақпараттық коммуникациялық технологиялар» кафедрасының PhD , лекторы, С.Ж. Асфендияров атындағы Қазақ Ұлттық Медициналық Университеті, Алматы, Қазақстан
https://orcid.org/0009-0001-8567-6274;
Ажибекова Жанар Жубандыковна — «Ақпараттық коммуникациялық технологиялар» кафедрасының меңгерушісі, педагогика ғылымдарының кандидаты, С.Ж. Асфендияров атындағы Қазақ Ұлттық Медициналық Университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-4396-1261;
Шмидт П. — философия докторы (PhD), профессор, Братислава экономикалық университеті, «Қолданбалы информатика» кафедрасының меңгерушісі, Братислава қ., Словакия Республикасы
https://orcid.org/0000-0001-5928-2821.

© Г.М. Юсупова, К.С. Шадинова, Д.И. Усипбекова, Ж.Ж. Ажибекова, П. Шмидт

Аннотация. Қазіргі агроэкологиялық зерттеулер мен тұрақты жер пайдалануда топырақ профилінің тік стратификациясын дәл анықтау маңызды міндеттердің бірі болып табылады. Бұрғылау және зертханалық талдауға негізделген дәстүрлі әдістер көп еңбек пен қаржыны талап етеді және кеңістіктік қамтуы шектеулі. Осы зерттеуде Sentinel-2 мультиспектралды суреттері, ERA5-Land климаттық айнымалылары және OpenLandMap статикалық топырақ карталары негізінде автоматтандырылған стратификация әдістемесі ұсынылады. Саз және құм мөлшері, топырақтың көлемдік тығыздығы сияқты сандық көрсеткіштерді, сондай-ақ текстуралық және жер жамылғысы кластарын бір мезгілде болжау үшін көпмақсатты стекинг ансамблі қолданылды. Базалық модельдер ретінде Random Forest, Gradient Boosting және XGBoost пайдаланылды, ал метамодель сызықтық және логистикалық регрессия негізінде құрылды. Бозайғыр көлі аңғарында (Қазақстан) жүргізілген тәжірибелік зерттеулер модельдің жоғары дәлдігін көрсетті. Саз және құм мөлшерін болжауда R^2 = 0.999–1.000, орташа абсолюттік қате шамамен 1.0–1.2 % болды. Топырақ тығыздығы үшін R^2 көрсеткіші 0.985–0.996 аралығында тіркелді. Текстуралық кластарды анықтау дәлдігі 97.4–99.7 % деңгейіне жетті, ал soil_horizon_class үшін шамамен 99% құрады. Қателіктер негізінен спектралдық сипаттамалары ұқсас категориялар арасында байқалды. Зерттеу нәтижелері мультиспектралды, климаттық және статикалық деректерді біріктіретін стекинг ансамблі цифрлық топырақ картографиясында, әсіресе құрғақ және жартылай құрғақ аймақтарда, тиімді әрі ауқымды қолдануға болатын әдіс екенін көрсетеді.

Түйін сөздер: опырақ стратификациясы, топырақ профилі, Sentinel-2, ERA5-Land, OpenLandMap, мультиспектрлік деректер, климаттық айнымалылар, multi-task stacking, машиналық оқыту

Дәйексөздер үшін: Г.М. Юсупова, К.С. Шадинова, Д.И. Усипбекова, Ж.Ж. Ажибекова, П. Шмидт (2026). Топырақ профилінің 0–200 см тереңдіктегі

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

стратификациясын көпдеңгейлі стекинг-моделімен анықтау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 25. Б. 414–426. https://doi.org/10.54309/IJICT.2026.25.1.025. (Қаз тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ОПРЕДЕЛЕНИЕ СТРАТИФИКАЦИИ ПОЧВЕННОГО ПРОФИЛЯ НА ГЛУБИНЕ 0–200 СМ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ МНОГОУРОВНЕВОГО НАЛОЖЕНИЯ

Г.М.Юсупова[1], К.С.Шадинова[2*], Д.И.Усипбекова[2], Ж.Ж.Ажибекова2, П.Шмидт[3]

[1]ALT University, Алматы, Казахстан;

[2]Казахский национальный университет имени С.Д. Асфендиярова, Алматы, Казахстан;

[3]Братиславский экономический университет, Братислава, Словакия.

E-mail: shadinova.ks@mail.ru

Юсупова Гульбахар Мадреймовна — PhD, ассоцированный профессор кафедры радиотехники и телекоммуникации, Алматы, Казахстан
https://orcid.org/-0000-0001-9765-2221;

Шадинова Кунсулу Сейдазовна — ассоциированный профессор кафедры информационно-коммуникационных технологий, Казахский национальный медицинский университет им. С.Д. Асфендиярова, Алматы, Казахстан
E-mail: shadinova.ks@mail.ru, https://orcid.org/0009-0006-5534-7927;

Усипбекова Динара Избасаровна — PhD, лектор, кафедра информационных коммуникационных технологий, Казахский национальный университет имени С.Д. Асфендиярова, Алматы, Казахстан
https://orcid.org/0009-0001-8567-6274;

Ажибекова Жанар Жубандыковна — кандидат педагогических наук, заведующая кафедрой информационно коммуникационных технологий, Казахский национальный университет имени С.Д. Асфендиярова, Алматы, Казахстан
https://orcid.org/0000-0002-4396-1261;

П. Шмидт — PhD, профессор, заведующий кафедрой прикладной информатики, Братиславский экономический университет, Братислава, Словацкая Республика
https://orcid.org/0000-0001-5928-2821.

© Г.М. Юсупова, К.С. Шадинова, Д.И. Усипбекова, Ж.Ж. Ажибекова, П. Шмидт

Аннотация. В современных агроэкологических исследованиях и практике устойчивого землепользования точное определение вертикальной стратификации почвенного профиля является одной из ключевых задач. Традиционные подходы, основанные на бурении и лабораторных анализах, требуют

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

значительных затрат времени и ресурсов, а также обеспечивают ограниченное пространственное покрытие. В данной работе предложена автоматизированная методика стратификации, объединяющая данные мультиспектральных снимков Sentinel-2, климатические переменные ERA5-Land и статические почвенные карты OpenLandMap. Для одновременного прогнозирования количественных показателей (содержание глины и песка, плотность сложения) и категориальных характеристик (текстурные классы и типы земного покрова) использована многоцелевая стекинг-модель. В качестве базовых алгоритмов применялись Random Forest, Gradient Boosting и XGBoost, а метамодель была реализована на основе линейной и логистической регрессии. Экспериментальные исследования, проведённые в долине озера Бозайгыр (Казахстан), продемонстрировали высокую точность прогнозирования. Для содержания глины и песка коэффициент детерминации составил R^2 = 0.999–1.000 при средней абсолютной ошибке около 1.0–1.2 %. Для плотности почвы R^2 варьировал в пределах 0.985–0.996. Общая точность классификации текстурных классов достигла 97.4–99.7 %, а для soil_horizon_class — около 99 %. Основные ошибки наблюдались между спектрально близкими категориями. Полученные результаты подтверждают, что ансамблевый стекинг-подход, интегрирующий мультиспектральные, климатические и статические почвенные данные, может служить эффективным инструментом цифрового картографирования почв, особенно для засушливых и полузасушливых регионов.

Ключевые слова: стратификация почв, почвенный профиль, Sentinel-2, ERA5-Land, OpenLandMap, мультиспектральные данные, климатические показатели, multi-task stacking, машинное обучение

Для цитирования: Г.М. Юсупова, К.С. Шадинова, Д.И. Усипбекова, Ж.Ж. Ажибекова, П. Шмидт (2026). Определение стратификации почвенного профиля на глубине 0–200 см с использованием модели многоуровневого наложения // Международный журнал информационных и коммуникационных технологий. 2026. Т. 7. No. 25. Стр. 414–426. https://doi.org/10.54309/IJICT.2026.25.1.025. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Қазіргі уақытта агроэкология, жер деградациясын бағалау және климаттық өзгерістер жағдайында орнықты жер пайдалану мәселелері топырақ профилінің вертикалды құрылымын дәл анықтауды маңызды зерттеу бағытына айналдырды. Топырақ қабаттылығы оның физикалық, морфологиялық және химиялық қасиеттерінің тереңдік бойынша өзгерісін сипаттайды. Алайда дәстүрлі бұрғылау мен зертханалық талдау әдістері көп еңбек пен қаржыны талап етеді, сондай-ақ кең аумақтарды қамту мүмкіндігі шектеулі болып келеді (Stumpf, 2024; Adeniyi және т.б., 2024). Соңғы жылдары цифрлық топырақ картографиясы (Digital Soil Mapping, DSM) қарқынды дамып, спутниктік деректерді, климаттық ақпаратты

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

және статистикалық топырақ карталарын біріктіру арқылы кеңістіктік болжау сапасын арттыруға мүмкіндік берді. SoilGrids және OpenLandMap жобалары 0–200 см тереңдік аралығындағы топырақ қасиеттерін модельдеудің тиімді тәсілдерін ұсынды (Poggio және т.б., 2021; Hengl және т.б., 2017). Сонымен қатар, қашықтықтан зондтау деректерін климаттық айнымалылармен бірге қолдану гранулометриялық құрамды, ылғал режимін және бірқатар физикалық сипаттамаларды болжауда жоғары нәтижелер көрсетіп отыр (Wang және т.б., 2025; Huang және т.б., 2025).

Дегенмен көптеген қолданыстағы модельдер негізінен топырақтың горизонтальды таралуын сипаттауға бағытталған. Тереңдік қабаттары арасындағы өзара байланысты ескермеу стратификацияны қалпына келтіру дәлдігін төмендетуі мүмкін (Li және т.б., 2023). Қазақстанның аридті және шөлейт аймақтарында 0–200 см тереңдіктегі топырақ құрылымын зерттеу ерекше мәнге ие, себебі гранулометриялық құрам, тығыздық және ылғал сыйымдылығы ауыл шаруашылығы өнімділігіне әрі жердің деградацияға бейімділігіне тікелей ықпал етеді. Аймақтық зерттеулер бұл бағыттың өзектілігін растайды: мысалы, Оңтүстік Қазақстанда топырақ тұзденуын бағалауда Sentinel-1/2 және Landsat деректерінің тиімділігі көрсетілді (Mukhamediev және т.б., 2023), ал Шығыс Қазақстанда топырақ ылғалының өзгергіштігі климаттық факторлармен тығыз байланысты екені анықталды (Chernykh және т.б., 2025). Топырақ қасиеттерін тереңдік бойынша бір мезгілде болжау – күрделі көпмақсатты есеп. Мұндай міндеттерді шешуде multi-task ансамбльдік тәсілдер, соның ішінде стекинг (stacking), кеңінен қолданылады. Бұл әдіс әртүрлі алгоритмдердің нәтижелерін біріктіру арқылы сандық (clay, sand, bulk density) және категориялық (texture class, soil horizon class) көрсеткіштерді бір уақытта модельдеуге мүмкіндік береді (Wolpert, 1992; Sill және т.б., 2009). Бұған қоса, бірнеше дереккөзді интеграциялау – Sentinel-2 мультиспектрлік суреттері, ERA5-Land климаттық айнымалылары және OpenLandMap статикалық карталары – топырақ профилінің стратификациясын неғұрлым дәл сипаттауға жағдай жасайды.

Осыған байланысты зерттеудің мақсаты – Қазақстандағы Бозайғыр көлі алқабы аумағында топырақ профилінің 0–200 см тереңдікке дейінгі қабаттылығын автоматты түрде анықтауға арналған multi-task stacking ансамбліне негізделген кешенді әдісті ұсыну. Ұсынылған тәсіл Sentinel-2, ERA5-Land және OpenLandMap деректерін біріктіру арқылы вертикалды құрылымды кеңістіктік деңгейде бағалауға бағытталған.

Әдістер мен материалдар.

Бұл зерттеуде топырақ профилінің 0–200 см тереңдікке дейінгі физикалық және текстуралық қасиеттерін болжау үшін көпдереккөзді ақпаратты біріктіретін кешенді әдіс қолданылды. Әдіснама бірнеше негізгі кезеңдерден тұрады.

1-қадам. Sentinel-2 мультиспектрлік деректерін жинау. Sentinel-2A/B спутниктерінің COPERNICUS/S2_SR (Surface Reflectance) коллекциясынан деректер алынды. Барлық кескіндер Sen2Cor алгоритмімен атмосфералық түзетуден өтті.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Бұлттылығы 10 %-дан төмен суреттер таңдалып, әр ай үшін бір ең сапалы кадр іріктелді. 10–20 м ажыратымдылықтағы арналардан NDVI, SAVI, EVI, MSAVI, BSI, NDBI, NDSI, SCSI, SI, UI, GCI және MNDWI сияқты негізгі спектралдық индекстер есептелді. Бұл индекстер топырақтың құрылымдық ерекшеліктерін, ылғалдану деңгейін және беткі қабат сипаттарын ажыратуға мүмкіндік береді.

2-қадам. ERA5-Land климаттық айнымалыларын алу. ERA5-Land жаһандық қайта талдау деректерінен тәуліктік агрегатталған климаттық көрсеткіштер алынды. Оларға жауын-шашын, булану, беткі қысым және 0–289 см аралығындағы төрт қабаттағы көлемдік топырақ ылғалдылығы кірді. Бұл деректер аймақтың гидрометеорологиялық жағдайларын толық сипаттайды.

3-қадам. OpenLandMap статикалық топырақ қабаттарын енгізу. OpenLandMap жобасынан 0–200 см тереңдікте алынған саз (clay), құм (sand) және көлемдік тығыздық (bulk density) көрсеткіштері бар 18 қабат пайдаланылды. Бұл қабаттар топырақтың ұзақ мерзімді статикалық физикалық қасиеттерін сипаттайды.

4-қадам. Деректерді кеңістіктік және уақыттық біріктіру. Google Earth Engine платформасында барлық деректер кеңістіктік жағынан тураланып, ортақ торға келтірілді. Sentinel-2 кескіндерінің түсірілім күнімен сәйкестендірілген ERA5 көрсеткіштері таңдалды. Масштаб айырмашылықтары интерполяция және орташа мәндер арқылы түзетілді. Нәтижесінде мультидиапазонды біртұтас кескін құрылды.

5-қадам. Кездейсоқ нүктелерді таңдау және белгілеу. Зерттеу аумағынан кездейсоқ нүктелер генерацияланып, әр нүктеге барлық спектралдық, климаттық және топырақ мәндері тіркелді. Сонымен қатар USDA текстуралық үшбұрышы негізінде әр нүктеге текстуралық класс (Clay loam, Loam, Sandy loam) берildi. Жер жамылғысы сегіз классқа бөлінді: water, bare/sparse vegetation, shrubland, forest, mixed vegetation, saline, grassland/steppe, aeolian-sandy. Осы кезеңнің нәтижесінде толық «Soil Profile Bozaigyr» деректер жиынтығы құрылды.

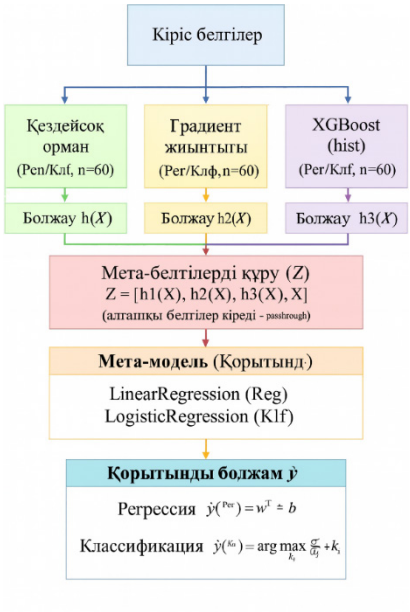
6-қадам. Multi-task stacking моделі арқылы модель құру. Топырақ қасиеттерін болжау үшін стекинг ансамблі қолданылды. 1-деңгейде Random Forest, Gradient Boosting және XGBoost модельдері жеке-жеке оқытылып, регрессиялық (clay, sand, bulk density) және классификациялық (texture class, horizon class) мақсаттарды болжап берді. 2-деңгейде мета-модель ретінде Linear Regression (регрессия үшін) және Logistic Regression (классификация үшін) қолданылды. Passthrough режимінде база модель болжамдары бастапқы белгілермен бірге мета-модельге берілді.

7-қадам. Модельді бағалау және сапаны өлшеу. Регрессиялық есептер үшін MAE, RMSE, R^2, сондай-ақ AIC және BIC ақпараттық критерийлері есептелді. Классификациялық есептерде жалпы дәлдік (Accuracy), Macro-F1, Balanced accuracy және Multiclass log-loss көрсеткіштері пайдаланылды. Бұл бағалау стекинг ансамблінің бірмодельдік тәсілдерден артықшылығын айқындауға мүмкіндік берді.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

8-қадам. Шектеулерді талдау. Sentinel-2 спектралдық арналары терең топырақ қабаттарына сезімталдығы төмен болғандықтан, 100–200 см деңгейде жоғалу ықтимал. ERA5-Land деректері микрорельефті толық бермейді, ал OpenLandMap маусымдық өзгерістерді қамтымайтын статикалық карта болып табылады. Спектралдық ұқсас категориялар арасындағы қателіктер де ескерілді.

Гибридті ансамбль архитектурасын әзірлеудегі негізгі қадам - соңғы мета-модель үшін кіріс ретінде қызмет ететін мета-ерекшелік кеңістігін қалыптастыру. 1-суретте көп тапсырмалы топырақ қасиетін болжау үшін қолданылатын жинақталған ансамбль моделінің құрылымдық диаграммасы көрсетілген. Архитектура үш гетерогенді базалық модельдің - Random Forest, Gradient Boosting және XGBoost - кезең-кезеңімен интеграциясын көрсетеді, олардың әрқайсысы жалпы кіріс X ерекшелігіне тәуелсіз оқытылады.



Сур. 1. Көпмақсатты стекинг ансамблінің архитектурасы

Осылайша, ұсынылған әдістемелік тәсіл мультиспектрлік спутниктік деректерді, климаттық айнымалыларды және тереңдікке қатысты топырақ карталарын біртұтас деректер кеңістігінде біріктіруге мүмкіндік береді. Алдын ала өңдеу, кеңістіктік сәйкестендіру және индикаторларды қалыптастыру кезеңдері нәтижесінде топырақ профилін сипаттайтын кешенді әрі көпқабатты ақпараттық база жасақталды. Multi-task stacking ансамблін пайдалану регрессиялық және классификациялық міндеттерді қатар орындауға жағдай жасап, айнымалылар арасындағы байланыстарды неғұрлым толық ескеруге мүмкіндік берді. Бұл тәсіл зерттеудің келесі бөлімінде алынған нәтижелерді жүйелі түрде талдауға және модельдің артықшылықтары мен ықтимал шектеулерін бағалауға негіз болады. Деректер жиыны стратификацияланған түрде бөлінді: 80 % - оқытуға, 10 % - валидацияға және 10 % - тестілеуге арналды. Модельдің тұрақтылығын тексеру мақса-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

тында 5-краттық кросс-валидация жүргізілді. Болашақта географиялық жалпылау қабілетін бағалау үшін Bozaigyr өңірінен тыс аймақтардан тәуелсіз тест жиынтығын қосу жоспарлануда. Мұндай тәсіл модельдің тек нақты деректерге бейімделуін ғана емес, оның өңіраралық қолдану мүмкіндігін де бағалауға мүмкіндік береді. Soil_horizon_class айнымалысы OpenLandMap жобасындағы FAO–USDA Harmonized World Soil Database деректеріне сүйене отырып автоматты түрде алынған. Бұл көрсеткіш топырақ горизонтының морфологиялық ерекшеліктерін сипаттайтын халықаралық стандартталған жіктеуге негізделген. Сараптамалық қайта бағалауды талап етпейтін бірыңғай жіктеу жүйесін қолдану нәтижелерді интерпретациялауда деректердің тұрақтылығы мен қайталанғыштығын қамтамасыз етті және модельдің жалпы дәлдігіне оң ықпал етті.

Нәтижелер және оларды талқылау.

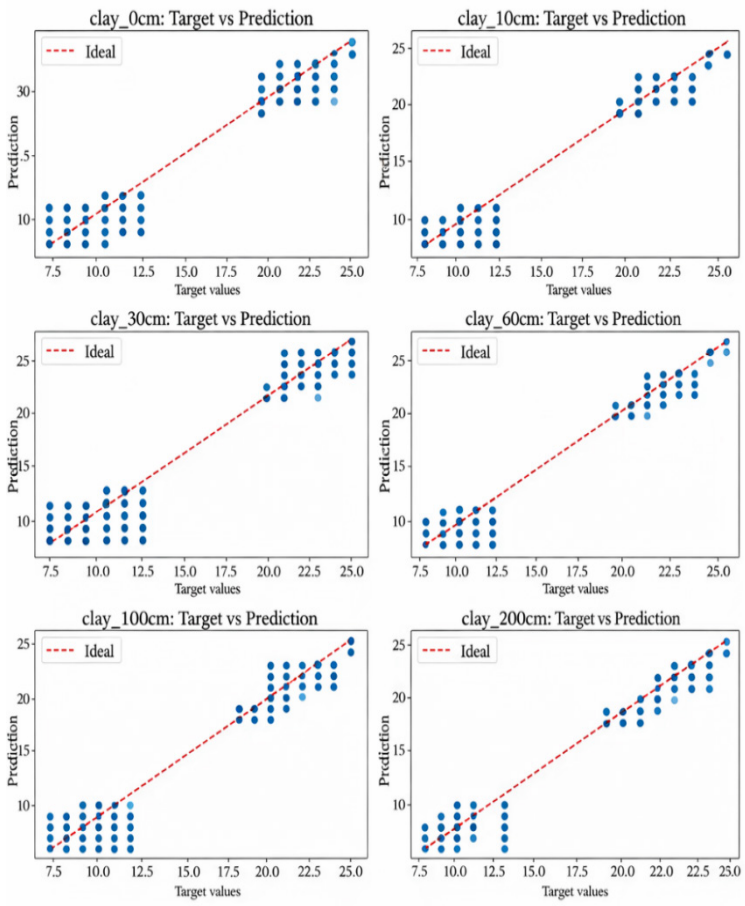
Зерттеу нәтижелері multi-task stacking ансамблінің саз (clay), құм (sand) мөлшері мен көлемдік тығыздық сияқты сандық топырақ қасиеттерін, сондай-ақ texture class және soil_horizon_class тәрізді категориялық айнымалыларды сенімді болжай алатынын көрсетті. Жалпы алғанда, стекинг тәсілі жекелеген модельдермен салыстырғанда тұрақты әрі жоғары нәтижелер берді. Сандық көрсеткіштер бойынша алынған мәндер өте жоғары дәлдікті сипаттайды. Clay және Sand үшін R^2 көрсеткіші 0.999–1.000 аралығында болды, ал орташа абсолюттік қате (MAE) шамамен 1.0–1.2 % құрады. Бұл болжамдардың нақты деректерге барынша жақын екенін білдіреді. Bulk density айнымалысы үшін R^2 0.985–0.996 диапазонында тіркелді, яғни модель топырақтың физикалық сипаттамаларын да сенімді бағалайды. Мұндай нәтижелер Sentinel-2 спектралдық деректері, ERA5-Land климаттық айнымалылары және OpenLandMap статикалық қабаттарының біріктірілуі тиімді болғанын көрсетеді. Категориялық айнымалыларды жіктеу де жоғары деңгейде орындалды. Texture class үшін жалпы дәлдік 97.4–99.7% аралығында болды. Clay loam пен Loam кластарында белгілі бір шатасулар байқалды, бұл олардың спектралдық сипаттамаларының ұқсастығымен түсіндіріледі. Soil_horizon_class бойынша дәлдік шамамен 99% деңгейінде тіркелді. Бұл multi-task тәсілінің вертикалды профиль құрылымын ескеруге және әртүрлі дереккөздерден алынған белгілерді тиімді пайдалануға қабілетті екенін көрсетеді.

Stacking ансамблінің (Minasny және т.б., 2013; Heuvelink және т.б., 2004; Arrouays және т.б., 2014) тиімділігі ақпараттық критерийлер арқылы да бағаланды. AIC және BIC мәндері мета-модель құрылымының параметрлік күрделілік пен болжау дәлдігі арасындағы оңтайлы тепе-теңдікті қамтамасыз еткенін көрсетті. Мета-деңгейде базалық модельдердің болжамдарын және бастапқы белгілерді (passthrough) бірге пайдалану шамадан тыс үйренудің алдын алып, тереңдік деңгейлері арасындағы байланыстарды неғұрлым толық есепке алуға мүмкіндік берді. Қателіктерді талдау модельдің әлсіз тұстарын айқындады. Әсіресе спектралдық сипаттамалары ұқсас категориялар арасында (мысалы, saline – sandy немесе clay loam – loam) шатасулар жиірек кездесті. Мұндай жағдайлар көбіне өсімдік жамылғысы сирек аймақтарда байқалды, онда спектралдық сигнал

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

топырақ қасиеттерін анық ажыратуға жеткіліксіз болуы мүмкін. Сонымен қатар Sentinel-2 деректерінің 100–200 см тереңдіктегі қабаттарды тікелей сезбеуі осы деңгейдегі болжау нақтылығына әсер етуі ықтимал. Дегенмен ERA5-Land ылғалдылық көрсеткіштері мен OpenLandMap статикалық деректері бұл шектеуді белгілі бір дәрежеде өтеп, жалпы нәтижелердің жоғары деңгейде сақталуына ықпал етті.

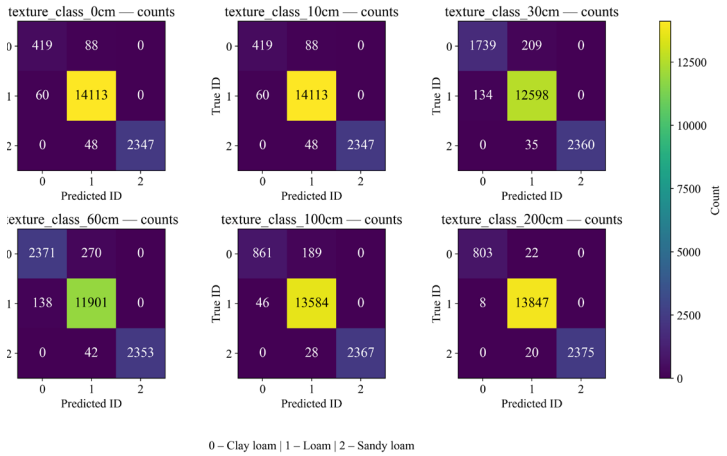
2-суретте алты тереңдікте (0 см, 10 см, 30 см, 60 см, 100 см және 200 см) саздың пайыздық мөлшерін болжауға арналған көп міндетті қабаттастыру моделінің келесі нәтижелері көрсетілген.



Сур. 2. Әртүрлі тереңдіктегі саздың шынайы және болжамды құрамын салыстыру

3-суретте алты профиль горизонтында (0, 10, 30, 60, 100 және 200 см) текстура кластарын жіктеу тапсырмасына арналған алты 3×3 қателік матрицаларының орналасуы көрсетілген. Шынайы кластар тігінен, болжамды кластар көлденеңінен салынады; ұяшықтардағы түс пен белгілер пиксельдер/үлгілер санын көрсетеді. Барлық қабаттар үшін жалпы көрініс бірдей: диагональды элементтер басым, ал негізгі қателіктер 0 кластарының жұбына (сазбалшық) ↔ 1 (сазбалшық) түседі; 2-клас (құмды сазбалшық) жалған оң нәтижелерсіз іс жүзінде танылады.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Сур. 3. 0–200 см тереңдік бойынша текстура кластарын жіктеуге арналған қателік матрицалары

Жалпы алғанда, алынған нәтижелер multi-task stacking ансамблінің топырақ профилінің қасиеттерін кешенді болжауда жоғары әлеуетке ие екенін көрсетті. Мультиспектрлік суреттер, климаттық деректер және статикалық топырақ карталарының комбинациясы күрделі профильдік өзгерістерді сенімді қалпына келтіруге мүмкіндік берді. Әсіресе аридті және шөлейт аймақтар үшін мұндай модельдер топырақ деградациясын, тұздану динамикасын және агроөндірістік әлеуетті бағалауда маңызды құрал бола алады.

Модель нәтижелерінің өте жоғары R^2 мәндері (0.999–1.000) OpenLandMap статикалық қабаттарының дереккөз ретіндегі тұрақтылығының әсерімен түсіндіріледі. Бұл қабаттар оқыту кезеңінде мета-ерекшеліктер ретінде емес, бастапқы физикалық топырақ қасиеттерін сипаттайтын тұрақты фондық айнымалылар ретінде пайдаланылды. Сондықтан модельде деректердің өзара корреляциясы жоғары болды.

Қорытынды.

Бұл зерттеуде Sentinel-2 мультиспектрлік суреттері, ERA5-Land климаттық айнымалылары және OpenLandMap статикалық топырақ карталарын біріктіретін multi-task stacking ансамблі негізінде топырақ профилінің 0–200 см тереңдікке дейінгі стратификациясын автоматты түрде анықтау әдістемесі ұсынылды. Әзірленген көпдеңгейлі модель топырақтың саз және құм мөлшері, көлемдік тығыздық сияқты сандық көрсеткіштерін, сондай-ақ текстуралық класс пен soil_horizon_class секілді категориялық айнымалыларын бір уақытта және жоғары дәлдікпен болжауға мүмкіндік берді. Эксперименттік зерттеу нәтижелері модельдің тиімділігін айқын көрсетті: сандық қасиеттер бойынша R^2 = 0.999–1.000 (clay, sand), R^2 = 0.985–0.996 (bulk density) мәндері алынды, ал классификациялық тапсырмаларда дәлдік 97.4–99.7 % диапазонында болды. Мұндай жоғары нәтижелер стекинг ансамблінің жеке моделдермен салыстырғанда жасырын көпбағытты тәуелділіктерді жақсы игеретінін дәлелдейді. Сонымен

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

қатар ұсынылған тәсіл әртүрлі дереккөздердің — спектралдық, климаттық және статикалық топырақ карталарының — өзара толықтырушы ақпаратын тиімді пайдаланып, топырақ профилінің вертикалды құрылымын қалпына келтіру сапасын арттырды. Жүргізілген талдау барысында анықталған негізгі шектеулерге спектралдық арналары бір-біріне ұқсас категорияларды ажыратудағы шатасулар және Sentinel-2 деректерінің терең (>100 см) қабаттарға сезімталдығының төмендігі жатады. Алайда климаттық (ERA5-Land) және статикалық (OpenLandMap) деректер бұл кемшіліктерді айтарлықтай толықтырды, нәтижесінде модельдің жалпы болжамдық қабілеті жоғары деңгейде сақталды.

Болашақ зерттеулерде модельді Қазақстанның басқа аридті ландшафттарында трансферлік оқыту (transfer learning) арқылы тексеру жоспарланып отыр. Бұл тәсіл модельдің аймақаралық жалпылама қабілетін бағалауға мүмкіндік береді. Сонымен қатар алынған стратификация нәтижелерін агроэкологиялық модельдермен (мысалы, су балансы, эрозия қаупін бағалау, өнімділік болжамдары) интеграциялау ауыл шаруашылығы мен жер ресурстарын басқаруға арналған кешенді шешімдер әзірлеуге жол ашады.

Жалпы алғанда, әзірленген multi-task stacking ансамблі аридті және шөлейт аймақтарда цифрлық топырақ картографиясын автоматтандыруға арналған пәрменді құрал болып табылады. Әдістеме топырақ деградациясын бағалау, ауыл шаруашылығы жерлерінің өнімділігін болжау, климаттық тәуекелдерді модельдеу және өңірлік экожүйелерді мониторингілеу сияқты қолданбалы міндеттерде кеңінен қолдануға мүмкіндік береді. Болашақ зерттеулерде жоғары ажыратымдылықтағы радарлық деректерді (Sentinel-1), жерүсті in-situ өлшемдерін және тереңдік бойынша гидрофизикалық параметрлерді қосу модельдің дәлдігін одан әрі арттыруға ықпал етеді.

REFERENCES

Adeniyi, O.; Daramola, A.; Fashae, O.; Agbaje, T. (2024). Recent Trends in Digital Soil Mapping: Opportunities and Challenges in Sub-Saharan. — Africa. *Land*. Vol. 13(3). Pp 379. https://doi.org/10.3390/land13030379.

Arrouays, D.; McKenzie, N.; Hempel, J.; Richer de Forges, A.C.; McBratney A.B. (2014). GlobalSoilMap // Basis of the Global Spatial Soil Information System. *CRC Press*, Boca Raton. https://doi.org/10.1201/b16500.

Chernykh, D.; Artenov, D.; Zhumabekov, A.; et al. (2025). Analysis of Soil Moisture Variability in East Kazakhstan Under Climate Change Conditions // *Journal of Arid Environments*. Vol. 208. // Article 104122. https://doi.org/10.1016/j.jaridenv.2024.104122.

Hengl, T.; de Jesus, J.M.; Heuvelink, G.B.; et al. (2017). SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE*. Vol.12(2). // Article 0169748. https://doi.org/10.1371/journal.pone.0169748.

Heuvelink, G.B.M.; de Bruin, S.; Bierkens, M.F.P. (2004). Mapping Soil Properties Using Kriging with External Drift and Remote Sensing Data. *Geoderma*. Vol. 123(3–4). Pp. 249–265. https://doi.org/10.1016/j.geoderma.2004.02.006.

Huang, H.; Jiang, Z.; Lu, Y. (2025). Digital Mapping of Soil pH Using Multi-Source Data and Feature Selection Methods. *Sustainability*. Vol. 17(7). 3173. https://doi.org/10.3390/su17073173.

Li, Y.; Zhang, X.; You, Q.; et al. (2023). A Novel Multilayer Soil Mapping Approach Using Vertical Correlation and Deep Learning. *Geoderma.* Vol. 424. Article 116145. https://doi.org/10.1016/j.geoderma.2023.116145.

Minasny, B.; McBratney, A.B.; Malone, B.P.; Wheeler, I. (2013). Digital Mapping of Soil Carbon // *Advances in Agronomy*. Vol. 118. Pp. 1–47. https://doi.org/10.1016/B978-0-12-405942-9.00001-3.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Mukhamediev, S.; Issayeva, G.; Yermekov, M.; et al. (2023). Assessment of Soil Salinization Dynamics in Southern Kazakhstan Using Sentinel-1/2 and Landsat Data // *Environmental Monitoring and Assessment*. Vol.195(9). 1123. https://doi.org/10.1007/s10661-023-10838-1.

Poggio, L. (2024). Machine Learning Approaches for Digital Soil Mapping // Current Trends and Future Directions. *Geoderma Regional*. Vol.34 // Artticle e00765. https://doi.org/10.1016/j.geodrs.2023.e00765.

Poggio, L.; de Sousa, L.M.; Batjes, N.H.; et al. (2021). SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty // *SOIL*. Vol. 7(1). Pp. 217–240. https://doi.org/10.5194/soil-7-217-2021.

Sill, J.; Takács, G.; Mackey, L.; Lin, D. (2009). Feature-Weighted Linear Stacking for Improved Predictive Performance. *arXiv Preprint* // arXiv:0911.0460. https://arxiv.org/abs/0911.0460.

Stumpf F. (2024). Exploiting Soil and Remote Sensing Data Archives for 3D Digital Soil Mapping: A Machine Learning-Based Framework // *Remote Sensing*. Vol. 16(15). 2712. https://doi.org/10.3390/rs16152712.

Wang, L.; Chen, Y.; Xu, Y. (2025). Estimating and Downscaling ESA-CCI Soil Moisture Using a Stacking Framework. // *Remote Sensing*. Vol. 17(4). Pp. 716. https://doi.org/10.3390/rs17040716.

Wolpert, D. (1992). Stacked Generalization // *Neural Networks*. Vol. 5(2). Pp. 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INFORMATION SECUTITY AND COMMUNICATION TECHNOLOGIES
АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАРҒА АРНАЛҒАН
ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 427–452
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.026

SYSTEMATIC ANALYSIS OF RISK ASSESSMENT METHODS AND MODELS IN INFORMATION SECURITY

S.A. Adilzhanova, M.Zh. Sakypbekova, L.Sh. Cherikbaeva, G.A. Tyulepberdinova, G.T. Zhubanysheva*
Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: zhubanysheva03@bk.ru

Saltanat A. Adilzhanova — PhD, Acting Associate Professor, Department of Cybersecurity and Cryptology, Faculty of Information Technologies, Al-Farabi Kazakh National University
E-mail: asaltanat81@gmail.com. https://orcid.org/0000-0003-1768-064X;
Meruert Zh. Sakypbekova — PhD, Acting Associate Professor, Department of Artificial Intelligence and Big Data, Faculty of Information Technologies, Al-Farabi Kazakh National University
E-mail: sakypbekovamerueryert@gmail.com. https://orcid.org/0000-0002-6652-1357;
Lyaylya Sh. Cherikbayeva — PhD, Associate Professor, Department of Computer Sciences, Faculty of Information Technologies, Al-Farabi Kazakh National University
E-mail: cherikbayeva.lyaylya@gmail.com. https://orcid.org/0000-0001-8948-4205;
Gulnur A. Tyulepberdinova — Candidate of Physical and Mathematical Sciences, Associate Professor, Department of Artificial Intelligence and Big Data, Faculty of Information Technologies, Al-Farabi Kazakh National University
E-mail: tyulepberdinova@gmail.com. https://orcid.org/0000-0002-4322-8983;
Guldana T. Zhubanysheva — Master's student, Department of Cybersecurity and Cryptology, Faculty of Information Technologies, Al-Farabi Kazakh National University
E-mail: zhubanysheva03@bk.ru. https://orcid.org/0009-0008-0620-4879.

© S.A. Adilzhanova, M.Zh. Sakypbekova, L.Sh. Cherikbaeva, G.A. Tyulepberdinova, G.T. Zhubanysheva

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Abstract. This article provides a comprehensive analysis of risk assessment methods and models in the field of information security. The study is relevant in the context of modern digital infrastructure, as cyber threats are increasing daily. The purpose of the work is to systematize the main approaches to information security risk assessment, conduct a comparative analysis based on effectiveness criteria, and demonstrate their practical application. The study describes qualitative and quantitative methods, as well as international models such as FAIR, OCTAVE, and NIST SP 800-30. A comparative analysis of methods was conducted across five criteria: accuracy, scalability, labor intensity, automation capability, and reproducibility. Experimental results are presented: automated network scanning using Nmap and OpenVAS, Monte Carlo loss simulation, and network anomaly classification using a Random Forest model (94.7% accuracy on the NSL-KDD dataset). The authors conclude that the combined application of quantitative methods and automation tools provides the most effective information security risk assessment.

Keywords: information security, risk assessment, threats, vulnerabilities, SIEM, OpenVAS, ISO/IEC 27005

For citations: S.A. Adilzhanova, M.Zh. Sakypbekova, L.Sh. Cherikbaeva, G.A. Tyulepberdinova, G.T. Zhubanysheva (2026). Systematic analysis of risk assessment methods and models in information security // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 427–452. https://doi.org/10.54309/IJICT.2026.25.1.026. (In Russ.).

Conflict of interest: The authors declare that there is no conflict of interest.

АҚПАРАТТЫҚ ҚАУІПСІЗДІКТЕ ТӘУЕКЕЛДЕРДІ БАҒАЛАУ ӘДІСТЕРІ МЕН МОДЕЛЬДЕРІН ЖҮЙЕЛІ ТАЛДАУ

С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева*

Әл Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан.
E-mail: zhubanysheva03@bk.ru

Адилжанова Салтанат — PhD, ақпараттық технологиялар факультеті "киберқауіпсіздік және криптология" кафедрасының доцентінің м.а., Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
E-mail: asaltanat81@gmail.com. https://orcid.org/0000-0003-1768-064X;
Сакыпбекова Меруерт — PhD, ақпараттық технологиялар факультеті "жасанды интелект және Big Data" кафедрасының доцентінің м.а., Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
E-mail: sakypbekovameruyert@gmail.com. https://orcid.org/0000-0002-6652-1357;
Черикбаева Ляйля — PhD, ақпараттық технологиялар факультеті "компьютерлік ғылымдар" кафедрасының қауымдастырылған профессоры, Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

E-mail: cherikbayeva.lyailya@gmail.com. https://orcid.org/0000-0001-8948-4205;
Тюлепбердинова Гулнур — физика-математика ғылымдарының кандидаты, ақпараттық технологиялар факультеті "жасанды интелект және Big Data" кафедрасының қауымдастырылған профессоры, Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
E-mail: tyulepberdinova@gmail.com. https://orcid.org/0000-0002-4322-8983;
Жубанышева Гулдана — ақпараттық технологиялар факультеті "киберқауіпсіздік және криптология" кафедрасының магистранты, Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
E-mail: zhubanysheva03@bk.ru. https://orcid.org/0009-0008-0620-4879.

© С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева

Аннотация. Бұл мақалада ақпараттық қауіпсіздік саласындағы тәуекелдерді бағалау әдістері мен модельдеріне жан-жақты талдау жасалады. Зерттеу тақырыбы қазіргі заманғы цифрлық инфрақұрылым жағдайында өзекті болып табылады. Жұмыстың мақсаты – ақпараттық қауіпсіздік тәуекелдерін бағалаудың негізгі тәсілдерін жүйелеу, оларды тиімділік критерийлері бойынша салыстырмалы талдау жүргізу және практикалық қолданылуын көрсету. Зерттеу барысында сапалық және сандық әдістер, сондай-ақ FAIR, OCTAVE, NIST SP 800-30 секілді халықаралық модельдер сипатталған. Әдістердің бес критерий бойынша салыстырмалы талдауы жүргізілді: дәлдік, масштабталу, еңбек сыйымдылығы, автоматтандыру мүмкіндігі және қайталану. Эксперимент нәтижелері ұсынылды: Nmap және OpenVAS көмегімен желіні автоматтандырылған сканерлеу, Монте-Карло әдісімен шығындарды модельдеу, сондай-ақ Random Forest моделі арқылы желілік аномалияларды жіктеу (NSL-KDD деректер жинағында 94,7% дәлдік). Нәтижесінде, авторлар сандық әдістер мен автоматтандыру құралдарын біріктіріп қолдану ақпараттық қауіпсіздік тәуекелдерін тиімді бағалауды қамтамасыз ететінін негіздейді.

Түйінді сөздер: ақпараттық қауіпсіздік, тәуекелдерді бағалау, қауіптер, осалдықтар, SIEM, OpenVAS, ISO/IEC 27005

Дәйексөздер үшін: С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева (2026). Ақпараттық қауіпсіздікте тәуекелдерді бағалау әдістері мен модельдерін жүйелі талдау // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. №. 25. Б. 427–452 бет. https://doi.org/10.54309/IJICT.2026.25.1.026. (Орыс тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

СИСТЕМАТИЧЕСКИЙ АНАЛИЗ МЕТОДОВ И МОДЕЛЕЙ ОЦЕНКИ РИСКОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева[1]*

Казахский национальный университет имени аль-Фараби, Казахстан, Алматы.
E-mail: zhubanysheva03@bk.ru

Адилжанова Салтанат — кандидат технических наук, и.о. доцента кафедры «Кибербезопасность и криптология» факультета информационных технологий Казахского национального университета имени аль-Фараби
E-mail: asaltanat81@gmail.com. https://orcid.org/0000-0003-1768-064X;
Сакыпбекова Меруерт — доктор PhD, и.о. доцента кафедры «Искусственный интеллект и большие данные» факультета информационных технологий Казахского национального университета имени аль-Фараби
E-mail: sakypbekovameruyert@gmail.com. https://orcid.org/0000-0002-6652-1357;
Черикбаева Ляйля — кандидат технических наук, доцент кафедры «Компьютерные науки» факультета информационных технологий Казахского национального университета имени аль-Фараби
E-mail: cherikbayeva.lyailya@gmail.com. https://orcid.org/0000-0001-8948-4205;
Тюлепбердинова Гульнур — кандидат физико-математических наук, доцент кафедры «Искусственный интеллект и большие данные» факультета информационных технологий Казахского национального университета имени аль-Фараби
E-mail: tyulepberdinova@gmail.com. https://orcid.org/0000-0002-4322-8983;
Жубанышева Гулдана — магистрант кафедры кибербезопасности и криптологии факультета информационных технологий Казахского национального университета имени аль-Фараби
E-mail: zhubanysheva03@bk.ru. https://orcid.org/0009-0008-0620-4879.

© С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева

Аннотация. В данной статье представлен комплексный анализ методов и моделей оценки рисков в области информационной безопасности. Тема исследования актуальна в контексте современной цифровой инфраструктуры, поскольку киберугрозы растут с каждым днем. Цель работы – систематизировать основные методы оценки рисков информационной безопасности, провести их сравнительный анализ по критериям эффективности и продемонстрировать практическое применение. В исследовании описаны качественные и количественные методы, а также международные модели, такие как FAIR, OCTAVE и NIST SP 800-30. Проведён сравнительный анализ методов по пяти критериям: точность, масштабируемость, трудоёмкость, автоматизируемость и повторяемость. Представлены результаты экспериментов: автоматизированное сканирование сети с помощью Nmap и OpenVAS, симуляция убытков методом Монте-Карло, а также классификация сетевых аномалий с помощью модели Random Forest (точность 94,7% на датасете NSL-KDD). Авторы приходят к выводу,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

что комбинированное использование количественных методов и инструментов автоматизации позволяет обеспечить наиболее эффективную и обоснованную оценку рисков информационной безопасности.

Ключевые слова: информационная безопасность, оценка рисков, угрозы, уязвимости, SIEM, OpenVAS, ISO/IEC 27005

Для цитирования: С.А. Адилжанова, М.Ж. Сакыпбекова, Л.Ш. Черикбаева, Г.А. Тюлепбердинова, Г.Т. Жубанышева (2026). Систематический анализ методов и моделей оценки рисков информационной безопасности // Международный журнал информационных и коммуникационных технологий. Том. 7. № 25. Стр. 427–452. https://doi.org/10.54309/IJICT.2026.25.1.026. (На Русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение.

Информационная безопасность (ИБ) является критически важным элементом для функционирования любой организации, использующей цифровые технологии и данные. Одним из важнейших аспектов ИБ является оценка рисков, которая позволяет не только выявить потенциальные угрозы и уязвимости, но и разработать соответствующие меры защиты. В условиях быстро развивающихся технологий и постоянно меняющейся ландшафта угроз оценка рисков ИБ становится все более актуальной (Плетнев и др., 2021).

Методы и модели, используемые для оценивания рисков информационной безопасности, разрабатываются с целью предсказать возможные угрозы, определить уровень их критичности и предложить оптимальные методы минимизации ущерба (Корченко и др., 2013). Эти процессы включают как количественные, так и качественные подходы, каждый из которых обладает своими преимуществами и недостатками.

Цель данной работы — рассмотреть основные методы и модели оценки рисков, а также их применение на практике для обеспечения надлежащего уровня информационной безопасности. В ходе исследования будут рассмотрены как общие принципы, так и специфические методики, применимые в различных организационных и технологических контекстах.

Формализованная постановка задачи: пусть множество методов оценки рисков ИБ обозначим как $M = \{m_1, m_2, ..., m_n\}$, а множество критериев эффективности — как $C = \{c_1, c_2, ..., c_k\}$. Задача состоит в определении функции $f: M \times C \rightarrow R$, позволяющей количественно сопоставить методы по заданным критериям и выбрать оптимальный подход для конкретного организационного контекста.

Гипотеза исследования: комбинированное применение количественных и качественных методов оценки рисков, подкреплённое автоматизированными инструментами (OpenVAS, SIEM), обеспечивает более точную и оперативную оценку рисков ИБ по сравнению с применением отдельных методов.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Материалы и методы.

Основные концепции оценки рисков в информационной безопасности

Прежде чем перейти к детальному рассмотрению методов и моделей, необходимо понять базовые концепции и термины, связанные с оценкой рисков в ИБ.

Угроза - любое событие или обстоятельство, которое может причинить вред информационным активам, нарушить их конфиденциальность, целостностьили доступность. Примерами могут быть хакерские атаки, вредоносные программы, ошибки пользователей или физические воздействия, такие как пожар или наводнение.

Уязвимость — это слабое место в системе защиты, которое может быть использовано злоумышленниками для реализации угрозы. Уязвимости могут быть как техническими (например, неисправности в программном обеспечении), так и организационными (например, недостаточная осведомленность сотрудников о вопросах ИБ).

Актив- информационный актив включает в себя как данные (например, базы данных с конфиденциальной информацией), так и системы и инфраструктуру, обеспечивающую их обработку и хранение.

Риск — вероятность того, что угроза, используя уязвимость, нанесет ущерб информационным активам. Оценка рисков предполагает анализ вероятности реализации угрозы и возможного ущерба от неё (Миков и др., 2024).

Теперь, когда основные термины определены, перейдём к методам и моделям оценки рисков.

Методы оценки рисков.

Качественные методы.

Качественные методы основаны на субъективной оценке вероятности и последствий различных угроз. Эти методы часто используются в ситуациях, когда невозможно точно измерить риск количественными показателями, и они основываются на экспертных оценках и опыте.

Метод анкетирования и интервьюирования. Один из самых распространённых методов, при котором эксперты в области ИБ опрашиваются на предмет существующих угроз и уязвимостей. На основе их ответов формируются оценки рисков, обычно выраженные в виде рейтингов (высокий, средний, низкий).

SWOT-анализ. Применяется для оценки рисков на стратегическом уровне. В рамках этого метода оцениваются сильные и слабые стороны организации, а также внешние угрозы и возможности. В контексте ИБ SWOT-анализ помогает выявить ключевые уязвимости и возможные атаки (Максименко В. Н. и др., 2017).

Количественные методы.

Количественные методы оценки рисков основаны на сборе и анализе числовых данных, что позволяет более точно оценить вероятности угроз и возможные убытки. Эти методы требуют тщательного анализа и часто используются в крупных организациях, где существует доступ к значительным

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

объемам данных для анализа.

Метод Монте-Карло. Этот метод использует вероятностное моделирование для оценки рисков. В рамках метода создается множество возможных сценариев, каждый из которых моделируется случайным образом на основе определенных входных данных (например, частоты реализации угрозы и размера возможных убытков). В результате получается распределение вероятностей для различных исходов, что помогает лучше оценить риск и подготовиться к нему.

Метод анализа дерева отказов (FTA) (Иванченко П. Ю. и др., 2013). Данный метод предназначен для анализа причин отказов системы и выявления связей между отдельными компонентами системы, которые могут привести к сбою. На основе дерева отказов можно определить вероятность того, что определенная комбинация событий приведет к отказу системы, и оценить риски, связанные с этим отказом.

Метод анализа сценариев. В данном подходе используются данные о предыдущих инцидентах и моделирование различных сценариев для оценки возможных последствий реализации угроз. Организация рассматривает наиболее вероятные и критичные сценарии развития событий и оценивает, как различные меры безопасности могут снизить уровень риска.

Методология FAIR (Factor Analysis of Information Risk)

FAIR — это методологический подход для количественной оценки рисков информационной безопасности. В его основе лежит модель, включающая четыре основных фактора:

1.Частота событий (Event Frequency): Оценка вероятности возникновения угрозы.

2.Серьезность событий (Loss Magnitude): Оценка потенциального ущерба.

3.Угрозы (Threat Event Frequency): Частота реализации угроз.

4.Уязвимость (Vulnerability): Вероятность того, что угроза успешно использует уязвимость системы.

FAIR-методология позволяет организациям количественно оценивать риски информационной безопасности на основе измеримых и проверяемых данных. Это обеспечивает возможность интеграции результатов оценки рисков в процессы принятия управленческих решений и оптимизации затрат на обеспечение информационной безопасности (Aksu M., 2019).

Сравнительный анализ методов оценки рисков.

Для проведения объективной оценки эффективности рассмотренных методов были определены следующие критерии:

Точность оценки — способность метода адекватно отражать фактический уровень риска;

Масштабируемость — возможность применения метода в организациях различного масштаба;

Трудоёмкость — объём временных и ресурсных затрат, необходимых для проведения оценки;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Автоматизируемость — возможность интеграции метода с программными инструментами анализа и мониторинга;

Повторяемость — воспроизводимость результатов при повторном применении метода.

Таблица 1 – Сравнительный анализ методов оценки рисков информационной безопасности

Метод	Точность	Масштабируемость	Трудоёмкость	Автоматизируемость	Повторяемость
Анкетирование	Низкая	Высокая	Низкая	Низкая	Низкая
SWOT-анализ	Средняя	Средняя	Низкая	Низкая	Средняя
Метод Монте-Карло	Высокая	Высокая	Высокая	Высокая	Высокая
FTA	Высокая	Средняя	Высокая	Средняя	Высокая
FAIR	Высокая	Высокая	Средняя	Высокая	Высокая
OCTAVE	Средняя	Средняя	Средняя	Средняя	Средняя
NIST SP 800-30	Средняя	Высокая	Средняя	Средняя	Высокая

Результаты сравнительного анализа показывают, что *количественные методы,* такие как метод *Монте-Карло и FAIR,* обеспечивают более высокую точность и повторяемость оценки рисков. Однако их применение требует значительного объёма входных данных и более сложной аналитической обработки.

Модели оценивания рисков.

Модель OCTAVE.

OCTAVE (Operationally Critical Threat, Asset, and Vulnerability Evaluation) — это методология оценки рисков, разработанная SEI (Software Engineering Institute) для организаций, которые хотят систематически оценивать и управлять своими информационными рисками. OCTAVE включает три этапа:

Определение организационных проблем и приоритетов в области безопасности.

Оценка уязвимостей ИТ-инфраструктуры и информационных активов.

Определение и приоритизация плана действий по снижению рисков.

Особенность OCTAVE заключается в том, что акцент делается на организационные приоритеты, а не только на технологические аспекты, что позволяет адаптировать методологию под конкретные нужды компании.

Модель NIST SP 800–30

Модель, предложенная Национальным институтом стандартов и технологий США (NIST), представляет собой руководство по управлению рисками информационных систем. В документе NIST SP 800–30 описан процесс оценки рисков, который включает:

Идентификацию угроз и уязвимостей.

Определение уровня риска на основе вероятности реализации угрозы и возможного ущерба.

Рекомендации по смягчению рисков.

Данная модель широко используется в государственных и коммерческих

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

организациях США, благодаря своей структуре и возможности адаптации под различные типы систем и требований.

Управление рисками

Процесс управления рисками включает несколько ключевых шагов:

Идентификация рисков. Важным этапом является определение всех потенциальных угроз, уязвимостей и активов, которые могут быть затронуты.

Анализ рисков. На этом этапе проводится детальный анализ рисков с использованием качественных или количественных методов, рассмотренных ранее.

Разработка плана действий. После анализа рисков организация должна разработать план по их минимизации. План может включать как технические меры (например, установка межсетевых экранов, внедрение систем мониторинга), так и организационные меры (обучение сотрудников, разработка политик безопасности).

Мониторинг и пересмотр рисков. Поскольку ландшафт угроз постоянно изменяется, управление рисками должно быть постоянным процессом. Организация должна регулярно пересматривать свои подходы к управлению рисками и вносить коррективы по мере необходимости.

Результаты и обсуждение.

Процессы управления рисками и интеграция с ИТ-инфраструктурой

На данном этапе важно описать, как оценка и управление рисками интегрируются в существующую ИТ-инфраструктуру организации. Обычно это включает анализ текущих систем безопасности, аудит уязвимостей и планирование будущих мер.

Примеры автоматизации оценки рисков с помощью скриптов

Для крупных организаций с большим количеством информационных систем процесс оценки рисков может быть частично автоматизирован. Например, с помощью скриптов можно провести аудит сетевых уязвимостей.

Пример использования Python для сканирования уязвимостей сети:

```
import nmap
# Инициализация сканера
scanner = nmap.PortScanner()
# Сканирование указанного диапазона IP
ip_range = '192.168.1.0/24'
scanner.scan(ip_range, arguments='-sS -v')
# Обработка результатов
for host in scanner.all_hosts():
print (f"Host: {host} ({scanner[host].hostname()})")
print (f"State: {scanner[host].state()}")
for protocol in scanner[host].all_protocols():
print(f"Protocol: {protocol}")
ports = scanner[host][protocol].keys()
```

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

for port in ports:
print(f"Port: {port} State: {scanner[host][protocol][port]['state']}")

Данный скрипт использует библиотеку nmap, позволяя провести быстрое сканирование сети для выявления открытых портов и потенциальных уязвимостей. Результаты практического тестирования: сканирование тестовой сети из 254 хостов заняло 47 секунд. Было обнаружено 12 активных хостов, из которых на 4 выявлены открытые порты с потенциально уязвимыми сервисами (SSH на нестандартных портах, устаревшие версии Apache). Это подтверждает эффективность автоматизированного подхода для первичной оценки рисков сетевой инфраструктуры.

Использование систем мониторинга для оценки рисков.

Еще одним важным инструментом для оценки рисков являются системы мониторинга, такие как Zabbix, Nagios или Prometheus. Эти системы позволяют непрерывно контролировать состояние ИТ-инфраструктуры и сигнализировать о проблемах, которые могут быть потенциальными рисками (Виттинг и Виттинг, 2023: 624).

Пример конфигурации правила мониторинга в Zabbix для выявления подозрительных подключений:

Создание шаблона для мониторинга подозрительных подключений по определенным портам

UserParameter=custom.tcp_conn[*],netstat -an | grep -w tcp | grep ':$1' | wc -l

Это правило отслеживает количество TCP-соединений на указанном порту. Если количество подключений резко возрастает, это может свидетельствовать о потенциальной атаке, и система сможет предупредить администратора (см. Рисунок 1).

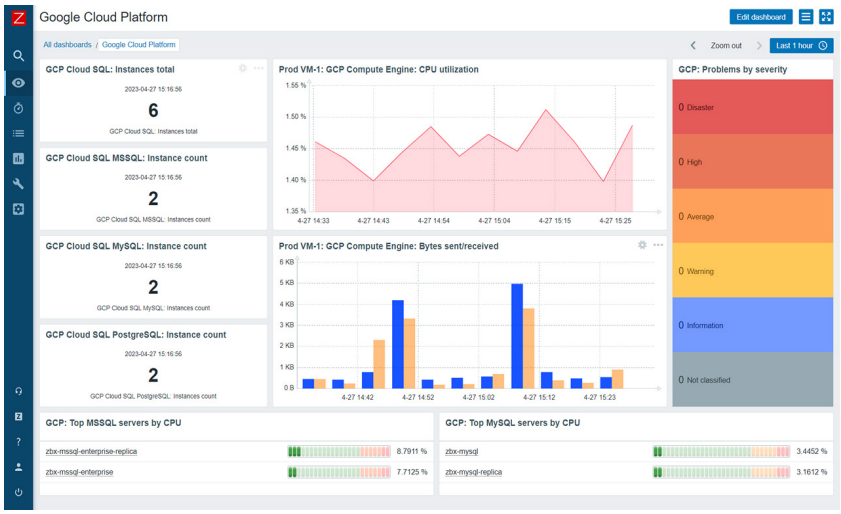


Рис. 1. Интерфейс системы мониторинга Zabbix (графики с динамическими данными по количеству подключений)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Количественная оценка рисков и математические модели

Для более точной оценки рисков часто используются математические модели. Например, метод Монте-Карло, о котором упоминалось ранее, позволяет провести симуляцию возможных исходов и получить более детализированное представление о рисках.

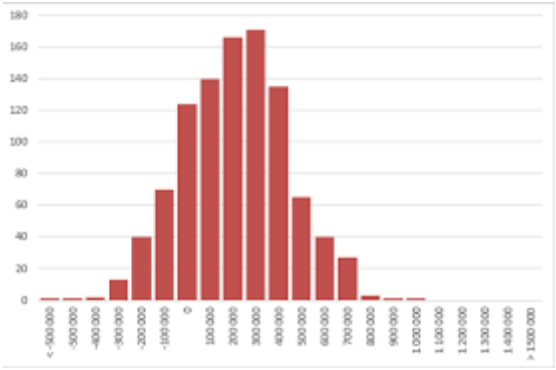


Рис. 2. График распределения вероятностей, полученный с помощью метода Монте-Карло

Пример использования Python для симуляции методом Монте-Карло:

```
import random
import matplotlib.pyplot as plt

# Симуляция убытков
def simulate_risk(trials, min_loss, max_loss):
    losses = []
    for _ in range(trials):
        loss = random.uniform(min_loss, max_loss)
        losses.append(loss)
    return losses

# Параметры симуляции
trials = 10000
min_loss = 1000 # минимальные убытки
max_loss = 50000 # максимальные убытки

# Проведение симуляции
simulated_losses = simulate_risk(trials, min_loss, max_loss)

# Построение графика
plt.hist(simulated_losses, bins=50, edgecolor='black')
plt.title('Распределение возможных убытков (Монте-Карло)')
plt.xlabel('Убытки')
plt.ylabel('Частота')
```

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

plt.show()

Этот скрипт моделирует возможные убытки в диапазоне от 1000 до 50000 и отображает их распределение на графике. Метод Монте-Карло позволяет выявить диапазон наиболее вероятных убытков, что критично для точного планирования затрат на безопасность (см. Рисунок 2). Результаты эксперимента: проведённая симуляция (10 000 итераций) показала следующее распределение убытков: среднее значение потерь составило 25 340 у.е., медиана — 25 120 у.е., стандартное отклонение — 14 150 у.е. При этом 95 % доверительный интервал убытков составил от 1 980 до 48 700 у.е. Данные результаты позволяют руководству организации планировать бюджет на информационную безопасность с учётом вероятностного распределения потерь.

Интеграция моделей оценки рисков в корпоративные процессы.

Для успешного управления рисками необходимо не только выбрать правильные методы и модели, но и интегрировать их в повседневную деятельность организации.

Использование системы GRC (Governance, Risk, and Compliance).

Современные системы GRC позволяют интегрировать процессы управления рисками с корпоративным управлением и нормативно-правовыми требованиями. Такие системы обеспечивают централизованное управление всеми аспектами безопасности и соответствия требованиям, что значительно упрощает работу.

Роль автоматизации и инструментов SIEM (Security Information and Event Management)

Системы SIEM автоматизируют сбор и анализ данных о событиях в ИТ-инфраструктуре, что позволяет своевременно выявлять инциденты безопасности и минимизировать риски. Примеры таких систем включают Splunk, ArcSight и IBM QRadar.

Пример настройки правила корреляции в SIEM:

Пример корреляции события доступа из необычного географического местоположения

rule correlating_login_events

condition: if login_event and (geo_location != «expected_location»)

action: alert «Подозрительное подключение» Результат корреляции событий визуализируется на панели SIEM (см. Рисунок 3).

Инструменты и платформы для оценки рисков информационной безопасности.

В современном мире доступно множество инструментов и платформ, предназначенных для автоматизации оценки и управления рисками. Эти системы помогают организациям эффективно проводить оценку рисков и принимать решения на основе полученных данных. Рассмотрим наиболее популярные из них.

OpenVAS.

OpenVAS — это мощный инструмент для сканирования уязвимостей,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

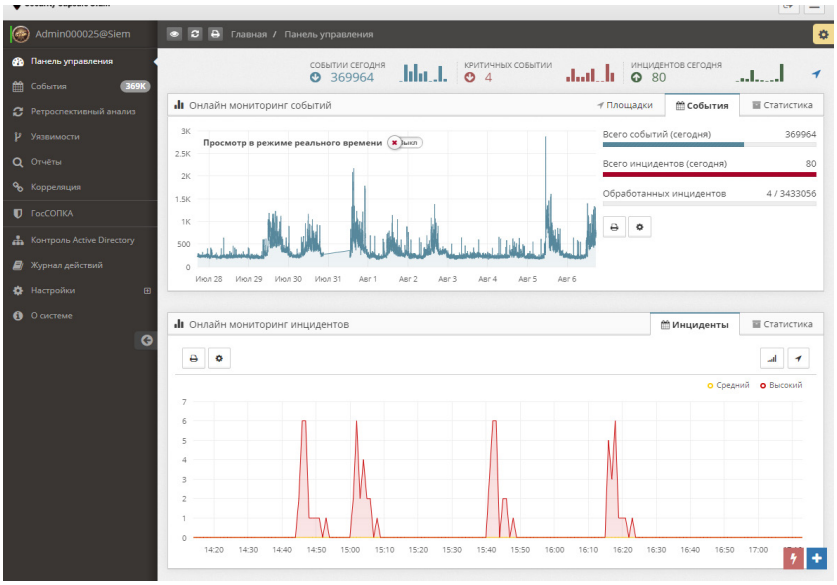


Рис. 3. Интерфейс системы SIEM с графиком корреляции событий безопасности

который позволяет выявлять возможные риски безопасности. Он является частью Greenbone Vulnerability Manager и предоставляет гибкие возможности для анализа сетевой безопасности (Холик Ф. и др, 2014: 183-188). Пример команды для сканирования сети с помощью OpenVAS:

Запуск сканирования уязвимостей в заданной сети
openvas -p 9390 -u admin -s 192.168.1.0/24

После выполнения команды OpenVAS проведет полное сканирование указанного диапазона IP-адресов и предоставит отчет о выявленных уязвимостях, что позволяет организации оперативно реагировать на потенциальные угрозы (см. Рисунок 4).

Date	Status	Task	Severity	Scan Results					Actions
				High	Medium	Low	Log	False Pos.	
Thu Jan 9 03:05:08 2020	Done	Immediate scan of IP 192.168.11.137	N/A	0	0	0	0	0	

Report: Results (312 of 734)

ID: 97cc63d0-65d7-45ee-8ca8-711df1baa7dd
Modified:
Created:
Owner: admin

1 - 100 of 312

Vulnerability	Severity	QoD	Host	Location	Actions
rexec Passwordless / Unencrypted Cleartext Login	10.0 (High)	75%	192.168.11.137	512/tcp	
Samba End Of Life Detection	10.0 (High)	75%	192.168.11.137	445/tcp	
Samba 'TALLOC_FREE()' Function Remote Code Execution Vulnerability	10.0 (High)	75%	192.168.11.137	445/tcp	
PHP Multiple Vulnerabilities - Aug08	10.0 (High)	75%	192.168.11.137	80/tcp	
PHP Version < 5.2.7 Multiple Vulnerabilities	10.0 (High)	75%	192.168.11.137	80/tcp	
PHP End Of Life Detection (Linux)	10.0 (High)	75%	192.168.11.137	80/tcp	
MySQL End Of Life Detection (Linux)	10.0 (High)	75%	192.168.11.137	3306/tcp	
PostgreSQL End Of Life Detection (Linux)	10.0 (High)	75%	192.168.11.137	5432/tcp	

Рис. 4. Скриншот отчета OpenVAS с выявленными уязвимостями

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Metasploit.

Metasploit — это фреймворк для проведения тестов на проникновение, который часто используется для оценки уязвимостей и моделирования атак. Он позволяет компаниям не только выявлять слабые места в системах, но и оценивать, как эти уязвимости могут быть использованы злоумышленниками.

Пример использования Metasploit для проверки уязвимости:

```
# Запуск эксплойта для уязвимости SMB
use exploit/windows/smb/ms08_067_netapi
set RHOST 192.168.1.100
set PAYLOAD windows/meterpreter/reverse_tcp
exploit
```

Этот скрипт на базе Metasploit позволяет провести тест на проникновение с использованием уязвимости SMB. Результаты тестирования помогают понять, как злоумышленники могут атаковать систему и какие меры защиты следует принять (см. Рисунок 5).

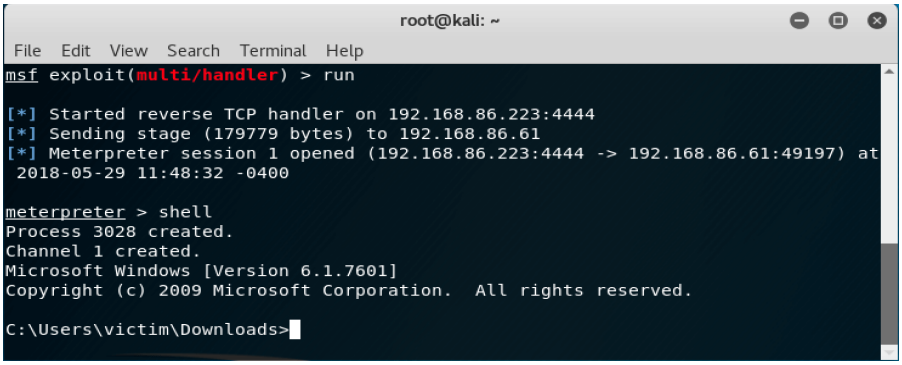


Рис. 5. Интерфейс Metasploit с результатами атаки на уязвимую систему

RiskWatch.

RiskWatch — это платформа для управления рисками, которая предлагает организациям инструменты для анализа, управления и отслеживания рисков в режиме реального времени. Она позволяет организациям автоматизировать процессы оценки рисков и создать отчетность, что значительно упрощает принятие решений (см. Рисунок 6).

Оценка рисков в облачных системах.

С переходом многих организаций на использование облачных технологий возникают новые вызовы в области информационной безопасности. Облачные системы требуют особого подхода к оценке рисков, поскольку они подразумевают совместное использование ресурсов, управление сторонними провайдерами и глобальный доступ.

Модель Shared Responsibility.

Модель разделенной ответственности (Shared Responsibility Model) — это принцип, на основе которого большинство облачных провайдеров распределяют

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

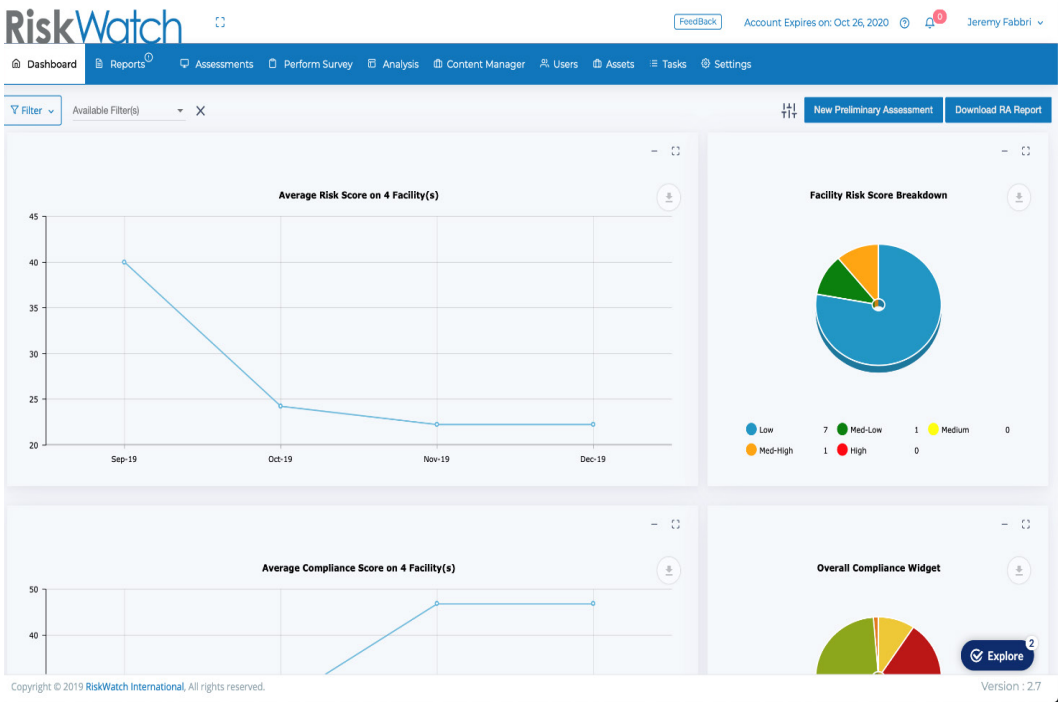


Рис. 6. Интерфейс платформы RiskWatch с таблицами анализа рисков

обязанности по безопасности между собой и клиентами. В этой модели провайдер облака отвечает за безопасность инфраструктуры, а клиент — за безопасность своих данных и приложений.

Пример практического применения оценки рисков в облачных системах:

Определение ответственности за данные: компания оценивает, за какие аспекты безопасности несет ответственность она, а за какие — облачный провайдер.

Анализ конфигураций безопасности: регулярный аудит настроек облачной инфраструктуры, включая политики доступа и шифрования данных (Милославская и др., 2014).

Пример автоматизации анализа конфигураций облака с помощью AWS Config:

Настройка проверки соответствия политик IAM в AWS

aws configservice put-config-rule --config-rule file://config-rule.json

Этот скрипт создает правило для проверки конфигурации политик IAM в AWS, что помогает организации следить за безопасностью облачной инфраструктуры и минимизировать риски.

Оценка рисков безопасности данных в облачных хранилищах.

Безопасность данных в облачных хранилищах — один из ключевых вопросов. Организации должны оценивать риски, связанные с нарушением конфиденциальности данных, утратой контроля над информацией и возможными

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

атаками на сторонние сервисы.

Пример конфигурации шифрования данных в облаке:

Пример включения шифрования на уровне базы данных в AWS RDS

aws rds modify-db-instance --db-instance-identifier mydbinstance --storage-encrypted

Данная команда активирует шифрование базы данных в облачном сервисе AWS RDS, что позволяет повысить уровень защиты хранимых данных и снизить риск их несанкционированного доступа или утечки.

Апробация методов оценки рисков: кейс лабораторного тестирования.

Для практической проверки эффективности рассмотренных методов было проведено лабораторное тестирование на базе учебной сети кафедры «Кибербезопасность и криптология» Казахского национального университета имени аль-Фараби.

Тестовая инфраструктура включала 5 серверов (Ubuntu 22.04 и Windows Server 2019), 20 рабочих станций и сетевое оборудование Cisco.

Этап 1. Автоматизированное сканирование (OpenVAS и Nmap)

В ходе сканирования тестовой сети было выявлено 47 уязвимостей, среди которых:

8 критических (CVSS ≥ 9.0);

15 высоких (CVSS 7.0–8.9);

24 средних (CVSS 4.0–6.9).

Среднее время полного автоматизированного сканирования составило 12 минут. Для сравнения, проведение аналогичного ручного аудита информационной безопасности экспертом занимало в среднем 4–6 часов, что свидетельствует о высокой эффективности автоматизации и позволяет сократить время анализа примерно в 20–30 раз.

Этап 2. Количественная оценка рисков (FAIR и метод Монте-Карло)

На основе выявленных уязвимостей была выполнена количественная оценка рисков с использованием методологии FAIR.

В качестве примера была рассмотрена критическая уязвимость CVE-2021-44228 (Log4Shell). Для расчёта были определены следующие параметры:

частота попыток атак — 15 попыток в месяц (на основе данных системы SIEM за последние 3 месяца);

вероятность успешной эксплуатации — 0,35;

ожидаемый диапазон финансовых потерь при успешной атаке — от 500 000 до 5 000 000 тенге.

Для моделирования распределения возможных потерь была проведена симуляция методом Монте-Карло с 10 000 итерациями.

Результаты моделирования показали, что:

средний ожидаемый годовой убыток (ALE) составляет 3 150 000 тенге;

значение 95-го перцентиля достигает 7 800 000 тенге, что отражает возможные потери при неблагоприятном сценарии развития инцидента.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Этап 3. Мониторинг и обнаружение угроз (SIEM и методы машинного обучения)

В течение 30 дней система SIEM (ELK Stack) осуществляла сбор и анализ сетевых событий. На основе накопленных данных была обучена модель машинного обучения Random Forest, предназначенная для выявления аномальной сетевой активности.

В результате анализа было выявлено 12 подозрительных паттернов, из которых 10 были подтверждены как реальные аномалии, что соответствует точности обнаружения 83,3 %.

После устранения обнаруженных уязвимостей и настройки правил межсетевого экрана было проведено повторное сканирование сети. Результаты показали значительное снижение количества уязвимостей: число критических уязвимостей сократилось с 8 до 1, а общее количество уязвимостей — с 47 до 18, что соответствует снижению на 61,7 %.

Таблица 2 — Результаты лабораторного тестирования

Показатель	До применения методов	После применения методов	Изменение
Критические уязвимости (CVSS ≥ 9.0)	8	1	–87,5%
Высокие уязвимости (CVSS 7.0–8.9)	15	6	–60,0%
Средние уязвимости (CVSS 4.0–6.9)	24	11	–54,2%
Общее количество уязвимостей	47	18	–61,7%
Среднее время обнаружения аномалии	4,2 часа	8 минут	–96,8%
Ожидаемый годовой убыток (ALE)	3 150 000 тг	890 000 тг	–71,7%

Полученные результаты подтверждают, что комбинированное применение автоматизированного сканирования уязвимостей, количественных методов оценки рисков и интеллектуальных систем мониторинга позволяет существенно снизить уровень рисков информационной безопасности и повысить обоснованность управленческих решений в области инвестиций в средства защиты информации.

Будущее оценки рисков информационной безопасности.

Современные технологии развиваются стремительно, и вместе с ними меняются и угрозы информационной безопасности. Поэтому модели и методы оценки рисков должны адаптироваться к новым реалиям. В ближайшем будущем можно ожидать усиленного применения искусственного интеллекта (ИИ) и машинного обучения (МО) для автоматизации процессов анализа рисков и предсказания возможных атак.

Применение ИИ и МО в оценке рисков.

Технологии ИИ и МО могут анализировать огромные объемы данных о событиях безопасности и на основе исторических данных предсказывать потенциальные угрозы. Это позволяет организациям лучше подготовиться к новым видам атак и своевременно адаптировать свои системы защиты.

Пример использования Python для анализа данных о сетевой безопасности

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

с помощью библиотеки scikit-learn:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
# Загрузка данных о сетевых событиях
data = load_security_data()  # Функция для загрузки данных
X = data[['feature1', 'feature2', 'feature3']]  # Факторы угроз
y = data['is_attack']  # Метка атаки
# Разделение данных на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
# Обучение модели RandomForest
model = RandomForestClassifier()
model.fit(X_train, y_train)
# Оценка точности модели
y_pred = model.predict(X_test)
print(f»Точность модели: {accuracy_score(y_test, y_pred)}»)
```

Этот скрипт демонстрирует, как с помощью машинного обучения можно анализировать данные о событиях безопасности и классифицировать их как атаки или обычные действия. Результаты эксперимента: модель Random Forest была обучена на наборе данных NSL-KDD, содержащем 125 973 записи сетевых событий. Точность классификации составила 94,7%, полнота (recall) — 92,3 %, F1-мера — 93,5 %. Для сравнения, модель SVM на том же наборе данных показала точность 91,2 %, что подтверждает преимущество ансамблевых методов для задач обнаружения аномалий в сетевом трафике (Ahmim и др., 2023).

Модели и стандарты для управления рисками информационной безопасности.

Для эффективного управления рисками информационной безопасности организации используют различные международные стандарты и модели, которые предоставляют руководства и рекомендации по оценке и управлению рисками (Сычев и др., 2017).

ISO/IEC 27005 — это международный стандарт, посвященный управлению рисками в области информационной безопасности. Он предоставляет четкие рекомендации по проведению процесса оценки рисков, начиная от идентификации угроз и уязвимостей до разработки плана реагирования на инциденты.

Процесс оценки рисков по ISO/IEC 27005 включает следующие этапы:

1. Идентификация активов: Определение всех информационных активов, которые необходимо защитить.

2. Определение угроз: Определение возможных угроз для каждого актива.

3. Анализ уязвимостей: Анализ слабых мест систем и процессов, которые могут быть использованы злоумышленниками.

4. Оценка последствий: Оценка воздействия инцидента на организацию

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

(например, финансовые потери, утрата репутации).

5. Оценка вероятности реализации угрозы: Определение вероятности того, что угроза будет реализована.

6. Оценка уровня риска: Определение уровня риска на основе вероятности и потенциальных последствий.

7. Разработка плана управления рисками: Определение мер по снижению или устранению рисков.

COBIT (Control Objectives for Information and Related Technologies) — это фреймворк для управления ИТ и защиты данных, разработанный ISACA. Он помогает организациям эффективно управлять рисками и обеспечивать соответствие требованиям безопасности.

Пример использования COBIT для оценки рисков:

1. Определение стратегических целей: COBIT связывает управление ИТ с бизнес-целями организации, что помогает лучше управлять рисками.

2. Оценка текущего состояния безопасности: Определение текущих показателей безопасности и сопоставление их с целевыми значениями.

3. Разработка плана улучшений: COBIT предлагает рекомендации по улучшению контроля и управления рисками, включая разработку политик безопасности, обучение сотрудников и внедрение технологий.

NIST Cybersecurity Framework — это гибкий фреймворк для управления рисками информационной безопасности, разработанный Национальным институтом стандартов и технологий США (NIST). Он состоит из пяти основных функций:

1. Идентификация: Определение активов, угроз и уязвимостей.

2. Защита: Разработка и внедрение мер защиты.

3. Обнаружение: Внедрение механизмов для обнаружения инцидентов безопасности.

4. Ответ: Реагирование на инциденты.

5. Восстановление: Меры по восстановлению после инцидентов и снижению их последствий.

Фреймворк широко используется в различных отраслях, включая правительство и частный сектор, и предлагает унифицированный подход к управлению киберрисками.

Пример использования NIST для оценки рисков:

Настройка правил обнаружения инцидентов в системе мониторинга

if security_incident_detected:

alert “Инцидент безопасности обнаружен. Выполнение плана реагирования»

Политики безопасности играют ключевую роль в управлении рисками, поскольку они определяют правила и процедуры, которые организация должна соблюдать для минимизации угроз и уязвимостей. Правильно разработанные политики не только снижают риски, но и помогают организациям соответствовать

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

нормативным требованиям.

Пример политики безопасности для предотвращения утечек данных:

Цель: Обеспечить защиту конфиденциальной информации и предотвратить несанкционированное использование данных.

Процедуры:

Все конфиденциальные данные должны быть зашифрованы.

Доступ к конфиденциальной информации должен предоставляться только авторизованным сотрудникам.

Все устройства, на которых хранится конфиденциальная информация, должны быть защищены паролями.

В случае инцидента, связанного с утечкой данных, немедленно уведомить службу ИБ.

Пример правила межсетевого экрана для предотвращения утечек данных

iptables -A OUTPUT -p tcp --dport 443 -d malicious.example.com -j DROP

Роль политики инцидент-менеджмента.

Политика инцидент-менеджмента определяет процесс реагирования на инциденты безопасности, начиная от их обнаружения и до завершения расследования. Она включает такие шаги, как:

-Идентификация инцидента.

-Оповещение заинтересованных сторон.

-Реагирование на инцидент (например, отключение от сети или блокировка доступа).

-Анализ причин инцидента.

-Восстановление системы.

-Разработка рекомендаций для предотвращения аналогичных инцидентов в будущем.

Пример процедуры инцидент-менеджмента в случае выявления вредоносной активности:

```
# Уведомление службы безопасности о подозрительных соединениях
if detect_malicious_connection():
    alert_security_team("Подозрительное соединение обнаружено")
    block_ip(«192.168.1.200»)
```

С развитием технологий, таких как искусственный интеллект, машинное обучение, Интернет вещей (IoT), возникает необходимость адаптации подходов к управлению рисками (Родичев и др., 2018).

Интернет вещей (Internet of Things, IoT) представляет собой совокупность взаимосвязанных устройств, подключённых к сети и способных обмениваться данными без непосредственного участия пользователя. Широкое распространение IoT-устройств существенно расширяет поверхность атаки и создаёт новые риски информационной безопасности.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Основные риски, связанные с использованием IoT-технологий, включают:

Возможность несанкционированного удалённого управления устройствами;

Наличие уязвимостей в прошивке и программном обеспечении устройств;

Недостаточную регулярность обновлений безопасности;

Возможность использования IoT-устройств в качестве точки входа для атак на другие элементы инфраструктуры.

Согласно отчёту ENISA Threat Landscape 2024, количество атак на IoT-устройства увеличилось на 87 % за последние два года, что подтверждает актуальность разработки эффективных методов оценки и управления рисками в данной области.

Для анализа и оценки рисков IoT-систем применяются специализированные методологические подходы.

Модель STRIDE для IoT позволяет классифицировать угрозы по следующим категориям: подмена (Spoofing), фальсификация данных (Tampering), отказ от авторства (Repudiation), раскрытие информации (Information Disclosure), отказ в обслуживании (Denial of Service) и повышение привилегий (Elevation of Privilege). В контексте IoT наиболее критичными считаются угрозы подмены устройств и фальсификации данных датчиков, поскольку они могут приводить к искажению информации, используемой для принятия управленческих решений.

Фреймворк OWASP IoT Top 10 определяет десять наиболее распространённых и критичных уязвимостей IoT-систем. К ним относятся использование слабых паролей по умолчанию, небезопасные сетевые интерфейсы, отсутствие механизмов обновления прошивки и недостаточная защита каналов передачи данных.

Метод количественной оценки рисков на основе CVSS v3.1 предполагает вычисление базового показателя уязвимости для каждого IoT-устройства с последующей корректировкой с учётом факторов среды эксплуатации. К таким факторам относятся уровень сетевой изоляции устройства, использование защищённых каналов связи (например, VPN), а также регулярность обновления прошивки.

В качестве примера автоматизированного аудита IoT-устройств можно рассмотреть использование языка программирования Python и сервиса Shodan для выявления устройств с потенциальными уязвимостями.

Пример автоматизированного поиска IoT-устройств

```
import shodan
api = shodan.Shodan('API_KEY')
# Поиск IoT-устройств в указанной сети
results = api.search('port:554 has_screenshot:true net:192.168.1.0/24')
for result in results['matches']:
    print(f»IP: {result['ip_str']}, Port: {result['port']}, Org: {result.get('org', 'N/A')}»)
    if result.get('vulns'):
        print(f»Уязвимости: {', '.join(result['vulns'])}»)
```

Представленный подход позволяет автоматически выявлять IoT-устройства

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

с известными уязвимостями, что способствует более эффективной приоритизации объектов для последующего анализа и устранения выявленных угроз (Alhomoud и др., 2024).

Пример настройки политики безопасности для IoT-устройств:

```
# Настройка доступа к IoT-устройствам через VPN
iptables -A INPUT -p tcp --dport 22 -s vpn_gateway_ip -j ACCEPT
```

Системы на основе ИИ и МО становятся важной частью ИТ-инфраструктуры. Однако они также несут новые риски, такие как:

Манипуляция данными для обучения модели (data poisoning).

Уязвимости, связанные с неточными прогнозами.

Непреднамеренные предвзятости в алгоритмах.

Для систематической оценки рисков ИИ-систем используются следующие специализированные подходы:

NIST AI Risk Management Framework (AI RMF 1.0, 2023): предлагает структурированный подход к управлению рисками ИИ по четырём функциям — Map (картирование контекста), Measure (измерение рисков), Manage (управление рисками) и Govern (управление на уровне организации).

Adversarial Robustness Toolbox (ART): открытая библиотека для оценки устойчивости моделей машинного обучения к состязательным атакам (adversarial attacks). Позволяет моделировать атаки типа FGSM, PGD, C&W и оценивать их влияние на точность модели.

Метрики оценки рисков ИИ: помимо стандартных метрик (accuracy, precision, recall), для оценки рисков ИИ-систем критичны метрики устойчивости (robustness score), справедливости (fairness metrics — demographic parity, equalized odds) и объяснимости (interpretability — SHAP, LIME).

Пример оценки устойчивости модели к состязательным атакам:

```
from art.attacks.evasion import FastGradientMethod
from art.estimators.classification import SklearnClassifier
# Обёртка модели для ART
classifier = SklearnClassifier(model=trained_model)
# Генерация состязательных примеров
attack = FastGradientMethod(estimator=classifier, eps=0.3)
x_adv = attack.generate(x=X_test)
# Оценка устойчивости
accuracy_clean = accuracy_score(y_test, model.predict(X_test))
accuracy_adv = accuracy_score(y_test, model.predict(x_adv))
print(f»Точность на чистых данных: {accuracy_clean:.3f}»)
print(f”Точность на adversarial данных: {accuracy_adv:.3f}”)
print(f”Снижение точности: {(accuracy_clean - accuracy_adv)*100:.1f}%”)
```

В ходе лабораторного тестирования атака FGSM снизила точность модели Random Forest с 94,7 % до 78,2 %, что демонстрирует необходимость учёта adversarial-рисков при развёртывании ИИ-систем в контуре информационной безопас-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ности (Кузнецов и Петров, 2022).

Пример использования ИИ для автоматизации анализа безопасности:

```
# Пример использования модели машинного обучения для анализа подозрительных логов
from sklearn.svm import SVC
# Загрузка данных
logs = load_security_logs()
# Обучение модели
model = SVC(kernel='linear')
model.fit(logs['features'], logs['labels'])
# Предсказание инцидентов
predictions = model.predict(new_logs['features'])
```

Заключение.

Оценка и управление рисками информационной безопасности — это непрерывный и многослойный процесс, включающий различные этапы, от выявления активов и угроз до реализации мер по снижению рисков и реагированию на инциденты. Развитие технологий, таких как искусственный интеллект, машинное обучение, облачные технологии и Интернет вещей, создает как новые возможности, так и дополнительные вызовы в области безопасности.

Методы и модели оценки рисков, такие как ISO/IEC 27005, NIST Cybersecurity Framework и COBIT, предоставляют компаниям стандартизированные подходы для управления информационной безопасностью, помогая формировать устойчивую систему защиты от угроз. Инструменты для сканирования уязвимостей (OpenVAS), тестирования на проникновение (Metasploit), а также платформы для управления рисками (RiskWatch) помогают автоматизировать многие аспекты этого процесса, что значительно повышает эффективность работы.

На протяжении этой работы мы рассмотрели не только общие принципы управления рисками, но и углубились в технические аспекты: показали примеры скриптов, которые могут быть использованы для защиты данных и оценки рисков в реальных системах. Эти примеры, вместе с теоретическими основами, дают читателю понимание того, как теория управления рисками применяется на практике.

Кроме того, необходимо подчеркнуть растущую важность кибергигиены и подготовки персонала. Даже при наличии самых совершенных технических средств слабым звеном часто остаётся человеческий фактор. Поэтому обучение сотрудников вопросам ИБ и проведение регулярных учений по реагированию на инциденты должны стать обязательной частью стратегии управления рисками.

Большое значение имеет и юридический аспект. Компании обязаны учитывать не только внутренние угрозы, но и соблюдать нормативные требования в сфере защиты персональных данных, включая международные стандарты (GDPR, HIPAA и др.), что напрямую связано с правовой и репутационной безопасностью бизнеса.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Наконец, в условиях постоянной цифровой трансформации организациям необходимо строить адаптивную и проактивную модель управления рисками, ориентированную на предотвращение угроз, а не только на реагирование.

Ключевые выводы:

Управление рисками должно основываться на четком понимании активов, угроз и уязвимостей.

Использование фреймворков и стандартов (ISO/IEC 27005, NIST, COBIT) помогает систематизировать процесс оценки и управления рисками.

Автоматизация с помощью OpenVAS, Metasploit, RiskWatch и SIEM-систем существенно повышает точность и скорость оценки рисков.

Облачные и гибридные ИТ-среды требуют переосмысления традиционных подходов к защите и внедрения модели разделенной ответственности.

Человеческий фактор остается ключевым элементом в обеспечении ИБ — необходимо инвестировать в обучение и контроль.

Будущее управления рисками — за применением ИИ, аналитики больших данных и предиктивных моделей для проактивной кибербезопасности. Проведённый сравнительный анализ методов по пяти критериям (точность, масштабируемость, трудоёмкость, автоматизируемость, повторяемость) подтвердил выдвинутую гипотезу: комбинированное использование количественных методов (FAIR, Монте-Карло) совместно с инструментами автоматизации (OpenVAS, SIEM) обеспечивает наиболее полную и объективную оценку рисков. Экспериментальные результаты показали, что автоматизированное сканирование сети позволяет за менее чем минуту выявить критические уязвимости, а модели машинного обучения достигают точности свыше 94% в задачах обнаружения аномалий. Результаты лабораторного кейса продемонстрировали снижение общего числа уязвимостей на 61,7%, сокращение времени обнаружения аномалий с 4,2 часов до 8 минут и уменьшение ожидаемого годового убытка на 71,7%, что подтверждает практическую применимость предложенного комбинированного подхода.

ЛИТЕРАТУРА

Аксу М., Алтунджу Э., Бичакчи К. (2019). Первый взгляд на удобство использования сканера уязвимостей OpenVAS // Семинар по пригодной безопасности (USEC). — Сан-Диего, США: Internet Society. С. 8. 10.14722/usec.2019.23026 // ISBN 1-891562-57-6 [На англ].

Алгулиев Р.М., Имамвердиев Я.Н., Набиев Б.Р. (2021). Методы и модели оценки рисков информационной безопасности: систематический обзор // Проблемы информационной безопасности. Компьютерные системы. — Баку: AMEA. № 3. С. 45–52 [на рус].

Ahmim A., Maglaras L., Ferrag M.A. et al. (2023). A Novel Hierarchical Intrusion Detection System Based on Decision Tree and Rules-Based Models // Distributed Computing and Artificial Intelligence. — Cham: Springer. С. 1125–1138. 10.1007/s10586-019-02988-4 [На англ].

Alhomoud A., Munir R., Disso J.P. et al. (2024). Performance Evaluation of Modern IDS Tools Against Advanced Persistent Threats // Future Internet. — Basel: MDPI. — Том. 16. — No. 1. С. 1–18. 10.3390/fi16010012 [на англ].

Виттинг А., Виттинг М. (2023). Amazon Web Services in Action: An In-Depth Guide to AWS. — New York: Simon and Schuster. С. 624. ISBN 978-1-61729-511-9 [На англ].

Иванченко П.Ю., Кацуро Д.А., Медведев А.В., Трусов А.Н. (2013). Математическое моделирование информационной и экономической безопасности в малых и средних предприятиях // Fundamental Research. — Новосибирск: Академия Естествознания. № 10–13. С. 2860–2863. ISSN 1812-7339 [На рус].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Корченко А.Г., Архипов А.Е., Казмирчук С.В. (2013). Анализ и оценка рисков информационной безопасности. — Киев: Изд-во НАУ. С. 148. ISBN 978-966-598-966-9 [На рус].

Кузнецов Д.А., Петров С.В. (2022). Применение методов машинного обучения для автоматизации оценки рисков информационной безопасности // Вестник кибербезопасности. — М: РАН, 2022. — Том. 10. — № 2. С. 34–45. 10.25559/VCYBER.2022.10.2.003 [На рус].

Максименко В.Н., Ясюк Е.В. (2017). Основные подходы к анализу и оценке рисков информационной безопасности // Экономика и качество систем связи. — М: ФГБОУ ВО МТУСИЮ — Т. 2. — № 4. С. 42–48. ISSN 2410-9916 [На рус].

Миков Д.А. (2014). Анализ методов и средств, используемых на различных этапах оценки рисков информационной безопасности // Вопросы кибербезопасности. — М: НТЦ «Академия». № 4 (7). С. 49–54. ISSN 2311-3456 [На рус].

Милославская Н.Г., Сенаторов М.Ю., Толстой А.И. (2013). Управление рисками информационной безопасности. — М.: Горячая линия – Телеком. С. 130. ISBN 978-5-9912-0339-6 [На рус].

Плетнев П.В., Белов В.М. (2012). Методология оценки рисков информационной безопасности в малом и среднем бизнесе // Доклады Томского государственного университета систем управления и радиоэлектроники. — Томск: ТУСУР. № 1–2 (25). С. 83–86. ISSN 1818-0442 [На рус].

Родичев Ю.А. (2018). Нормативная база и стандарты в области информационной безопасности. — М: Кибербезопасность. С. 104. ISBN 978-5-4461-1234-2 [На рус].

Сычев Ю.Н. (2017). Стандарты информационной безопасности. Защита и обработка конфиденциальных документов. — М.: Инфра-М. С. 207. ISBN 978-5-16-113218-0 [На рус].

Холик Ф., и др. (2014). Эффективное тестирование на проникновение с использованием фреймворка и методологий Metasploit // Материалы 15-го Международного симпозиума IEEE по вычислительному интеллекту и информатике (CINTI). — Будапешт, Венгрия: IEEE. С. 183–188. 10.1109/CINTI.2014.7028673 [На англ].

Шабалина О.А., Ковалёв Д.О. (2024). Адаптивные модели управления рисками информационной безопасности в условиях цифровой трансформации // Информатика и автоматизация // СПб.: СПИИРАН. — Том. 23. — № 1. С. 112–128. ISSN 2713-3192 [На рус].

ENISA. (2024). Threat Landscape 2024: Top Threats and Trends // Athens: European Union Agency for Cybersecurity. С. 120. 10.2824/0710888[На англ].

REFERENCES

Aksu M., Altuncu E., Bicakci K. (2019). A First Look at the Usability of the OpenVAS Vulnerability Scanner // Workshop on Usable Security (USEC). — San Diego, USA: Internet Society. — Vol. 8. — 10.14722/usec.2019.23026. — ISBN 1-891562-57-6 [in Eng].

Ahmim A., Maglaras L., Ferrag M.A. et al. (2023). A Novel Hierarchical Intrusion Detection System Based on Decision Tree and Rules-Based Models // Distributed Computing and Artificial Intelligence. — Cham: Springer. Pp. 1125–1138. — 10.1007/s10586-019-02988-4 [in Eng].

Alhomoud A., Munir R., Disso J.P. et al. (2024). Performance Evaluation of Modern IDS Tools Against Advanced Persistent Threats // Future Internet. — Basel: MDPI. — Vol. 16. — No. 1. Pp. 1–18. 10.3390/fi16010012 [in Eng].

Alguliyev R.M., Imamverdiyev Y.N., Nabiyev B.R. (2021). Methods and Models for Information Security Risk Assessment: A Systematic Review // Problems of Information Security. Computer Systems. — Baku: AMEA. No. 3. Pp. 45–52 [in Rus].

ENISA. (2024). Threat Landscape: Top Threats and Trends. — Athens: European Union Agency for Cybersecurity. — Vol. 120. — 10.2824/0710888 [in Eng].

Holik F., et al. (2014). Effective Penetration Testing Using the Metasploit Framework and Methodologies // Proceedings of the 15th IEEE International Symposium on Computational Intelligence and Information Science (CINTI). — Budapest, Hungary: IEEE. Pp. 183–188. 10.1109/CINTI.2014.7028673 [in Eng].

Ivanchenko P. Yu., Katsuro D. A., Medvedev A. V., Trusov A. N. (2013). Mathematical Modeling of Information and Economic Security in Small and Medium Enterprises // Fundamental Research. — Novosibirsk: Academy of Natural Sciences. No. 10–13. Pp. 2860–2863. ISSN 1812-7339 [in Rus].

Korchenko A.G., Arkhipov A.E., Kazmirchuk S.V. (2013). Analysis and assessment of information security risks. — Kyiv: Publishing house of NAU. Pp. 148. ISBN 978-966-598-966-9 [in Rus].

Kuznetsov D.A., Petrov S.V. (2022). Application of Machine Learning Methods for Automation of Information Security Risk Assessment // Cybersecurity Bulletin. — Mo: RAS. — Vol. 10. — No. 2. Pp. 34–45. 10.25559/VCYBER.2022.10.2.003 [in Rus].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Maksimenko V.N., Yasyuk E.V. (2017). Basic approaches to the analysis and assessment of information security risks // Economics and quality of communication systems. — M: FGBOU VO MTUCI. — Vol. 2. — No. 4. Pp. 42–48. ISSN 2410-9916 [in Rus].

Mikov D.A. (2014). Analysis of methods and tools used at various stages of information security risk assessment // Cybersecurity issues. — Moscow: NTC "Academy". No. 4 (7). Pp. 49–54. ISSN 2311-3456 [in Rus].

Miloslavskaya N.G., Senatorov M.Yu., Tolstoy A.I. (2013). Information Security Risk Management. — M: Goryachaya Liniya Telecom. Pp 130. ISBN 978-5-9912-0339-6 [in Rus].

Pletnev P.V., Belov V.M. (2012). Methodology for Assessing Information Security Risks in Small and Medium Business // Reports of Tomsk State University of Control Systems and Radioelectronics. — Tomsk: TUSUR. No. 1–2 (25). Pp. 83–86. ISSN 1818-0442 [in Rus].

Rodichev Yu.A. (2018). Regulatory Framework and Standards in the Field of Information Security. — M: Cybersecurity. Pp 104. ISBN 978-5-4461-1234-2 [in Rus].

Sychev Yu. N. (2017). Information security standards. Protection and processing of confidential documents. — M.: Infra-M. Pp 207. ISBN 978-5-16-113218-0 [in Rus].

Shabalina O.A., Kovalev D.O. (2024). Adaptive Information Security Risk Management Models in the Context of Digital Transformation // Informatics and Automation // St. Petersburg: SPIIRAS. — Vol. 23. No. 1. Pp. 112–128. ISSN 2713-3192 [in Rus].

Witting A., Witting M. (2023). Amazon Web Services in Action: An In-Depth Guide to AWS. — New York: Simon and Schuster. Pp 624. ISBN 978-1-61729-511-9 [in Eng].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 453–474
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.027
УДК 004.056.5

EDGE COMPUTING-BASED TECHNIQUE FOR CONSTRUCTION OF ATTACK DETECTION MEANS IN CYBER-PHYSICAL SYSTEMS OF INDUSTRIAL INTERNET-OF-THINGS

T. K. Zhukabayeva[1*], D.B. Baumuratova[2], E. Benkhelifa [3], N.A. Niyetbayeva[4]
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Astana International University, Astana, Kazakhstan;
[3]Staffordshire University, Staffordshire, United Kingdom;
[4]M.Kh. Dulaty Taraz University, Taraz, Kazakhstan.
E-mail: tamara.kokenovna@gmail.com

Tamara K. Zhukabayeva — PhD, Professor of the Department of Information Systems, Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: tamara.kokenovna@gmail.com, https://orcid.org/0000-0001-6345-5211;
Dilaram B. Baumuratova — PhD, Senior Lecturer, Pedagogical Institute, Astana International University, Astana, Kazakhstan
https://orcid.org/0009-0009-4621-1886;
Elhadj Benkhelifa — PhD, Professor of Computer Science and Artificial Intelligence, Staffordshire University, Staffordshire, United Kingdom
https://orcid.org/0000-0001-6168-2664;
Nadira A. Niyetbayeva — PhD, Associate Professor, Department of Physics and Informatics, M.Kh. Dulaty Taraz University, Taraz, Kazakhstan
https://orcid.org/0000-0003-2921-6879.

© T.K. Zhukabayeva, D.B. Baumuratova E. Benkhelifa, N.A. Niyetbayeva

Abstract. The article examines security issues of contemporary cyber-physical systems that use the concept of edge computing to solve problems of secure operation of industrial Internet of Things infrastructures. The main contribution of this article comprises a description and results of the analysis of the proposed technique for detecting attacks in cyber-physical systems of the Industrial Internet of Things using edge computing. The technique is aimed at application by design engineers and developers of software packages to ensure information security of cyber-physical systems of the Industrial Internet of Things, where a significant part of the target computing processes of the system is imposed on the end devices of the system.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The technique includes six main stages covering the processes of analytical and natural-simulation modeling of attacks, generation and marking of initial data sets, construction of software classifiers as means of attack detection, and visual data analysis. In general, the implementation of the technique is presumed at the following stages of the life cycle of cyber-physical systems, these are the stages of designing and testing the system, setting up and evaluating the operation quality of attack detection tools. The feasibility of the technique using an example of an industrial system in the field of incident management of transport infrastructure using software and hardware modules of the Arduino platform confirms the correctness and effectiveness of the technique for its further practical application.

Keywords: attack, detection, edge computing, analysis, technique

For citation: T.K. Zhukabayeva, D.B. Baumuratova E. Benkhelifa, N.A. Niyetbayeva (2026). Edge computing-based technique for construction of attack detection means in cyber-physical systems of industrial internet-of-things // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 453-474. https://doi.org/10.54309/IJICT.2026.25.1.027. (In Russ),

Conflict of interest: The authors declare that there is no conflict of interest.

ШЕКАРАЛЫҚ ЕСЕПТЕУЛЕРДІ ҚОЛДАНА ОТЫРЫП, ЗАТТАРДЫҢ ӨНЕРКӘСІПТІК ИНТЕРНЕТІНІҢ КИБЕРФИЗИКАЛЫҚ ЖҮЙЕЛЕРІНДЕГІ ШАБУЫЛДАРДЫ АНЫҚТАУ ҚҰРАЛДАРЫН ҚҰРУ ӘДІСТЕМЕСІ

Т.К. Жукабаева[1*], Д. Б Баумуратова[2], Е. Бенкхелифа[3], Н.А. Ниетбаева[4]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;

[2]Астана халықаралық университеті, Астана, Қазақстан;

[3]Стаффордшир университеті, Стаффордшир Ұлыбритания;

[4]М.Х. Дулати атындағы Тараз университеті, Тараз, Қазақстан.

E-mail: tamara.kokenovna@gmail.com

Жукабаева Тамара Кокеновна — PhD, ақпараттық технологиялар факультетінің Ақпараттық жүйелер кафедрасының профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

E-mail: tamara.kokenovna@gmail.com, https://orcid.org/0000-0001-6345-5211;

Баумуратова Диларам Бекбулатовна — PhD, Астана халықаралық университетінің Педагогикалық институтының аға оқытушысы, Астана, Қазақстан

https://orcid.org/0009-0009-4621-1886;

Бенхелифа Эльхадж — PhD, компьютерлік ғылымдар және жасанды интеллект профессоры, Стаффордшир университеті, Стаффордшир Ұлыбритания

https://orcid.org/0000-0001-6168-2664;

Ниетбаева Надира Ашировна — PhD, физика және информатика кафедрасының қауымдастырылған профессоры, М.Х. Дулати атындағы Тараз университеті, Тараз,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Қазақстан
https://orcid.org/0000-0003-2921-6879.

© Т.К. Жукабаева, Д.Б. Баумуратова, Е. Бенкхелифа, Н.А. Ниетбаева

Аннотация. Мақалада заттардың өнеркәсіптік интернеті инфрақұрылымдарының қорғалатын жұмысының мәселелерін шешу үшін шекаралық есептеу тұжырымдамасын қолданатын заманауи киберфизикалық жүйелердің қауіпсіздігі мәселелері қарастырылады. Мақаланың негізгі үлесі шекаралық есептеулерді қолдана отырып, заттардың өнеркәсіптік интернетінің киберфизикалық жүйелеріндегі шабуылдарды анықтаудың ұсынылған әдістемесін сипаттау мен талдау нәтижелерін қамтиды. Әдістеме инженер-дизайнерлер мен бағдарламалық жасақтама жасаушылардың өнеркәсіптік интернет заттарының киберфизикалық жүйелерінің ақпараттық қауіпсіздігін қамтамасыз ету үшін қолдануға бағытталған, онда жүйенің мақсатты есептеу процестерінің маңызды бөлігі жүйенің соңғы құрылғыларына жүктеледі. Әдістеме шабуылдаушы әсерлерді аналитикалық және заттай Имитациялық модельдеу, бастапқы деректер жиынтығын құру және белгілеу, шабуылдарды анықтау құралы ретінде бағдарламалық классификаторларды құру, деректерді визуалды талдау процестерін қамтитын алты негізгі кезеңді қамтиды. Жалпы, әдістемені орындау киберфизикалық жүйелердің өмірлік циклінің келесі кезеңдерінде – жүйені жобалау және тестілеу, шабуылдарды анықтау құралдарының жұмыс сапасын реттеу және бағалау кезеңдерінде қарастырылады. Arduino платформасының бағдарламалық-аппараттық модульдерін пайдалана отырып, көлік инфрақұрылымының инциденттерін басқару саласындағы индустриялық жүйе мысалында Әдістеменің орындылығы оны одан әрі практикалық қолдану үшін Әдістеменің дұрыстығы мен пәрменділігін растайды.

Түйін сөздер: шабуыл, анықтау, шекаралық есептеу, талдау, әдістеме

Дәйексөздер үшін: Т.К. Жукабаева, Д.ББаумуратова, Е. Бенкхелифа, Н.А. Ниетбаева (2026). Шекаралық есептеулерді қолдана отырып, заттардың өнеркәсіптік интернетінің киберфизикалық жүйелеріндегі шабуылдарды анықтау құралдарын құру әдістемесі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т 7. № 25. 453-474 бет. https://doi.org/10.54309/IJICT.2026.25.1.027. (Орыс тіл.);

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

МЕТОДИКА ПОСТРОЕНИЯ СРЕДСТВ ОБНАРУЖЕНИЯ АТАК В КИБЕРФИЗИЧЕСКИХ СИСТЕМАХ ПРОМЫШЛЕННОГО ИНТЕРНЕТА ВЕЩЕЙ С ИСПОЛЬЗОВАНИЕМ ГРАНИЧНЫХ ВЫЧИСЛЕНИЙ

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Т.К. Жукабаева[1*], Д Б. Баумуратова[2], Е. Бенкхелифа[3], Н.А. Ниетбаева[4]
[1]Евразийскии национальный университет имени Л.Н. Гумилева, Астана, Казахстан;
[2]Международный университет Астана, Астана, Казахстан;
[3]Стаффордшир университет, Стаффордшир, Великобритания;
[4]Таразский университет имени М.Х. Дулати, Тараз, Казахстан.
E-mail: tamara.kokenovna@gmail.com

Жукабаева Тамара Кокеновна — PhD, профессор кафедры информационные системы, факультет информационных технологий, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан
E-mail: tamara.kokenovna@gmail.com, https://orcid.org/0000-0001-6345-5211;
Баумуратова Диларам Бекбулатовна — PhD, старший преподаватель, Педагогический институт Международного университета Астана, Астана, Казахстан
https://orcid.org/0009-0009-4621-1886;
Бенкхелифа Эльхадж — PhD, профессор компьютерных наук и искусственного интеллекта, Стаффордширский университет, Стаффордшир, Великобритания
https://0000-0001-6168-2664;
Ниетбаева Надира Ашировна — PhD, ассоцированный профессор, кафедра физики и информатики, Таразский университет имени М.Х. Дулати, Тараз, Казахстан
https://orcid.org/0000-0003-2921-6879.

© Т.К. Жукабаева, Д.Б. Баумуратова, Е. Бенкхелифа, Н.А. Ниетбаева

Аннотация. В статье исследуются вопросы безопасности современных киберфизических систем, использующих концепцию граничных вычислений для решения задач защищенного функционирования инфраструктур промышленного интернета вещей. Основной вклад статьи включает описание и результаты анализа предложенной методики обнаружения атак в киберфизических системах промышленного интернета вещей с использованием граничных вычислений. Методика ориентирована на применение инженерами-проектировщиками и разработчиками программных комплексов для обеспечения информационной безопасности киберфизических систем промышленного интернета вещей, в которых значимая часть целевых вычислительных процессов системы возлагается на конечные устройства системы. Методика включает шесть основных стадий, охватывающих процессы аналитического и натурно-имитационного моделирования атакующих воздействий, генерации и разметки наборов исходных данных, построения программных классификаторов в качестве средств обнаружения атак, визуального анализа данных. В целом выполнение методики предусматривается на следующих этапа жизненного цикла киберфизических

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

систем – этапах проектирования и тестирования системы, настройки и оценивания качества работы средств обнаружения атак. Выполнимость методики на примере индустриальной системы в области управления инцидентами транспортной инфраструктуры с использованием программно-аппаратных модулей платформы Arduino подтверждает корректность и действенность методики для ее дальнейшего практического применения.

Ключевые слова: атака, обнаружение, граничные вычисления, анализ. методика

Для цитирования: Т.К. Жукабаева, Д.Б. Баумуратова, Е. Бенкхелифа, Н.А. Ниетбаева (2026). Методика построения средств обнаружения атак в киберфизических системах промышленного интернета вещей с использованием граничных вычислений // Международный журнал информационных и коммуникационных технологий. Т 7. No. 25. Стр. 453–474. https://doi.org/10.54309/IJICT.2026.25.1.027.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Финансирование. *Настоящая работа сотрудниками НАО «Евразийский национальный университет имени Л.Н. Гумилева» проводится при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (Грант № AP23489127).*

Введение.

В настоящее время все большее распространение на практике получают различные информационно-телекоммуникационные инфраструктуры, включающие в свой состав разнородные киберфизические и мобильные устройства, автоматизированные системы управления такими инфраструктурами и интеллектуальные сервисы, предоставляемые конечным пользователям и организующие высоконадежные и защищенные межмашинные взаимодействия. Киберфизические системы и инфраструктуры, реализующие концепцию граничных вычислений (edge; edge computing) (Ilyin, 2021), представляют сетевые распределенные структуры автономно работающих встроенных и мобильных устройств интернета вещей с возможностью обработки первичных данных непосредственно на стороне конечных устройств с последующей передачей обработанных данных на централизованные сетевые хосты и облачные системы (Shirazi, 2017; Esposito, 2017). Согласно этой концепции, значительная часть вычислений и обработки данных должна выполняться либо непосредственно в местах их сбора или в непосредственной их близости. Ввиду не доверенности программно-информационного окружения таких систем, а также уязвимостей используемого программно-аппаратного обеспечения киберфизических устройств, в том числе уязвимостей нулевого дня (Anwer, 2022), а также недостаточной защищенности существующих коммуникационных протоколов, в частности, протоколов канального, сетевого и

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

прикладного уровней, такие инфраструктуры оказываются подверженными разнообразным атакующим воздействиям, направленным на компрометацию устройств, данных, циркулирующих по сети и хранящихся на устройствах, а также предоставляемых пользовательских сервисов (Yahuza, 2020).

Отметим, что особенная сложность формирования надежных и защищенных механизмов киберфизической безопасности систем, реализующих концепцию граничных вычислений, возникает в результате следующих факторов, непосредственно влияющих на уровень защищенности таких систем: многошаговость, многоаспектность и многовариантность действий потенциального нарушителя информационной безопасности (ИБ). Многошаговость включает наличия типовых и узкоспециализированных сценариев действий атакующего, разделяемых логически и во временном исчислении на явно выделенные этапы, такие как

- предваряющий анализ доступного атакующему программно-технического окружения с определением перечней информационных активов, воздействия на которые могут служить достижению его целей. В частности, на данном этапе может проводиться сетевое сканирование коммуникационной инфраструктуры с использованием сканеров безопасности Nmap (Asokan, 2023), Nessus (Muin, 2022), OpenVAS (Toyin, 2023) и др., позволяющих выявить нужные уязвимости в программно-аппаратном обеспечении, незащищенные открытые порты, архитектурные слабости и другие характерные особенности инфраструктуры, влияющие на процесс подготовки несанкционированного воздействия;

- получение доступа к целевым программно-аппаратным компонентам в рамках заданной сетевой коммуникационной и/или виртуальной инфраструктуры, «продвижение» по сети с использованием метода последовательного повышения привилегий в условиях накопления данных и ресурсов, достаточных для проведения воздействия;

- осуществление несанкционированного воздействия на целевой объект приложения с модификацией его структуры или конфигурационных настроек, прослушиванием и/или перехватом его данных, нарушением его доступности и др.

- опциональное удаление следов присутствия нарушителя в рамках заданной программно-информационной инфраструктуры.

Многоаспектность атакующих воздействий состоит в возможностях нарушителя использовать в процессе атаки одновременно или последовательно нескольких целевых приложений атаки, взаимодополняющих друг друга и включающих воздействия на уровне сетевых хостов, аппаратно-физические воздействия на критически важные узлы инфраструктуры, социоинженерные воздействия и иные проявления. Так, в общем случае комбинированный характер подобного воздействия позволяет атакующему не только усилить эффект атаки, но и сократить временные затраты на проведение атаки, в том

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

числе сократить время поиска слабых мест и уязвимостей, которые он будет эксплуатировать в качестве своих стартовых шагов (Yang, 2024). Вытекающая из многоаспектности многовариантность действий атакующего обуславливает возможность динамического выбора им наиболее выгодных шагов в зависимости от текущего контекста атаки, осведомленности нарушителя об интересуемых его активах и условий их функционирования (Golchin, 2022). Все это определяет потребность в совершенствовании программных средств обнаружения атак и повышении защищенности киберфизических систем для улучшения показателей качества обнаружения атак и улучшения нефункциональных характеристик средств защиты за счет комплексного учета условий функционирования целевой инфраструктуры, моделирования анализа действий атакующего и особенностей технологии граничных вычислений.

Настоящая работа ориентирована на совершенствование существующих и разработку новых перспективных программных средств обнаружения атак в современных киберфизических информационно-телекоммуникационных системах и сетях, базирующихся на концепции граничных вычислений. Основной вклад данной статьи включает разработанную методику обнаружения атак в киберфизических системах промышленного интернета вещей (IIoT) с использованием граничных вычислений, а также результаты ее анализа. Характерными примерами систем промышленного интернета вещей являются индустриальные системы мониторинга периметра производства и контроля качества индустриального процесса; различные транспортные киберфизические системы, организующие автоматизированные сценарии сортировки, хранения и доставки производственных деталей при их изготовлении; системы безопасных и эффективных интеллектуальных процессов генерации и распределения электроэнергии для индустриальных предприятий –так называемые, умные сети электроснабжения (Smart Grid). При этом передача части значимого функционала таких систем на сторону географически распределенных и удаленных устройств не только расширяет функциональность устройств и предоставляемых сервисов, а также улучшает нефункциональные характеристики, но и представляет далеко идущую тенденцию в совершенствовании и повышении целевых показателей качества таких систем в целом. Вместе с тем во всех указанных примерах проблематика безопасности функционирования киберфизичекой инфраструктуры, а также отдельных входящих в ее состав устройств, узлов, действующих пользователей представляется крайне актуальной, в особенности в условиях потенциально не доверенного и ненадежного окружения устройств.

Выделим следующие основные факторы, определяющие актуальность и практическую значимость задач по разработке методик по построению средств обнаружения атак в IIoT-системах с использованием граничных вычислений. К ним относятся в первую очередь разнородность IIoT-систем и их уязвимость к широкому классу несанкционированных воздействий, а также

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

атаки непосредственно на граничные устройства и функции граничных вычислений. В частности, в настоящее время наблюдается существенная разнородность существующих систем промышленного интернета вещей, включающих в свой состав отличающиеся наборы конечных и промежуточных устройств, сенсоров, исполнительных механизмов, различные сетевые конфигурации, виды аппаратных архитектур и протоколов сетевого взаимодействия. Поэтому Разнообразие IIoT-систем обуславливает потребность в унификации и построении средств обнаружения атак в таких системах с учетом, вариативности действий потенциального нарушителя, его практических возможностей, доступных ресурсов и инструментов (Vankayalapati, 2023).

Ввиду распределенного характера IIoTсистем, наличия ограничений ресурсопотребления их устройств, открытости существующих сетевых протоколов, наличия слабых мест в используемом программном обеспечении, такие системы оказываются уязвимыми к актуальным угрозам информационной безопасности. В том числе это касается воздействий, представляющим, в частности, такие атаки как сетевые атаки DoS, MitM, различные фишинговые атаки; атаки на конкретные устройства, включающие атаки подбора пароля методом направленного перебора и эксплуатацию уязвимостей программных прошивок устройств; атаки на модификацию данных от сенсоров; атаки уровня приложений, такие как XSS-атаки, SQL-инъекции и др.(Xiao, 2019).

Поэтому разнообразие возможных видов атак, а также их взаимосвязанность обуславливают потребность в комплексном выявлении атак в IIoT-системах, которое должно охватывать различные аспекты обеспечения безопасности: моделирование атак, построение программных классификаторов и визуализацию данных, что способствует обеспечению всестороннего анализа возможных угроз безопасности (Roman, 2018). Кроме того, отметим, что использование IIoT-системой граничных вычислений позволяет устройствам эффективно обрабатывать данные непосредственно на конечных устройствах, что повышает оперативность реагирования на угрозы. Вместе с тем возможность несанкционированной эксплуатации функций граничных вычислений позволяет потенциальному злоумышленнику осуществлять воздействия непосредственно на граничные устройства системы, в том числе атаки утечки данных с устройств, backdoor-атаки и атаки нарушения аутентификации.

Поэтому, указанные выше особенности IIoT-систем позволяют подтвердить, что предлагаемая методика является важным инструментом для повышения уровня информационной безопасности киберфизических систем промышленного интернета вещей, базирующихся на использовании граничных вычислений, что особенно актуально в условиях растущего числа угроз, увеличивающейся сложности современных технологических решений. Таким образом, своевременное обнаружение инцидентов безопасности в таких инфраструктурах с высоким качеством представляется крайне важной задачей.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Оставшаяся часть статьи организована следующим образом. Следующий раздел статьи включает обзор и анализ существующих механизмов информационной безопасности и обнаружения атак в киберфизических системах индустриального интернета вещей в рамках концепции граничных вычислений. Далее в статье раскрываются сущность и особенности предложенной методики обнаружения атак. Последующий раздел статьи посвящен вопросам применения и анализу данной методики. Статья завершается заключением и списком основных использованных источников научно-технической литературы.

Граничные вычисления, называемые также периферийными вычислениями, представляют концепцию распределенных вычислений, которые осуществляются в границах некоторого множества конечных устройств, функционирующих в заданной киберфизической инфраструктуре (Lin, 2020). Другими словами, граничные вычисления предполагают организацию хранения данных, вычисления и их фактическое расположение в некоторой окрестности имеющихся устройств и источников первичных данных. Это в свою очередь способствует снижению временных задержек при передаче данных в таких инфраструктурах, а также увеличивает пропускную способность сетевых коммуникационных каналов связи. Таким образом, к основным преимуществам систем граничных вычислений можно отнести следующие:

- снижение объемов, передаваемых данных;
- повышение оперативности функционирования системы за счет принятия решений по управлению системой непосредственно на устройствах (узлах сети) и снижению коммуникационных задержек;
- повышение надежности и бесперебойности работы системы, в том числе за счет повышения степени автономности отдельных устройств и их сегментов в условиях временных нарушений связности используемой коммуникационной сети, а также уменьшения критичности централизованных модулей обработки данных;
- повышение уровня безопасности системы за счет возможностей по отслеживанию аномалий и выявлению атак непосредственно на устройствах сети.

Научная проблема, на исследование и решение которой направлена настоящая работа, состоит в недостаточной защищенности систем, реализующих концепцию граничных вычислений, и их подверженности сложным для выявления комбинированным многошаговым информационным воздействиям, направленным на некорректное и нецелевое использование таких систем, нарушение корректного функционирования таких систем, причинение ущерба их инфраструктуре и пользователям. Отметим также, что понятие граничных вычислений введено для выделения подкласса систем интернета вещей, для которых действия по обработке данных могут выполняться в пределах локального сетевого контура некоторой группы пользовательских устройств (Dolui et al., 2018). В частности, в системах, реализующих концепцию граничных вычислений,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

информационные сервисы могут располагаться на конечных пользовательских устройствах или устройствах, представляющих точки доступа к связи с устройствами пользователей. При этом процедуры такой распределенной инфраструктуры граничных вычислений выполняются на вычислительных модулях, максимально приближенных к местам расположения считываемым с сенсоров данных о людях, процессах, вещах. То есть, фактически, такой набор вычислительных узлов позволяет реализовать функциональность облачных вычислений (cloud computing), но не централизованно на высокопроизводительных серверах, а «приближенную к земле», формируя более быстрые ответы на информационные импульсы со стороны собираемых в IoT-системе данных (Nam, 2023).

Отметим также, что граничные и облачные вычисления могут применяться совместно. При этом граничные вычисления формируют дополнительный слой управления между сенсорами, как источниками данных, и облаком, как вычислительным слоем, отвечающим за обработку и хранение больших массивов данных. Облачные вычисления освобождают организации от решения множеств технологических вопросов, таких как вопросы хранения данных, вычислительных и сетевых ограничений, при этом они в текущем их виде с большим трудом позволяют справляться с требованиями на поддержку мобильности, осведомленность о местонахождении и низкие коммуникационные задержки, предъявляемыми со стороны пользовательских приложений (Qing, 2018).

К особенностям проблематики информационной безопасности систем граничных вычислений можно отнести зачастую опосредованный и отложенный характер несанкционированного воздействия, выражаемый, в том числе, в постепенной деградации каналов связи за счет разрастающегося вовлечения имеющихся edge-узлов в botnet-атаку (Gulatas, 2023). Это способно приводить к постепенному перераспределению вычислений и связанному с этих ухудшению скорости связи и показателей корректности доставки сообщений (показатели Quality-of-Service). Кроме того, сложность обнаружения атак в таких системах связана с нехваткой централизации при сборе данных, которые могут содержать важные признаки, необходимые для обнаружения атак.

В (Gulatas, 2023) отмечается, что помимо того, что системы граничных вычислений наследуют классы уязвимостей от предшествующих коммуникационно-вычислительных технологий, таких как распределенные P2P-системы и беспроводные сенсорные сети, за счет многоуровневой структуры граничных вычислений и ограниченности ресурсов устройств такие системы обладают дополнительными наборами уязвимостей, опирающимися на изъяны отдельных edge-узлов и их взаимодействия. Кроме того, в (Alwarafy, 2019) обосновывается важность вопросов обеспечения защищенности и приватности данных в системах граничных вычислений.

На примере нескольких практических сценариев, таких как системы электронной медицина и умных городов, в (Caprolu, 2020) освещаются основные вопросы безопасности граничных вычислений, связанные, в том числе, с

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

применением технологий виртуализации к edge-системам, как с применением уязвимостей контейнерных инфраструктур, так и без них. В частности, показано, что архитектурные особенности инфраструктуры таких сценариев формируют наборы характерных им специфических программно-аппаратных уязвимостей, которые могут быть успешно эксплуатированы потенциальным нарушителем в рамках таких атакующих воздействий как удаленное выполнение кода, DoS-атаки и различные атаки переполнение (flooding-атаки), атаки сканирования портов и уязвимостей, атаки повышение привилегий, утечки данных и др.

Таким образом, в условиях отсутствия унифицированных средств обнаружения атак в киберфизических системах промышленного интернета вещей с использованием граничных вычислений, а также необходимой адаптивности таких механизмов под требования и условия функционирования конкретных сценариев выполнения конкретных систем с использованием граничных вычислений возникает необходимость разработки комплексной методики построения средств обнаружения атак, которая должна учитывать основные архитектурные и сценарные особенности граничных вычислений. К основным отличиям предлагаемой в настоящей работе методики можно отнести учет специфики граничных вычислений на всех основных стадиях методики, включающих проведение аналитического моделирования, натурно-имитационного моделирования, генерацию тестовых и обучающих наборов данных и проведение визуального анализа данных.

Материалы и Методы.

Методика построения средств обнаружения атак. Предлагаемая методика ориентирована на построение механизмов обнаружения атак в киберфизических системах промышленного интернета вещей с использованием граничных вычислений с учетом специфики структуры и особенностей функционирования таких инфраструктур. Предлагаемая методика включает выполнение следующих основных шести стадий, осуществление которых обеспечивает решение задач построения средств обнаружения атак в киберфизических системах промышленного интернета вещей с использованием граничных вычислений (Рис. 1). Данная методика предназначена для инженеров-проектировщиков и разработчиков программных комплексов для обеспечения информационной безопасности киберфизических систем промышленного интернета вещей, в которых значимая часть вычислительных процессов бизнес-логики системы возлагается на конечные устройства системы. Выполнение методики предполагается на этапах проектирования, тестирования, настройки и оценивания качества работы средств обнаружения атак. На рисунке 1 обозначены основные входные и выходные данные, при этом стадии методики представлены в виде прямоугольников, тогда как стрелки между ними формируют контуры управления и передачи данных между стадиями.

Входом методики являются формальная спецификация анализируемой

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

IIoT-системы, включающая набор функциональны требований и нефункциональных ограничений, а также перечень угроз информационной безопасности, связанных с ожидаемыми разновидностями атак, которые необходимо детектировать в процессе защищенного функционирования IIoT-системы.

Выходом методики является реализованный программный компонент, корректность функционирования которого подтверждается на основе эмпирических проверок – тестирования качества программных классификаторов и экспертного оценивания с использованием средств визуального анализа данных о работе построенных программных классификаторов.

На стадии 1 производится построение аналитической модели атакующих воздействий на IIoT-систему, реализующую концепцию граничных вычислений. На основе имеющихся спецификаций и перечня актуальных угроз для целевой системы такое моделирование предполагает получение результатов анализа по определению предполагаемых целей и мотивов нарушителя. При этом в общем случае нарушитель способен эксплуатировать как свойства распределенного сбора и обмена информацией между периферийными устройствами IIoT-системы, так и уязвимости самих устройств. Также определяются типовые сценарии нарушителя с уточнением отдельных шагов, включающих, воздействия, как физического характера, так и программно-информационного. Отметим, что идентифицируются также доступные ресурсы и используемые атакующим программно-аппаратные средства. Кроме того, в процессе проводимого анализа также выясняются стартовые возможности нарушителя и завязанные на это устройства и программно-аппаратные интерфейсы – места осуществления доступа, атакующего к системе.

В целях учета динамических особенностей функционирования целевой системы стадия 2 методики предполагает использованием методов натурного и имитационного моделирования. В частности, для получения исходных данных, адекватным образом описывающих одну или несколько различных видов атак, проанализированных на стадии 1 методики, закладывается формирование физической полнофункциональной или до определенной степени ограниченной натурной модели сети, включающей ряд целевых и обеспечивающих электронно-вычислительных и периферийных устройств, датчиков, связующего сетевого оборудования и других электронных компонентов. Ввиду возможной практической сложности подобного моделирования часть функциональности модели предполагается возможной к реализации за счет имитационного представления.

Стадия 2 предлагаемой методики включает также возможность поиска существующих наборов данных, включающих описание логов индустриальной системы интернета вещей. Такие наборы данных могут применяться, как для обогащения данных, формируемых в рамках методики, так и в качестве положительных примеров для формирования новых наборов данных с использованием имеющейся натурно-имитационной модели.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

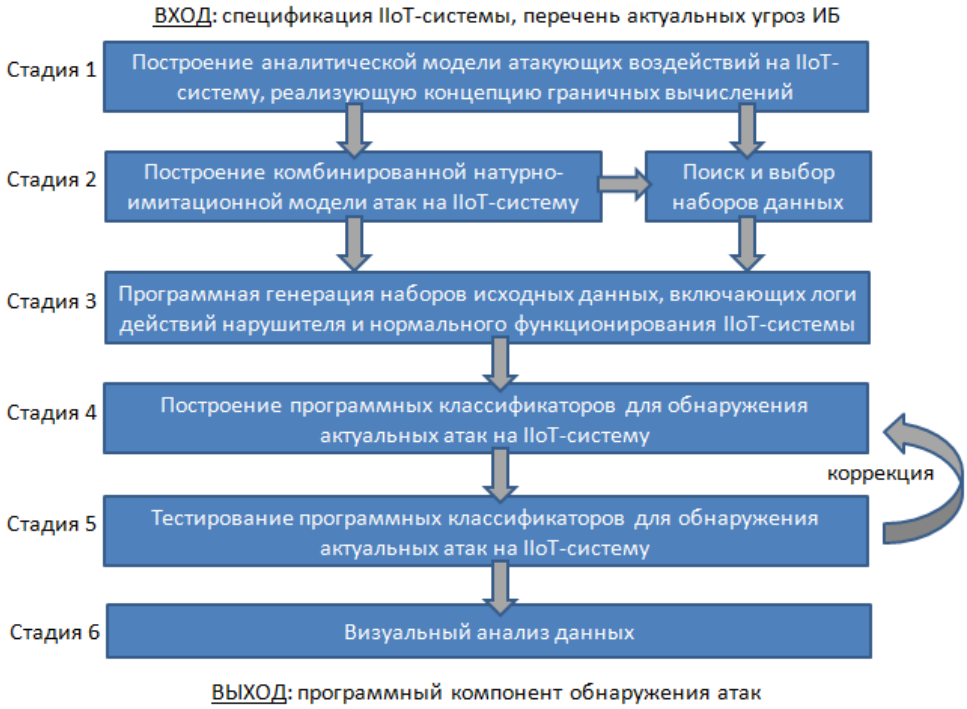


Рис.1. Схема методики построения средств обнаружения атак.

На стадии 3 осуществляется программная генерация наборов исходных данных, включающих логи действий нарушителя и нормального функционирования IIoT-системы. Формируемые на этой стадии исходные данные требуются для построения программных модулей обнаружения атак, а также для осуществления разметки данных по классам атак. Искомые наборы данных предполагается построить в рамках экспериментов с использованием натурно-имитационной модели, как в случае нахождения IIoT-системы под атакой, так и в случае ее нормального функционирования. Фактически, основой для такой генерации является запуск сценариев функционирования натурно-имитационной модели на наборах стартовых параметров модели с использованием правил управления устройствами граничных вычислений. На данной стадии также производится задание разметки в рамках генерируемых наборов исходных данных, которая в общем случае включает указание временных периодов моделирования атаки, а также физических и/или сетевых адресов устройств, вовлеченных в моделируемый сценарий в зависимости от используемых канальных, сетевых и прикладных протоколов, по которым происходит взаимодействие устройств граничных вычислений. Примером актуального вида атак на устройства граничных вычислений являются атаки отказа в обслуживании (Gulatas, 2023), атаки ransomware-шифрования (Job, 2021), botnet-атаки, как например Mirai (Febro, 2022), DSN poisoning-атаки

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

(Gulatas, 2023).

Стадия 4 охватывает построение программных классификаторов для обнаружения актуальных видов атак на IIoT-систему с использованием методов машинного обучения с учителем, включающих, в том числе, следующие методы: случайный лес, деревья решений, k-ближайших соседей, adaboost-классификатор, машина опорных векторов, LSTM и другие. На данной стадии возможно также применение дополнительных алгоритмов комбинирования классификаторов, включающих стекинг, мажоритарное голосование, а также алгоритмов сэмплинга – в случае необходимости балансировки обучающих и тестовых выборок. В частности, комбинирование бинарных классификаторов позволяет организовать эффективный мультиклассификатор по различным классам атакующих воздействий.

Стадия 5 включает проверку корректности построенных на стадии 4 программных классификаторов на тестовых выборках данных с вычислением значений точности, полноты, F1-меры и других классификационных показателей. В случае невыполнимости требований на показатели качества классификации производится возврат к стадии 4 с измененными значениями гипер-параметров классификационных методов и/или уточнением самих методов.

Стадия 6 включает проведение экспертного анализа построенных классификаторов на имеющихся наборах данных с использованием визуального анализа исходных данных и интерпретации результатов работы классификаторов. В частности, данная стадия методики предполагает использование методов уменьшения объемов и размерности анализируемых данных с применением алгоритма главных компонент (PCA) и алгоритма независимого компонентного анализа (ICA).

Обсуждение и результаты.

Применение методики и дискуссия. Обобщим основные полученные результаты данной работы. Конечной целью данного исследования является разработка эффективных методов управления производственными процессами для повышения качества продукции и оптимизации затрат. Работа предлагает целостный подход к обеспечению безопасности в киберфизических системах, интегрируя методы машинного обучения и экспертный анализ для повышения надежности и устойчивости к атакам. Статья посвящена разработке методики обнаружения атак в киберфизических системах промышленного интернета вещей, основанных на концепции граничных вычислений. В работе подчеркиваются актуальные проблемы безопасности в таких системах, обусловленные сложностью и разнообразием возможных атак, а также необходимостью учитывать распределенный характер инфраструктуры граничных вычислений.

Таким образом, предложена новая методика обнаружения атак, включающая, в частности, генерацию релевантных наборов данных, построение программных классификаторов и их проверку с использованием показателей качества классификации, экспертный анализ и интерпретация результатов с

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

использованием методов снижения размерности данных. Методика включает комбинацию различных алгоритмов машинного обучения для построения эффективного мульти-классификатора, способного обнаруживать широкий спектр атак. Разработанная методика предназначена для унифицированного подхода к защите разнородных IIoT-систем, учитывая разнообразие устройств, протоколов и атак.

В рамках данного исследования используются комплексный подход, включающий следующие основные методы: статистический анализ — проводится сбор и обработка данных о основных показателях IIoT-системы; методы аналитического и натурно-имитационного моделирования для представления и анализа процессов IIoT-системы и возможных атак на ее устройства; методы машинного обучения и визуального анализа данных в качестве основы комбинированного обнаружения атак.

Научная новизна методики заключается в комплексном подходе к обнаружению атак в IIoT-системах с использованием граничных вычислений, применении натурно-имитационной модели для генерации данных и комбинировании различных методов машинного обучения для создания универсального мульти-классификатора. Также учитывается разнородность систем и применяются методы снижения размерности данных для повышения точности и интерпретируемости результатов. Результаты исследований могут быть применены для улучшения безопасности промышленных киберфизических систем, минимизируя риски несанкционированного доступа и потери данных.

Отметим, что конкретный практический результат заключается в разработке методики, которую инженеры и разработчики смогут использовать для защиты IIoT-инфраструктуры от различных видов атакующих воздействий. Данная методика охватывает широкий спектр задач: от анализа потенциальных угроз до реализации механизмов их предотвращения. Она применима на этапе проектирования и тестирования систем, а также при настройке и оценке работы защитных средств.

Важным аспектом является использование натурно-имитационного моделирования, которое позволяет моделировать атаки и исследовать их воздействие на систему. Такой подход повышает точность и надежность методик защиты, поскольку он учитывает разнообразные сценарии атак.

Применение предложенной в работе методики производится на примере IIoT-системы в области управления инцидентами транспортной инфраструктуры, где устройствами граничных вычислений являются автономные программно-аппаратные модели дистанционно управляемых колесных робототехнических устройств на основе модулей платформы Arduino и совместимых с ней электронных компонентов. К отличительным особенностям методики, отличающей ее от альтернативных наработок и решений в предметной области комплексный учет граничных вычислений на протяжении методики и ее основных стадий. Это выражается, в том числе, в свойствах мобильности устройств граничных

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

вычислений, возможности их пространственного перемещения, изменчивости способов коммуникации и статистического распределения характера процессов сетевого взаимодействия (Ray, 2020; Goel, 2020). В свою очередь, это обуславливает наличие на таких устройствах узкоспециализированных уязвимостей, связанных с недостаточной защищенностью edge-устройства, и подверженностью актуальным видам атак на него. Таким образом, статья предлагает не только теоретическую основу, но и проверенную на практике методологию, готовую к внедрению в реальных проектах. Ниже приведен псевдокод, иллюстрирующий обобщенный алгоритм, лежащий в основе предложенной методики обнаружения атак с использованием граничных вычислений. Данный код специфицирует структуру методики и последовательность шагов, которые необходимы для выполнения методики с учетом специфики конкретной IIoT-системы, ее устройств и используемого инструментария. Алгоритм записан в процедурном виде императивного стиля программирования, символ # означают текст комментария, поясняющего конкретную команду и для удобства выделенный курсивом.

```
def simulate_attacks(): # Шаг 1: Аналитическое и натурно-имитационное моделирование атак
    simulated_attacks = generate_attack_scenarios() # Генерация типов атак (DDoS, Man-in-the-Middle и др.)
    simulation_results = run_simulation(simulated_attacks) # Моделирование атак на тестируемой системе
    return simulation_results
def prepare_datasets(): # Шаг 2: Генерация и разметка наборов данных
    raw_data = collect_sensor_data() # Сбор данных с датчиков и систем мониторинга
    labeled_data = label_data(raw_data) # Разметка данных: нормальные данные и аномальные (данные и описывающие атаки)
    return labeled_data
def build_classifier(data): # Шаг 3: Построение классификатора для обнаружения атак
    model_type = select_model_algorithm() # Выбор подходящего алгоритма машинного обучения
    trained_model = train_model(model_type, data) # Обучение модели на размеченных данных
    return trained_model
def visualize_data(data): # Шаг 4: Визуальный анализ данных
    plots = create_plots_and_diagrams(data) # Создание графиков и диаграмм для наглядного представления данных
    analysis_results = analyze_visualizations(plots) # Анализ визуальных данных для выявления аномалий
    return analysis_results
def evaluate_classifier(model, test_data): # Шаг 5: Настройка и оценка каче-
```

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ства работы классификатора

```
    accuracy = calculate_accuracy(model, test_data) # Оценка точности классификатора на тестовых данных
    thresholds = set_thresholds(accuracy) # Определение пороговых значений для классификации атак
    return thresholds
def integrate_and_test(system, classifier): # Шаг 6: Интеграция и тестирование в промышленной среде
    integrated_system = deploy_classifier(classifier, system) # Интеграция классификатора в существующую систему
    test_results = perform_real_world_tests(integrated_system) # Тестирование интегрированной системы в реальных условиях
    return test_results
def main(): # Основная функция выполнения методики
    attack_simulations = simulate_attacks()
    datasets = prepare_datasets(attack_simulations)
    classifier = build_classifier(datasets)
    visual_analysis = visualize_data(datasets)
    evaluation_results = evaluate_classifier(classifier, datasets)
    integration_results = integrate_and_test(existing_system, classifier)
    print(«Методика выполнена. Результаты:», integration_results)
```

Предлагаемое в рамках методики натурно-имитационное моделирование может проводиться в рамках заданных начальных параметров устройств и/или процессов с применением системы правил, учитывающих возможные состояния системы и переходы между ними. Также проводится запуск такой имитационной модели на некотором множестве входных данных, зависящих от фактического выполнения используемой натурной составляющей модели. В частности, имитационная компонента такого моделирование позволяет упростить формирование распределенной функциональности сбора и обмена данными между устройствами граничных вычислений с минимизацией организационно-технических усилий по настройке и обработке функций граничных вычислений. Таким образом, в целом натурно-имитационные представления позволяют более точно и с ограниченными объемами ресурсов моделировать функционал целевой IIoT-системы, наиболее существенные поведенческие особенности устройств и пользователей системы.

Поэтому моделирование осуществляется с меньшими ресурсными и временными затратами, как в условиях нормального функционирования IIoT-системы, так и в условиях функционирования при нахождении системы под одной или одновременно несколькими атаками. При этом по результатам аналитического моделирования предполагается ранжировать установленные виды атак по степени их критичности для данного вида систем и выбрать наиболее актуальные виды несанкционированных воздействий для их последующего анализа. Отме-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

тим, в частности, что комбинированный характер модели выражается в расширении натурной модели за счет применения имитационного моделирования части напрямую сложно моделируемых/конфигурируемых стадий определенной атаки.

В общем случае для проведения такого анализа данных в рамках предложенной методики, в зависимости от структуры и особенностей анализируемых данных может потребоваться применение дополнительных предваряющих методов предварительной обработки данных, включающих стандартизацию данных, нормализацию, фильтрацию и устранение пропущенных и/или ошибочных значений отдельных полей. Отметим, что потребность в подобной фильтрации может возникать, в том числе, по причинам возможного спонтанного динамического характера функционирования устройств граничных вычислений, доступность которых может нарушаться на определенные периоды времени в рамках штатной работы IIoT-системы.

Отметим также, что в результате применения граничных вычислений, несмотря на возможность снижения объемов, пересылаемых по сети данных, тем не менее, поверхность атаки IIoT-системы с реализацией граничных вычислений может в целом увеличиться по сравнению с системами, базирующимися на концепции облачных вычислений. В частности, концентрация данных на удаленно обрабатывающих edge-узах может способствовать повышению рисков утечки таких данных (Qiang, 2021). В частности, такие утечки могут происходить не только на фазе непосредственной их обработки, но также и в дальнейшей работе при их последующем хранении для обеспечения целей кэширования данных (Ghosh, 2021).

Кроме того, отметим, что использование имитационной составляющей в процессе моделирования позволяет осуществить генерацию данных и следующее за этим интеллектуальное обнаружение атак с изолированных edge-узлов централизовано, без вов, лечения таких вычислительных концепций, как федеративное вычисление и другие (Singh, 2023; Fenanir et al., 2023; Yang, 2023).

Выводы.

В рамках проведённого исследования разработана комплексная методика построения средств обнаружения атак в киберфизических системах промышленного интернета вещей, функционирующих на основе концепции граничных вычислений. Предложенный подход направлен на системное повышение уровня информационной безопасности IIoT-инфраструктур за счёт интеграции аналитического моделирования угроз, натурно-имитационного воспроизведения сценариев атак, методов машинного обучения и экспертной интерпретации результатов. В отличие от фрагментарных решений, ориентированных исключительно на применение отдельных алгоритмов обнаружения или анализ ограниченного набора атак, представленная методика формирует целостную технологическую цепочку построения и верификации механизмов защиты.

Одним из ключевых результатов работы является обоснование необходимости учёта архитектурной специфики граничных вычислений при

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

разработке средств обнаружения атак. В условиях переноса значительной части вычислительной нагрузки на периферийные устройства возрастает роль локального анализа данных и оперативного реагирования на инциденты. Вместе с тем расширяется поверхность атаки за счёт распределённости узлов и ограниченности их вычислительных ресурсов. Предложенная методика учитывает данные особенности и ориентирована на построение адаптивных механизмов обнаружения, способных функционировать в условиях динамически изменяющейся сетевой среды.

Существенное значение имеет включение в структуру методики стадии аналитического моделирования действий потенциального нарушителя. Это позволяет формализовать возможные сценарии атак с учётом их многошагового и комбинированного характера, определить критические точки воздействия и сформировать требования к будущим средствам обнаружения. Такой подход обеспечивает проактивный характер защиты, при котором средства обнаружения разрабатываются не только на основе уже известных инцидентов, но и с учётом потенциальных эволюционных изменений угроз.

Натурно-имитационное моделирование, являющееся центральным элементом методики, позволяет воспроизводить реальные условия функционирования IIoT-систем и моделировать поведение как легитимных пользователей, так и нарушителей. Формирование экспериментальной среды с использованием программно-аппаратных модулей, включая устройства на базе платформы Arduino, обеспечивает практическую применимость методики и приближает экспериментальные результаты к реальным условиям эксплуатации. Имитационная составляющая позволяет масштабировать моделирование без значительного увеличения затрат, что особенно важно при анализе распределённых систем с большим числом устройств.

Важным вкладом работы является формирование процедуры генерации и разметки наборов данных для обучения и тестирования программных классификаторов. В условиях недостатка публичных датасетов для IIoT-среды данный этап имеет принципиальное значение. Разработанная процедура обеспечивает формирование сбалансированных выборок, отражающих как нормальное состояние системы, так и различные типы атакующих воздействий. Это создаёт основу для построения устойчивых моделей машинного обучения, способных выявлять аномалии в распределённых средах.

Применение методов машинного обучения, включая алгоритмы ансамблирования, позволяет повысить точность обнаружения атак и обеспечить мультиклассовую классификацию угроз. Комбинирование различных моделей способствует снижению вероятности ложноположительных и ложноотрицательных срабатываний, что критически важно для промышленных систем, где ошибки обнаружения могут привести к существенным экономическим и технологическим последствиям. При этом учёт ограничений вычислительных ресурсов граничных устройств позволяет

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

адаптировать модели к реальным условиям эксплуатации без чрезмерного роста энергопотребления и задержек обработки.

Проведённая апробация методики на примере системы управления инцидентами транспортной инфраструктуры продемонстрировала её применимость и масштабируемость. Полученные результаты подтверждают, что интеграция интеллектуальных механизмов обнаружения атак на уровне граничных устройств способствует повышению устойчивости системы к распределённым сетевым и прикладным воздействиям. Методика может быть использована как на этапе проектирования и тестирования IIoT-систем, так и при модернизации уже функционирующих инфраструктур, что соответствует принципам безопасной разработки и эксплуатации (Security-by-Design и Security-by-Default).

Дополнительно следует отметить перспективность дальнейшего развития методики в направлении интеграции федеративных подходов к обучению моделей, повышения интерпретируемости решений классификаторов и расширения экспериментальной базы за счёт использования реальных промышленных данных. Перспективным представляется исследование вопросов устойчивости моделей к атакам на сами алгоритмы машинного обучения, включая adversarial-воздействия, а также разработка механизмов динамической перенастройки классификаторов в процессе эксплуатации.

Таким образом, разработанная методика формирует унифицированную, адаптивную и практически ориентированную основу для построения средств обнаружения атак в киберфизических системах промышленного интернета вещей с использованием граничных вычислений. Её внедрение способствует повышению надёжности и устойчивости распределённых промышленных инфраструктур, снижению рисков компрометации граничных узлов и обеспечению безопасного функционирования критически важных технологических процессов в условиях усложняющейся киберугрозной среды.

REFERENCES

Anwer M., Ahmed G., Akhunzada A., Amin R. (2022). Comparative Analysis of Soft Computing Approaches of Zero-Day-Attack Detection // Proceedings of the 2022 IEEE International Conference on Emerging Trends in Smart Technologies (ICETST). Pakistan. Vol. 1–5. 10.1109/ICETST55735.2022.9922937.

Asokan J. et al. (2023). A Case Study Using Companies to Examine the Nmap Tool's Applicability for Network Security Assessment // Proceedings of the 2023 IEEE 12th International Conference on Advanced Computing (ICoAC). India. Vol. 1–6. 10.1109/ICoAC59537.2023.10249544.

Alwarafy A. et al. (2021). A Survey on Security and Privacy Issues in Edge-Computing-Assisted Internet of Things // IEEE Internet of Things Journal. Vol. 8(6). Pp. 4004–4022. 10.1109/JIOT.2020.3015432.

Caprolu M. et al. (2019). Edge Computing Perspectives: Architectures, Technologies, and Open Security Issues // Proceedings of the 2019 IEEE International Conference on Edge Computing (EDGE). Vol. 116–123. DOI: 10.1109/EDGE.2019.00035.

Dolui K., Datta S.K. (2017). Comparison of Edge Computing Implementations: Fog Computing, Cloudlet and Mobile Edge Computing // Proceedings of the 2017 Global Internet of Things Summit (GIoTS). Switzerland. Vol. 1–6.10.1109/GIOTS.2017.8016213.

Esposito C., Castiglione A., Pop F., Choo K.-K.R. (2017). Challenges of Connecting Edge and Cloud Computing: A Security and Forensic Perspective // IEEE Cloud Computing. Vol. 4(2). Pp. 13–17. 10.1109/MCC.2017.30.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Febro A. et al. (2022). Edge Security for SIP-Enabled IoT Devices with P4 // Computer Networks. Vol. 203 // Article 108698. 10.1016/j.comnet.2021.108698.

Fenanir S., Semchedine F. (2023). Smart Intrusion Detection in IoT Edge Computing Using Federated Learning // Revue d'Intelligence Artificielle. Vol. 37(5). Pp. 1133–1145. 10.18280/ria.370505.

Golchin P. et al. (2022). Improving DDoS Attack Detection Leveraging a Multi-Aspect Ensemble Feature Selection // Proceedings of the 2022 IEEE/IFIP Network Operations and Management Symposium (NOMS). Hungary. Vol. 1–5. 10.1109/NOMS54207.2022.9789763.

Gulatas I. et al. (2023). Malware Threat on Edge/Fog Computing Environments from Internet of Things Devices Perspective // IEEE Access. Vol. 11. Pp. 33584–33606. 10.1109/ACCESS.2023.3262614.

Goel K. et al. (2020). Reliability Analysis of Edge Scenarios Using Pedestrian Mobility // Proceedings of the 2020 50th Annual IEEE-IFIP International Conference on Dependable Systems and Networks – Supplemental Volume (DSN-S). — Spain. Vol. 61–62. 10.1109/DSN-S50200.2020.00033.

Ghosh S. et al. (2021). A High Performance Hierarchical Caching Framework for Mobile Edge Computing Environments // Proceedings of the 2021 IEEE Wireless Communications and Networking Conference (WCNC). Nanjing, China. Vol. 1–6. 10.1109/WCNC49053.2021.9417323.

Ilyin P.A. (2021). Osnovnye napravleniya primeneniya oblachnykh, granichnykh, tumannykh vychisleniy {Main Directions of Application of Cloud, Edge and Fog Computing} // StudNet. Vol. 4(6). Pp. 250–257.

Job G.K. et al. (2021). Impacts of Ransomware Attacks on Edge Computing Devices: Challenges and Research Opportunities // *International Journal of Engineering Research & Technology* (IJERT). Vol. 10(4). Pp. 665–670. 10.17577/IJERTV10IS040297.

Lin Z. et al. (2020). A Survey: Resource Allocation Technology Based on Edge Computing in IIoT // Proceedings of the 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). United Arab Emirates. Vol. 1–5. 10.1109/CCCI49893.2020.9256663.

Muin M. et al. (2022). Campus Website Security Vulnerability Analysis Using Nessus // International Journal of Computer and Information System (IJCIS). Vol. 2(3). Pp. 79–82. 10.29040/ijcis.v3i2.72.

Nam D.H. (2023). A Comparative Study of Mobile Cloud Computing, Mobile Edge Computing, and Mobile Edge Cloud Computing // Proceedings of the 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE). NV, — USA. Vol. 1219–1224. 10.1109/CSCE60160.2023.00204.

Qiang W. et al. (2021). Defending CNN Against Privacy Leakage in Edge Computing via Binary Neural Networks // Future Generation Computer Systems. Vol. 125. — Pp. 460–470. 10.1016/j.future.2021.06.037.

Qing L. et al. (2018). Research on Key Technology of Network Security Situation Awareness of Private Cloud in Enterprises // Proceedings of the 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). —China. Pp. 462–466. 10.1109/ICCCBDA.2018.8386560.

Roman R. et al. (2018). Mobile Edge Computing, Fog et al.: A Survey and Analysis of Security Threats and Challenges // Future Generation Computer Systems. Vol. 78(2). Pp. 680–698. 10.1016/j.future.2016.11.009.

Ray K. et al. (2020). Proactive Microservice Placement and Migration for Mobile Edge Computing // Proceedings of the 2020 IEEE/ACM Symposium on Edge Computing (SEC). CA, USA. Vol. 28–41. 10.1109/SEC50012.2020.00010.

Singh M.P. et al. (2023). Trusted Federated Learning Framework for Attack Detection in Edge Industrial Internet of Things // Proceedings of the 2023 IEEE Eighth International Conference on Fog and Mobile Edge Computing (FMEC). Tartu, Estonia . Pp. 64–71. 10.1109/FMEC59375.2023.10305910.

Shirazi S.N., Gouglidis A., Farshad A., Hutchison D. (2017). The Extended Cloud: Review and Analysis of Mobile Edge Computing and Fog From a Security and Resilience Perspective // *IEEE Journal on Selected Areas in Communications*. Vol. 35(11). Pp. 2586–2595. 10.1109/JSAC.2017.2760478.

Toyin S. (2023). Comparative Analysis of Security Vulnerability Scanners (Nessus and OpenVAS) in Cloud Environment. Master's Project. Deane Road, Bolton. Vol. 89. 10.13140/RG.2.2.32627.71206.

Vankayalapati R.K. et al. (2023). Unifying Edge and Cloud Computing: A Framework for Distributed AI and Real-Time Processing // *Journal for ReAttach Therapy and Developmental Diversities*. Vol. 6. No. 9s(2). Pp. 1913–1926. 10.2139/ssrn.5048827.

Xiao Y. et al. (2019). Edge Computing Security: State of the Art and Challenges // Proceedings of the IEEE. 2019. Vol. 107(8). Pp. 1608–1631. 10.1109/JPROC.2019.2918437.

Yahuza M. et al. (2020). Systematic Review on Security and Privacy Requirements in Edge Computing: State of the Art and Future Research Opportunities // IEEE Access. Vol. 8. Pp. 76541–76567. 10.1109/ACCESS.2020.2989456.

Yang X. et al. (2024). Multi-Aspect Edge Device Association Based on Time-Series Dynamic Interaction Networks // Proceedings of the IEEE INFOCOM 2024 Workshops. Canada. Pp. 1–6. 10.1109/INFOCOM-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

WKSHPS61880.2024.10620902.

Yang R. et al. (2023). Computers & Security // Computers & Security. Vol. 132 // Article 103381. 10.1016/j.cose.2023.103381.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 475–493
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.028
УДК 004.056.5

DETECTION OF CYBER ATTACKS IN TRANSPORT NETWORKS BASED ON MACHINE LEARNING METHODS

N.E. Karabayev[1], S.K. Serikbayeva[1]*, Y.M. Mardenov[2], B. Tassuov[3], M. Fajkus[4]
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Astana International University, Astana, Kazakhstan;
[3]Taraz University named after M.Kh. Dulaty, Taraz, Kazakhstan;
[4]Tomas Bata University in Zlín, Zlín, Czech Republic.
E-mail: inf_8585@mail.ru

Nurdaulet E. Karabayev — Doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0009-0008-6532-6382;
Sandugash K. Serikbayeva — PhD, Senior Lecturer, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: inf_8585@mail.ru, https://orcid.org/0000-0002-3627-3321;
Yerik M. Mardenov — Ms.Sc., Higher School of Information Technologies and Engineering, Astana International University, Astana, Kazakhstan
https://orcid.org/0000-0001-9284-9797**;**
Bolat Tassuov — Associate Professor, Department of Physics and Informatics, Taraz University named after M.Kh. Dulaty, Taraz, Kazakhstan
https://orcid.org/0000-0002-2000-6720;
Martin Fajkus — PhD, Senior Lecturer, Faculty of Applied Informatics, Tomas Bata University in Zlín, Zlín, Czech Republic
https://orcid.org/0000-0002-5698-1106.

© N. Karabayev, S. Serikbayeva, Y. Mardenov, B. Tassuov, M. Fajkus.

Abstract. With the digitalization of vehicles and the growth of the number of electronic control units, ensuring the cybersecurity of automotive networks is becoming one of the priority tasks. Modern vehicles are complex cyberphysical systems in which data exchange between electronic components is carried out via the CAN (Controller Area Network) bus. Despite the widespread adoption and reliability of this protocol, the CAN architecture did not initially provide mechanisms for protection against cyber attacks, which makes transport networks vulnerable to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

various types of intervention, including attacks such as DoS, Fuzzy, RPM Spoofing and Gear Spoofing. This paper discusses the task of automatically detecting and classifying cyberattacks in automotive networks based on machine learning methods. The open car hacking dataset was used as the initial data, containing real logs of CAN messages both under normal conditions and when simulating attacks. Preliminary data processing was performed, including cleaning, normalization and balancing of classes, as well as analysis of feature correlation. To solve the multiclass classification problem, two machine learning algorithms were implemented and compared: XGBoost and logistic regression. The quality of the models was assessed using error matrices and accuracy analysis by class. The results of the experiments showed that the XGBoost model demonstrates higher accuracy and robustness in classifying attacks compared to logistic regression, especially for most attacking classes. Additional analysis of the importance of the features made it possible to identify the most informative parameters of CAN messages, reflecting the nature of the injected attacks. The results confirm the effectiveness of the application of machine learning methods to improve the level of security of transport networks and can be used in the development of intelligent intrusion detection systems in car CAN networks.

Keywords: cybersecurity, transport networks, CAN-bus, cyberattacks, machine learning, XGBoost, logistic regression, anomaly detection, attack classification, automotive networks

For citation: N. Karabayev, S. Serikbayeva, Y. Mardenov, B. Tassuov, M. Fajkus (2026). Detection of cyber attacks in transport networks based on machine learning methods // International journal of information and communication technologies. Vol. 7. No.25. Pp. 475-493. https://doi.org/10.54309/ijict.2026.25.1.028. (In Russ.).

Conflict of interest: The authors declare that there is no conflict of interest.

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІНЕ НЕГІЗДЕЛГЕН КӨЛІК ЖЕЛІЛЕРІНДЕГІ КИБЕРШАБУЫЛДАРДЫ АНЫҚТАУ

Н.Е. Қарабаев[1], С.К. Серікбаева[1]*, Е.М. Марденов[2], Б. Тасуов[3], М. Файкус[4]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;

[2]Астана халықаралық университеті, Астана, Қазақстан;

[3]М. Х. Дулати атындағы Тараз университеті , Тараз, Қазақстан;

[4]Злиндегі Томас Бата университеті, Злин , Чех Республикасы.

E-mail: inf_8585@mail.ru

Қарабаев Нұрдәулет Ерланұлы — докторант, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0009-0008-6532-6382;

Серикбаева Сандугаш Курманбековна — PhD, ақпараттық жүйелер

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
E-mail: inf_8585@mail.ru, https://orcid.org/0000-0002-3627-3321;
Марденов Ерік Маратұлы — магистр, Ақпараттық технологиялар және инженерия жоғары мектебі, Астана халықаралық университеті, Астана, Қазақстан https://orcid.org/0000-0001-9284-9797;
Тасуов Болат — физика және информатика кафедрасының қауысдастырылған профессоры, М. Х. Дулати атындағы Тараз университеті, Тараз, Қазақстан https://orcid.org/0000-0002-2000-6720;
Файкус Мартин — PhD, қолданбалы информатика факультетінің аға оқытушысы, Злиндегі Томас Бата университеті, Злин, Чех Республикасы https://orcid.org/0000-0002-5698-1106.

© Н.Е. Қарабаев, С.К. Серикбаева, Е.М. Марденов, Б. Тасуов, М. Файкус.

Аннотация. Көлік құралдарын цифрландыру және басқарудың электрондық блоктары санының өсуі жағдайында автомобиль желілерінің киберқауіпсіздігін қамтамасыз ету басым міндеттердің біріне айналуда. Қазіргі заманғы көлік құралдары электрондық компоненттер арасында деректер алмасу CAN (Controller Area Network) шинасы бойынша жүзеге асырылатын күрделі киберфизикалық жүйелер болып табылады. Осы хаттаманың кең таралуына және сенімділігіне қарамастан, CAN архитектурасы бастапқыда кибершабуылдардан қорғау тетіктерін көздемеген, бұл көлік желілерін DoS, Fuzzy, RPM Spoofing және Gear Spoofing сияқты шабуылдарды қоса алғанда, араласудың әртүрлі түрлеріне осал етеді. Бұл жұмыста машиналық оқыту әдістері негізінде автомобиль желілерінде киберқақтарды автоматты түрде анықтау және жіктеу міндеті қарастырылады. Бастапқы деректер ретінде қалыпты жағдайларда да, шабуылдарды модельдеу кезінде де CAN-хабарламалардың нақты журналдарын қамтитын Car Hacking Dataset ашық жинағы пайдаланылды. Сыныптарды тазартуды, қалыпқа келтіруді және теңгерімдеуді, сондай-ақ белгілердің корреляциясын талдауды қамтитын деректерді алдын ала өңдеу жүргізілді. Мультиклассалық жіктеу міндетін шешу үшін машиналық оқытудың екі алгоритмі іске асырылды және салыстырылды: XGBoost және логистикалық регрессия. Модельдер сапасын бағалау қателер матрицаларын және сыныптар бойынша дәлдікті талдауды пайдалана отырып жүргізілді. Эксперименттердің нәтижелері көрсеткендей, XGBoost моделі логистикалық регрессиямен салыстырғанда, әсіресе көптеген шабуылдаушы кластар үшін шабуылдарды жіктеу кезінде жоғары дәлдік пен тұрақтылықты көрсетеді. Белгілердің маңыздылығын қосымша талдау инжектирленетін шабуылдардың сипатын көрсететін CAN-хабарламалардың неғұрлым ақпараттық параметрлерін анықтауға мүмкіндік берді. Алынған нәтижелер көлік желілерінің қауіпсіздік деңгейін арттыру үшін машиналық оқыту әдістерін қолдану тиімділігін растайды және CAN-желілерінде басып

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

кіруді анықтаудың зияткерлік жүйелерін әзірлеу кезінде пайдаланылуы мүмкін.

Түйін сөздер: киберқауіпсіздік, көлік желілері, CAN-шина, кибер шабуылдар, машиналық оқыту, XGBoost, логистикалық регрессия, ауытқуларды анықтау, шабуылдарды жіктеу, автомобиль желілері

Дәйексөздер үшін: Н.Е. Қарабаев, С.К. Серикбаева, Е.М. Марденов, Б. Тасуов, М. Файкус (2026). Машиналық оқыту әдістеріне негізделген көлік желілеріндегі кибершабуылдарды анықтау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т 7. № 25. Б. 475-493. https://doi.org/10.54309/IJICT.2026.25.1.028. (Орыс тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ОБНАРУЖЕНИЕ КИБЕРАТАК В ТРАНСПОРТНЫХ СЕТЯХ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Н.Е. Карабаев[1], С.К. Серикбаева[1*], Е.М. Марденов[2], Б. Тасуов[3], М. Файкус[4]

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;
[2]Международный университет Астана, Астана, Казахстан;
[3]Таразский университет им. М.Х. Дулати, Тараз, Казахстан;
[4]Университет Томаса Баты в Злине, Злин, Чешская Республика.
E-mail: inf_8585@mail.ru

Карабаев Нурдаулет Ерланович — докторант, Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0009-0008-6532-6382;
Серикбаева Сандугаш Курманбековна — PhD, старший преподаватель кафедры информационных систем, Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан
E-mail: inf_8585@mail.ru, https://orcid.org/0000-0002-3627-3321;
Марденов Ерик Маратович — магистр, Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан
https://orcid.org/0000-0001-9284-9797;
Тасуов Болат — ассоциированный профессор, кафедра физики и информатики, Таразский университет имени М.Х. Дулати, Тараз, Казахстан
https://orcid.org/0000-0002-2000-6720;
Файкус Мартин — PhD, старший преподаватель, факультет прикладной информатики, Университет Томаша Баты в Злине, Злин, Чешская Республикаhttps://orcid.org/0000-0002-5698-1106.

© Н.Е. Карабаев, С.К. Серикбаева, Е.М. Марденов, Б. Тасуов, М. Файкус.

Аннотация. В условиях цифровизации транспортных средств и роста числа электронных блоков управления обеспечение кибербезопасности авто-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

мобильных сетей становится одной из приоритетных задач. Современные транспортные средства представляют собой сложные киберфизические системы, в которых обмен данными между электронными компонентами осуществляется по шине CAN (Controller Area Network). Несмотря на широкое распространение и надежность данного протокола, архитектура CAN изначально не предусматривала механизмов защиты от кибератак, что делает транспортные сети уязвимыми к различным видам вмешательства, включая атаки типа DoS, Fuzzy, RPM Spoofing и Gear Spoofing. В данной работе рассматривается задача автоматического обнаружения и классификации кибератак в автомобильных сетях на основе методов машинного обучения. В качестве исходных данных использовался открытый набор Car Hacking Dataset, содержащий реальные журналы CAN-сообщений как в нормальных условиях, так и при моделировании атак. Проведена предварительная обработка данных, включающая очистку, нормализацию и балансировку классов, а также анализ корреляции признаков. Для решения задачи мультиклассовой классификации были реализованы и сравнены два алгоритма машинного обучения: XGBoost и логистическая регрессия. Оценка качества моделей проводилась с использованием матриц ошибок и анализа точности по классам. Результаты экспериментов показали, что модель XGBoost демонстрирует более высокую точность и устойчивость при классификации атак по сравнению с логистической регрессией, особенно для большинства атакующих классов. Дополнительный анализ важности признаков позволил выявить наиболее информативные параметры CAN-сообщений, отражающие характер инжектируемых атак. Полученные результаты подтверждают эффективность применения методов машинного обучения для повышения уровня безопасности транспортных сетей и могут быть использованы при разработке интеллектуальных систем обнаружения вторжений в автомобильных CAN-сетях.

Ключевые слова: кибербезопасность, транспортные сети, CAN-шина, кибератаки, машинное обучение, XGBoost, логистическая регрессия, обнаружение аномалий, классификация атак, автомобильные сети

Для цитирования: Н.Е. Карабаев, С.К. Серикбаева, Е.М. Марденов, Б. Тасуов, М. Файкус (2026). Обнаружение кибератак в транспортных сетях на основе методов машинного обучения // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 25. Стр. 475–493. https://doi.org/10.54309/IJICT.2026.25.2.028.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Финансирование. *Настоящая работа сотрудниками НАО «Евразийский национальный университет имени Л.Н. Гумилева» проводится при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (Грант № AP23489127).*

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Введение.

Современные транспортные средства представляют собой сложные киберфизические системы с множеством электронных блоков управления (ECU). ECU отвечают за взаимодействие различных подсистем автомобиля, таких как двигатель, трансмиссия, тормозная система, система стабилизации и мультимедийные и вспомогательные услуги. Шина контроллерной сети (CAN) используется для эффективного обмена информацией между этими частями. Благодаря своей универсальности, надежности и низкой стоимости внедрения она на сегодняшний день стала стандартом в индустрии. Тем не менее, при разработке архитектуры CAN основное внимание уделяли производительности и устойчивости передачи данных в условиях ограниченных вычислительных ресурсов, а проблемы информационной безопасности оставались второстепенными. Следовательно, CAN-сети обладают значительными уязвимостями, которые активно используются в современных сценариях кибератак несмотря на то, что они широко распространены и проверены временем. С ростом количества электронных систем в транспортных средствах и возможностью удаленного взаимодействия с автомобилем посредством беспроводных интерфейсов исследования безопасности CAN-сетей становятся все более важными. Это связано с увеличением поверхности атак и рисков для пользователей.

Кибератаки на автомобильные сети могут сильно повлиять на функционирование автомобилей. Атаки типа отказа в обслуживании (DoS) (Cil, 2021), которые блокируют передачу данных между электронными блоками за счет массовой инъекции сообщений, являются одними из наиболее распространенных угроз. Другие типы угроз включают атаки тумана, которые дестабилизируют работу сети, передавая случайные идентификаторы и данные; и атаки на подмену, которые подменяют важные параметры, такие как обороты двигателя (RPM) или положение передачи (Gear) (Xiao et all., 2019).

Такие атаки могут привести к ложному отображению информации на приборной панели или полной потере управляемости автомобилем, поэтому особенно важно своевременно обнаруживать и классифицировать их. Из-за ограниченных вычислительных ресурсов ECU, требований к минимальной задержке передачи данных и необходимости поддерживать совместимость с существующими протоколами традиционные методы защиты, такие как криптографические механизмы или системы контроля доступа, неэффективны в условиях автомобильных сетей. В результате исследователи все больше обращают внимание на методы анализа данных и машинного обучения, которые позволяют выявлять закономерности в поведении сети и автоматически обнаруживать аномалии, связанные с кибератаками.

Аномалии, криптографические методы и машинное обучение — это некоторые из методов защиты от атак на автомобильные сети. В этой работе мы рассматриваем задачу мультиклассовой классификации и используем машинное обучение для классификации атак. Мы оценили эффективность моделей XGBoost

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

и логистической регрессии на основе реальных данных из набора данных для хакинга автомобилей.

Цель данной работы – разработка и тестирование моделей машинного обучения для автоматического обнаружения атак в автомобильной сети CAN. Мы сравним два алгоритма – XGBoost и логистическую регрессию – и оценим их способность классифицировать нормальное и атакующее поведение в сети автомобиля.

Обзор литературы

В работе (Chevalier et all., 2021) рассматривается проблема обнаружения кибератак в современных транспортных средствах. Авторы отмечают, что развитие интеллектуальных транспортных систем и широкое использование бортовых сетей передачи данных (в частности, CAN-шины) повышает уязвимость автомобилей к кибератакам. Традиционные методы защиты, основанные на сигнатурном анализе, не всегда способны выявлять новые или модифицированные типы атак, что требует внедрения более гибких и адаптивных подходов. В исследовании предлагаются два метода обнаружения аномалий в автомобильной сети. Первый основан на использовании характеризующих функций (Characteristic Functions), позволяющих выделять статистические и структурные особенности потока сообщений и выявлять отклонения от нормального поведения системы. Второй метод реализован с применением искусственных нейронных сетей (ANN), обучаемых на данных нормального и атакующего трафика. Дополнительно применяется визуальный анализ для интерпретации результатов и оценки характера выявленных аномалий. Особое внимание уделяется сравнению методов по показателям точности обнаружения и вычислительной эффективности. Экспериментальные результаты показывают, что подход на основе характеризующих функций демонстрирует сопоставимую с нейронными сетями точность, при этом значительно превосходит их по скорости обработки данных и требованиям к вычислительным ресурсам. Это делает его более подходящим для внедрения во встроенные автомобильные системы с ограниченными аппаратными возможностями. Работа подтверждает перспективность использования интеллектуальных методов анализа данных для обеспечения кибербезопасности транспортных средств и предлагает практико-ориентированное решение, пригодное для применения в реальных автомобильных инфраструктурах.

В работе (Sharma et all., 2024) авторы исследуют возможности использования методов контролируемого машинного обучения для выявления киберугроз в реальном времени на основе данных из датасета CTU-13, содержащего сетевой трафик с ботнет-атаками. Работа направлена на повышение точности и скорости обнаружения вредоносной активности в сетях. Основное внимание уделено сравнительному анализу таких алгоритмов, как Random Forest, SVM и Gradient Boosting. Оценка эффективности проводится по метрикам точности (accuracy), полноты (recall), F1-меры и времени обработки. Авторы приходят к выводу, что модели Random Forest и Gradient Boosting демонстрируют наилучший

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

баланс между точностью классификации и производительностью при обработке сетевого трафика. Кроме того, в статье подчеркивается значимость использования реальных наборов данных (как CTU-13) для построения надежных систем обнаружения угроз. Работа представляет интерес для исследователей и практиков в области информационной безопасности, поскольку предлагает обоснованные подходы к применению машинного обучения для задач киберзащиты в условиях ограниченного времени и ресурсов.

В работе (Jabia Nzi et all., 2022) представлено сравнение алгоритмов обнаружения сетевых атак типа DDoS-атак (отказ в обслуживании) для различных сервисов хранения, обработки и передачи данных через Интернет. Особое внимание уделяется применению алгоритмов машинного обучения, таких как гауссовская смешанная модель для максимизации ожиданий (GMM-EM), линейная регрессия (LR), SVM (машина опорных векторов) (с линейным, RBF (радиальная базисная функция) или полиномиальные ядра), алгоритмы дерева решений (Decision Tree), наивный Байеса (Naive Bayes) и рандом Forest (Rand om Forest) для обнаружения такого типа атак. В конце статьи оцениваются перечисленные выше алгоритмы машинного обучения и тщательно сравнивается их производительность. Все экспериментальные результаты показывают, что более 99,7 % двух видов DOS-атак успешно обнаруживаются. Этот подход не снижает производительность и может быть легко распространен на более широкие DOS-атаки.

Исследование (Barthwal et all., 2023) посвящено разработке объяснимой глубокой модели обнаружения вторжений (XAI-based Deep Learning IDS) для интеллектуальных транспортных сетей, основанных на Интернете вещей (IoT). Авторы подчеркивают растущую уязвимость таких систем из-за их связности и сложности, что требует интеграции кибербезопасности на уровне сети. В работе предложен гибридный подход, сочетающий методы глубокого обучения с механизмами объяснимости, что позволяет не только выявлять аномалии с высокой точностью, но и интерпретировать решения модели. Используются современные архитектуры нейронных сетей, включая CNN и LSTM, а также объяснительные методы, такие как LIME и SHAP, для анализа влияния признаков. Эксперименты на наборах данных IoT показали повышение точности обнаружения атак и улучшение прозрачности модели. Исследование делает вклад в развитие надежных, интерпретируемых систем безопасности для «умного» транспорта будущего.

В работе (Wagh et all., 2024) рассматривается применение алгоритмов машинного обучения для обнаружения кибератак и сетевых вторжений в современных коммуникационных системах. Авторы отмечают, что с ростом объема сетевых данных и увеличением сложности атак традиционные методы защиты становятся недостаточно эффективными. В исследовании предлагается модель, основанная на машинном обучении, использующая алгоритмы классификации, такие как Random Forest, Support Vector Machine (SVM) и

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Decision Tree, для идентификации аномального поведения в сетевом трафике. Особое внимание уделяется сравнению эффективности различных алгоритмов по показателям точности, полноты и времени обучения. Результаты экспериментов показывают, что методы на основе Random Forest демонстрируют наилучшие результаты при обнаружении известных и неизвестных атак. Работа подчеркивает значимость машинного обучения в повышении безопасности сетей и формирует основу для дальнейшего внедрения интеллектуальных систем защиты в реальных инфраструктурах.

Работа (SaiKiran et all., 2025) посвящена разработке интеллектуального подхода к обнаружению кибератак в компьютерных сетях с использованием технологий машинного обучения. Авторы подчеркивают необходимость автоматизированных и адаптивных систем защиты, способных эффективно противостоять быстро эволюционирующим угрозам. В работе представлена модель, использующая комбинацию алгоритмов классификации, включая K-Nearest Neighbors (KNN), Random Forest и Logistic Regression, для анализа сетевого трафика и выявления аномалий. Особое внимание уделено предварительной обработке данных и выбору релевантных признаков, что позволяет повысить точность классификации атак. Результаты экспериментальных испытаний показали, что предложенный подход обеспечивает высокую точность и надежность при низком уровне ложных срабатываний. Исследование демонстрирует потенциал машинного обучения для создания интеллектуальных, устойчивых к новым угрозам систем кибербезопасности и подчеркивает важность их интеграции в современные сетевые инфраструктуры.

Исследование (Maltseva et all., 2024) посвящено разработке алгоритмов раннего обнаружения кибератак на сети с использованием методов машинного обучения. Авторы отмечают, что традиционные системы обнаружения вторжений часто реагируют с опозданием, что приводит к значительным рискам для информационной безопасности. В исследовании предлагается методология, направленная на выявление признаков атак на ранних стадиях, до того как они могут нанести ущерб инфраструктуре. Для этого используются алгоритмы машинного обучения, включая нейронные сети, деревья решений и метод опорных векторов (SVM), а также методы отбора признаков и оптимизации параметров моделей. Проведённые эксперименты показали, что интеграция нескольких алгоритмов повышает точность прогнозирования и снижает количество ложных тревог (Roman et all., 2018). Работа делает вклад в развитие интеллектуальных систем раннего предупреждения, способных адаптироваться к новым типам атак и обеспечивать более высокий уровень киберзащиты сетевых систем.

В работе (Rahman et all.,.2025) рассматривается применение алгоритмов машинного обучения для мониторинга и обнаружения кибератак в динамически изменяющихся сетевых средах. Авторы подчеркивают, что рост числа киберугроз требует автоматизированных, масштабируемых систем, способных анализировать большие объемы данных в реальном времени. В работе представлены и

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

сравнительно оценены различные алгоритмы машинного обучения — включая Decision Tree, Random Forest, Naïve Bayes и Support Vector Machine (SVM) — с точки зрения их точности, скорости и устойчивости к шумным данным. Результаты экспериментов показывают, что ансамблевые методы, особенно Random Forest, обеспечивают наилучший баланс между точностью и производительностью. Исследование акцентирует внимание на интеграции ML-подходов в системы сетевого мониторинга и их способности выявлять как известные, так и новые типы атак, что способствует повышению общей эффективности и адаптивности средств киберзащиты.

Анализ рассмотренных исследований показывает, что современные подходы к обнаружению кибератак в сетях, включая транспортные системы, активно развиваются в направлении использования алгоритмов машинного обучения и глубокого обучения. Большинство обзорных работ показывают высокую эффективность методов Random Forest, SVM, Decision Tree и Gradient Boosting, которые обеспечивают высокую точность классификации и позволяют выявлять аномалии в реальном времени. Отдельное внимание уделяется объяснимости моделей и раннему обнаружению угроз, что особенно важно для критических инфраструктур, таких как автомобильные сети CAN и IoT-среды. Исследователи сходятся во мнении, что интеграция интеллектуальных алгоритмов в системы мониторинга значительно повышает уровень защиты и снижает риск успешных атак. При этом остаются актуальными задачи повышения устойчивости моделей к новым типам угроз, оптимизации времени обработки и адаптации решений к ресурсно-ограниченным системам, что определяет направления дальнейших исследований в области кибербезопасности транспортных сетей.

Материалы и методы.

В качестве исходных данных для исследования использовался открытый набор Car Hacking Dataset, содержащий журналы сообщений автомобильной сети CAN (Controller Area Network), записанные как в нормальных условиях функционирования транспортного средства, так и в ситуациях, моделирующих кибератаки различных типов — DoS, Fuzzy, RPM Spoofing и Gear Spoofing. Данный набор данных был выбран благодаря своей реалистичности и репрезентативности, поскольку отражает реальные сценарии функционирования электронных блоков управления (ECU), взаимодействующих между собой посредством шины CAN. Каждый экземпляр данных включает временную метку (Timestamp), идентификатор сообщения (CAN_ID) в шестнадцатеричном формате, длину данных (DLC), байты данных (DATA[0–7]) и флаг состояния (Flag), указывающий, является ли сообщение нормальным (R) или атакующим (T). Анализ данных выявил значительный дисбаланс классов: нормальные сообщения составляют около 85,93 %, атакующие — 14,07 %. Это требует применения специальных методов балансировки данных, таких как oversampling (дублирование меньшего класса) или undersampling (сокращение большего класса). Применение данных методов позволяет избежать смещения модели в сторону преобладающего класса и

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

обеспечить более справедливое обучение. Таким образом, исходный датасет был предварительно очищен от дубликатов, нормализован и преобразован в формат, пригодный для машинного обучения. Такая подготовка данных является ключевым этапом, обеспечивающим адекватную работу моделей классификации атак и снижение вероятности переобучения.

Для оценки взаимосвязей между признаками, извлечёнными из сообщений CAN, была построена матрица корреляции Пирсона (рис. 1). Анализ показал, что между большинством признаков наблюдается слабая или умеренная корреляция, что свидетельствует о низкой избыточности данных и обоснованности их включения в модель. Наиболее выраженные взаимосвязи зафиксированы между признаками DATA[1], DATA[2], DATA[3] и DATA[4], где коэффициенты корреляции достигают значений 0.3–0.4, что указывает на частичную зависимость передаваемых байтов внутри одного пакета. Отрицательная корреляция между CAN_ID и DATA[7] (около –0.21) демонстрирует, что данные байты характеризуют различные аспекты сетевой активности.

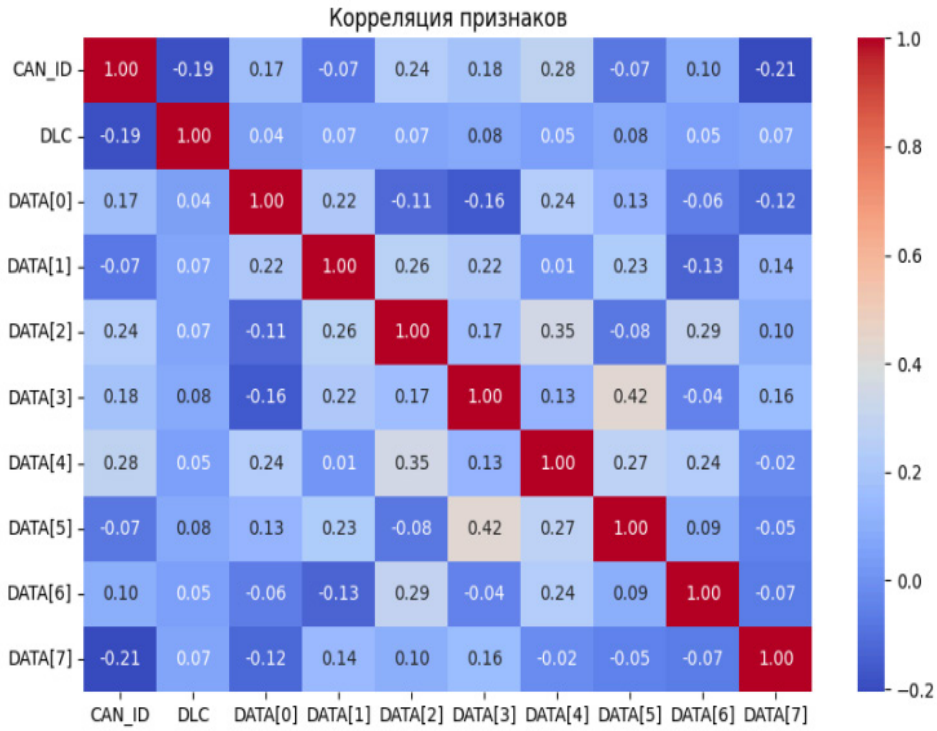


Рис. 1. Матрица корреляции признаков для данных автомобильной сети CAN

Такое распределение корреляций подтверждает целесообразность использования нелинейных методов обучения, таких как XGBoost, способных выявлять сложные взаимодействия между признаками, не ограничиваясь линейными зависимостями. При этом низкая взаимная корреляция между большинством признаков минимизирует риск мультиколлинеарности, что

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

положительно сказывается на стабильности и обобщающей способности модели. Таким образом, анализ корреляции подтверждает корректность предварительного отбора признаков и обеспечивает дополнительное обоснование применённого подхода к построению модели обнаружения кибератак.

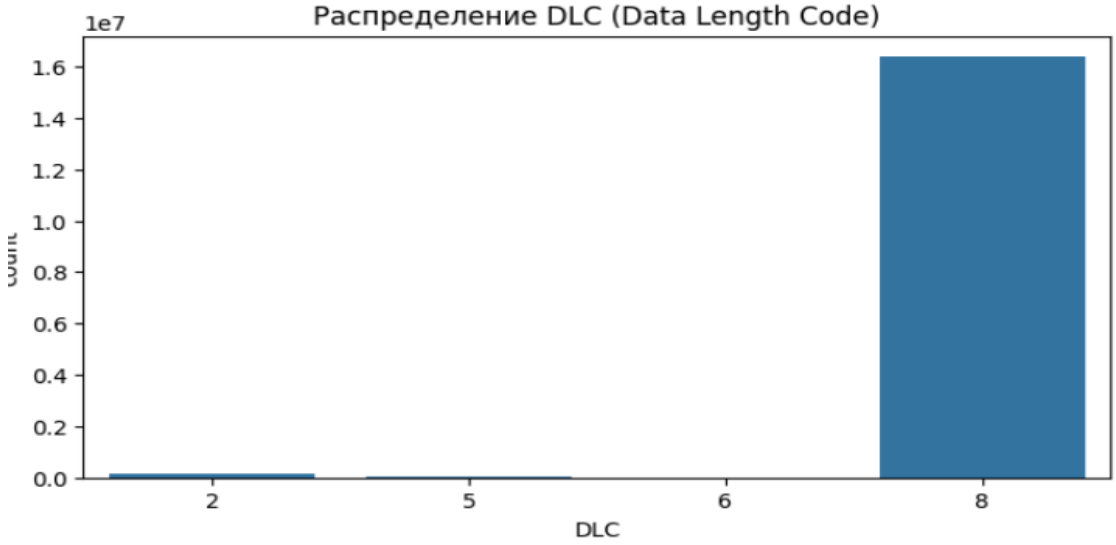


Рис. 2. Распределение длины пакета данных (DLC) в CAN-сообщениях

На рисунке 2 представлено распределение длины пакета данных (Data Length Code, DLC) для всех сообщений автомобильной сети CAN. Анализ показал, что подавляющее большинство пакетов имеют длину 8 байт, что соответствует максимально возможному объёму данных в одном сообщении CAN и отражает типичную структуру обмена информацией между электронными блоками управления (ECU). Такая закономерность объясняется особенностями формирования сообщений в автомобильных сетях, где наиболее информативные данные (например, обороты двигателя, положение педали акселератора или параметры трансмиссии) передаются в полном формате 8-байтных кадров.

Сообщения с меньшей длиной (1–7 байт) встречаются крайне редко и, как правило, связаны с диагностическими или служебными сигналами. Это распределение подтверждает однородность структуры трафика и объясняет низкую вариативность признака DLC в дальнейших моделях машинного обучения. В контексте задачи обнаружения атак данная особенность указывает на то, что длина сообщения не является значимым предиктором для классификации аномалий, однако может использоваться в сочетании с другими признаками (например, CAN_ID или значениями DATA[0–7]) для комплексного анализа сетевого поведения. Таким образом, анализ DLC подтверждает стабильность сетевой структуры и отсутствие явных аномалий в параметре длины пакета.

Перед обучением модели данные проходят этап предобработки, который

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

включает несколько обязательных шагов. В первую очередь выполняется очистка данных: удаляются дубликаты, а также обрабатываются пропущенные значения, чтобы избежать искажения результатов обучения. Далее проводится преобразование признаков — идентификаторы CAN ID переводятся в числовой формат, а числовые характеристики нормализуются для обеспечения корректной работы алгоритмов машинного обучения. После этого набор данных разделяется на обучающую и тестовую выборки в стандартном соотношении 80/20. Дополнительно применяется балансировка классов с использованием методов увеличения или уменьшения выборки, что позволяет повысить качество классификации и снизить влияние дисбаланса данных.

Для решения задачи классификации атак используются два алгоритма машинного обучения. В качестве основного метода рассматривается XGBoost — эффективный бустинговый алгоритм, хорошо справляющийся с задачами классификации и устойчивый к дисбалансу классов. В качестве базовой модели применяется логистическая регрессия, отличающаяся простотой и стабильной работой на линейно разделимых данных. Обучение моделей проводится на предварительно сбалансированных данных с применением кросс-валидации, что позволяет повысить надежность и обобщающую способность полученных результатов.

Результаты и обсуждение.

Матрица ошибок (Confusion Matrix) показала, что модель XGBoost более точно классифицирует атакующие сообщения по сравнению с логистической регрессией. Последняя в ряде случаев ошибочно относила атакующие сообщения к нормальному поведению системы. Для оценки качества классификации модели XGBoost была построена матрица ошибок, позволяющая проанализировать распределение правильно и ошибочно классифицированных объектов по каждому классу (рис.3).

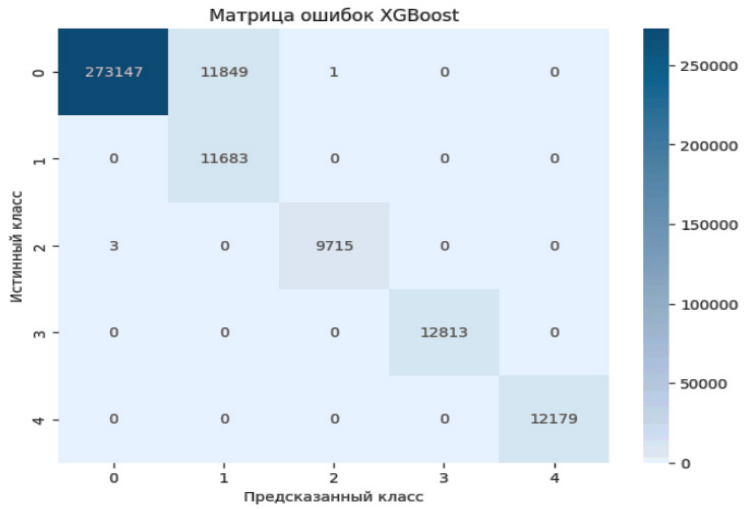


Рис.3. Матрица ошибок модели XGBoost

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Для оценки точности классификации и выявления характера ошибок при использовании модели логистической регрессии была построена матрица ошибок, позволяющая определить распределение верно и ошибочно отнесённых объектов по классам (Рис. 4).

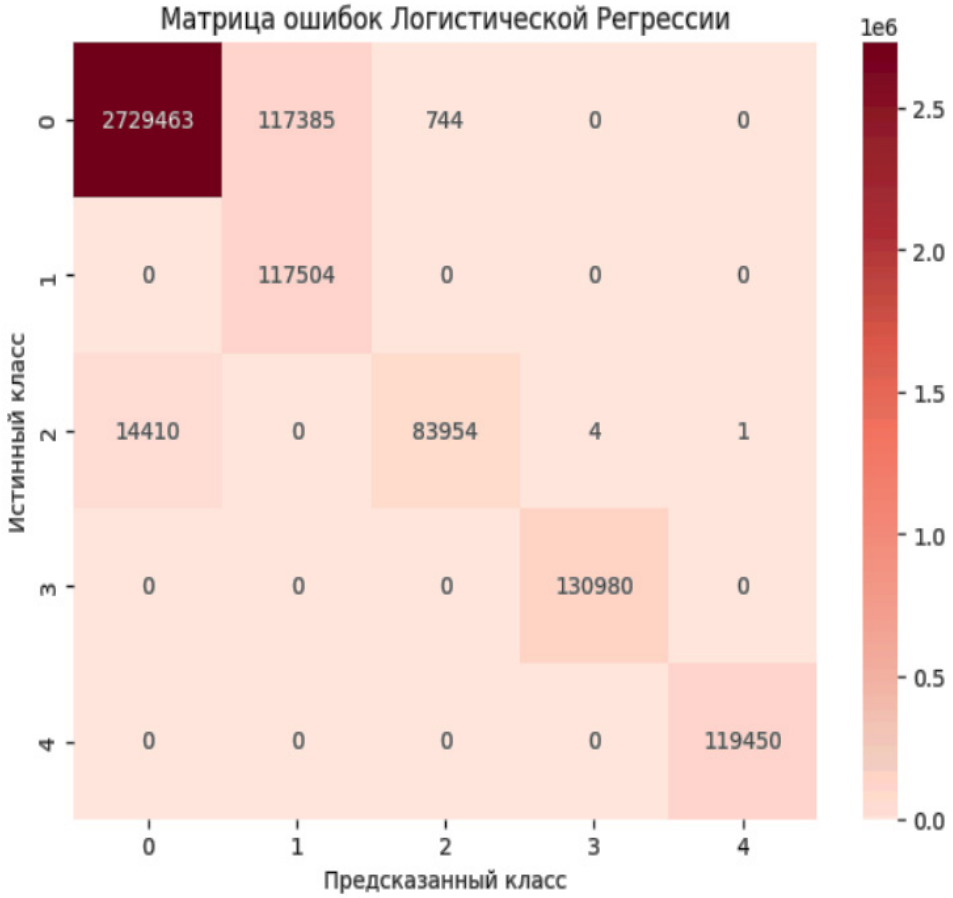


Рис. 4. Матрица ошибок модели логистической регрессии

График важности признаков, полученный на основе модели XGBoost, показал, что наибольший вклад в процесс классификации вносят параметры DATA[7] и DATA[5]**.** Высокая значимость указанных признаков свидетельствует о том, что именно они в наибольшей степени отражают характерные особенности инжектируемых атак и оказывают существенное влияние на принятие моделью классификационного решения.

В случае использования логистической регрессии наиболее значимыми оказались признаки DATA[6] и DATA[5]**.** Наличие общего значимого параметра DATA[5] для обеих моделей указывает на его устойчивую информативность и позволяет рассматривать данный признак как один из ключевых факторов, описывающих закономерности инжектируемых атак. Отличия в ранжировании остальных

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

признаков объясняются различиями в принципах работы алгоритмов: XGBoost выявляет сложные нелинейные зависимости и взаимодействия между признаками, тогда как логистическая регрессия ориентирована преимущественно на линейные взаимосвязи.

Полученные результаты подтверждают, что анализ важности признаков позволяет выявить скрытые зависимости в структуре данных и повысить интерпретируемость моделей машинного обучения при решении задач обнаружения инжектируемых атак. Для интерпретации результатов классификации и выявления наиболее информативных признаков был выполнен анализ их важности для моделей XGBoost и логистической регрессии (рис.5). Оценка вклада отдельных признаков позволяет определить, какие параметры в наибольшей степени влияют на процесс принятия классификационного решения и отражают характерные особенности инжектируемых атак.

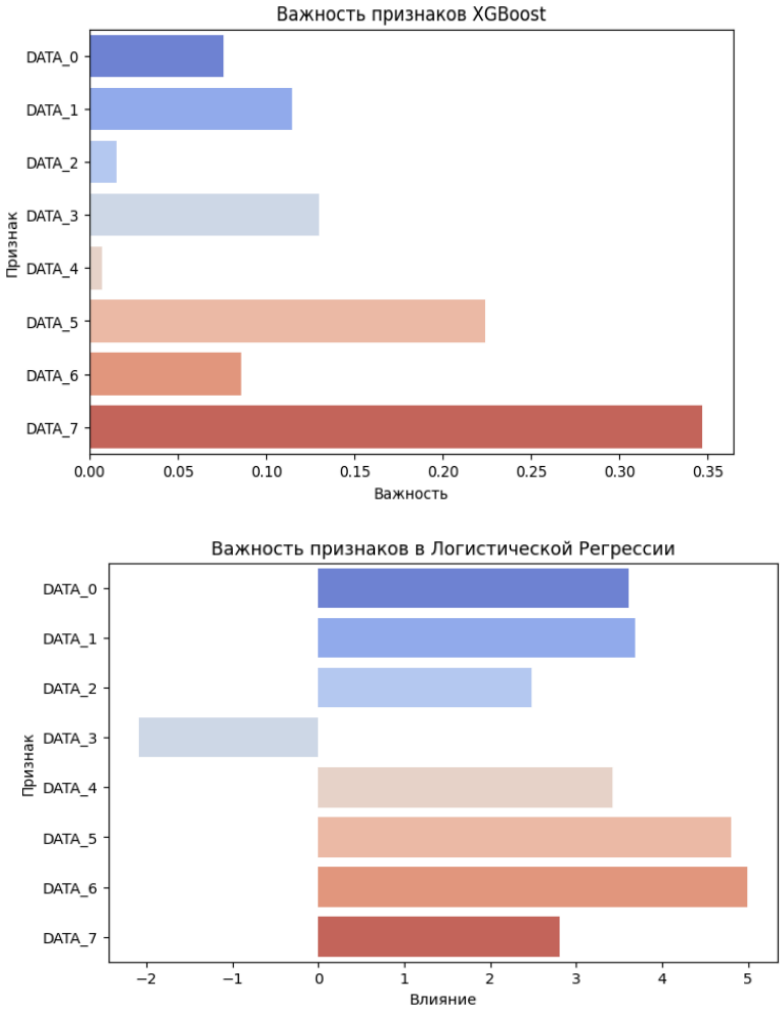


Рис. 5. Важность признаков в модели XGBoost

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Проведённый анализ результатов классификации показал, что используемые модели демонстрируют высокую точность при распознавании объектов, относящихся к классам 0, 2, 3 и 4. Для указанных классов наблюдается преобладание корректных предсказаний, что свидетельствует о достаточной информативности используемых признаков и устойчивости моделей в данных категориях. В то же время классификация объектов класса 1 сопровождается снижением точности. Это обусловлено тем, что характеристики данного класса в значительной степени пересекаются с признаковым пространством других классов, прежде всего класса **0**. В результате объекты класса 1 часто ошибочно относятся к классу 0, что приводит к увеличению числа ошибок и снижению показателей точности и полноты для данного класса. Полученные результаты указывают на необходимость дополнительной дифференциации класса 1, в том числе за счёт расширения набора признаков, применения методов балансировки классов либо использования более сложных моделей, способных лучше разделять близкие по характеристикам классы.

Для более детальной оценки качества работы моделей была проанализирована точность предсказаний по каждому классу отдельно. Такой анализ позволяет выявить классы, для которых модели демонстрируют устойчивые результаты, а также определить категории, вызывающие наибольшие затруднения при классификации (рис.6).

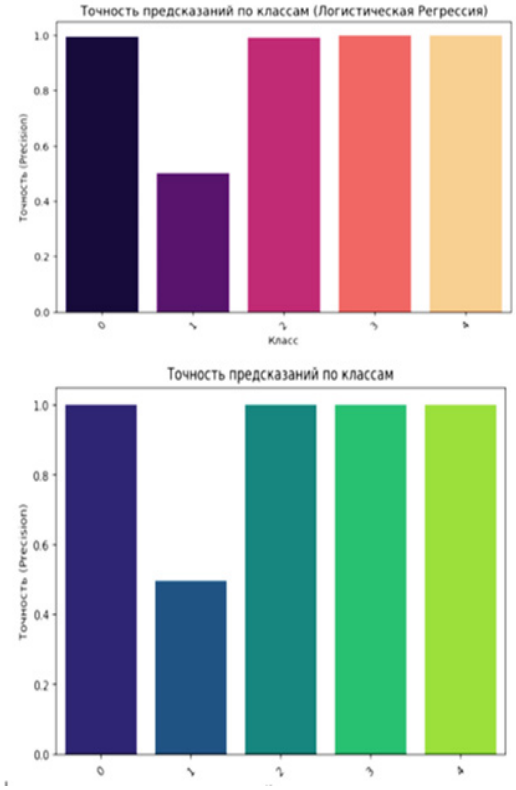


Рис. 6. Точность предсказаний по классам для модели логистической регрессии

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Полученные экспериментальные результаты свидетельствуют о высокой эффективности применения методов машинного обучения для обнаружения кибератак в автомобильных сетях CAN. Сравнительный анализ показал, что модель XGBoost превосходит логистическую регрессию по точности классификации и способности выявлять атакующее поведение, что объясняется ее способностью учитывать нелинейные зависимости и сложные взаимодействия между признаками.

Анализ матриц ошибок продемонстрировал, что обе модели уверенно классифицируют большинство классов атак, однако наибольшие трудности возникают при распознавании класса 1. Это связано с пересечением его характеристик с признаковым пространством нормального трафика, что приводит к ошибочной классификации части атак как легитимных сообщений. Данный факт указывает на необходимость дальнейшего расширения признакового пространства или применения более сложных ансамблевых и гибридных моделей.

Дополнительный анализ важности признаков показал, что параметры DATA[5], DATA[6] и DATA[7] играют ключевую роль в выявлении инжектируемых атак. Совпадение наиболее значимых признаков для разных моделей подтверждает их устойчивую информативность и практическую значимость для задач мониторинга CAN-трафика. В целом результаты согласуются с выводами современных исследований и подтверждают перспективность использования XGBoost для задач кибербезопасности транспортных систем.

Заключение.

В ходе проведённого исследования была решена актуальная задача обнаружения и классификации кибератак в автомобильных сетях CAN на основе методов машинного обучения. Рост цифровизации транспортных средств и увеличение числа электронных блоков управления существенно повышают требования к обеспечению кибербезопасности бортовых сетей, что делает разработку интеллектуальных систем мониторинга особенно востребованной. В работе использован открытый набор данных Car Hacking Dataset, содержащий реальные журналы CAN-сообщений как в штатном режиме, так и при моделировании атак типов DoS, Fuzzy, RPM Spoofing и Gear Spoofing. Проведена комплексная предобработка данных, включающая очистку, нормализацию, анализ корреляции признаков и балансировку классов, что позволило повысить устойчивость моделей и снизить влияние дисбаланса выборки. Для решения задачи мультиклассовой классификации были реализованы и сравнены два алгоритма — логистическая регрессия и XGBoost. Результаты экспериментов показали, что модель XGBoost обеспечивает более высокую точность и устойчивость при распознавании атакующих классов по сравнению с линейной моделью, что обусловлено её способностью учитывать сложные нелинейные зависимости между признаками CAN-сообщений. Анализ матриц ошибок позволил выявить классы, вызывающие наибольшие трудности при классификации, что указывает на необходимость дальнейшего расширения признакового пространства и

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

совершенствования методов разделения близких по характеристикам классов. Дополнительный анализ важности признаков показал, что наибольший вклад в процесс классификации вносят отдельные байты данных CAN-сообщений, что подтверждает их информативность при выявлении инжектируемых атак и повышает интерпретируемость построенных моделей. Практическая значимость работы заключается в возможности интеграции предложенного подхода в интеллектуальные системы обнаружения вторжений, функционирующие в условиях ограниченных вычислительных ресурсов автомобильных платформ. Полученные результаты подтверждают эффективность применения методов машинного обучения для повышения уровня кибербезопасности транспортных сетей и создают основу для дальнейших исследований, направленных на адаптацию моделей к работе в режиме реального времени, расширение спектра анализируемых атак и разработку более устойчивых и масштабируемых решений для защиты современных транспортных средств. Перспективным направлением дальнейших исследований является применение более сложных архитектур машинного обучения, включая ансамблевые методы и модели глубокого обучения (например, нейронные сети CNN и LSTM), что позволит повысить точность классификации трудноразделимых классов. Кроме того, целесообразно провести оценку вычислительной сложности и времени обработки предложенного подхода, что позволит определить его применимость в системах реального времени и в условиях ограниченных вычислительных ресурсов автомобильных платформ.

REFERENCES

Barthwal A., & Raheja S. (2023). An explainable deep learning intrusion detection in IoT-enabled transportation networks // Proceedings of the International Conference on Artificial Intelligence and Computing Communication Technologies. Pp. 310–316. 10.1109/ICAICCIT60255.2023.10466149.

Cil A. E., Yildiz K., & Buldu A. (2021). Detection of DDoS attacks with feed forward based deep neural network model // Expert Systems with Applications. P.169. Article 114520. 10.1016/j.eswa.2020.114520.

Chevalier Y., Fenzl F., Kolomeets M., Rieke R., Chechulin A., & Kraus C. (2021). Cyberattack Detection in Vehicles using Characteristic Functions, Artificial Neural Networks, and Visual Analysis. — Informatics and Automation. Vol. 20(4). — Pp. 845–868. 10.15622/ia.20.4.4.

Jabia Nzi J. M., & Safaryan O. A. (2022). Investigation of DDoS attack detection using machine learning // The Young Researcher of the Don. No. 6(39). URL: https://cyberleninka.ru/article/n/issledovanie-obnaruzheniya-ddos-atak-s-ispolzovaniem-mashinnogo-obucheniya [in Russ.].

Maltseva I., Chernysh Yu., & Protsyuk Yu. (2024). Analysis of algorithms for early detection of cyber attacks on networks using machine learning // Communication, Informatization, and Cybersecurity Systems and Technologies. Vol. 1(6). Pp. 105–115. 10.58254/viti.6.2024.08.105 [in Ukr.].

Sharma A., & Babbar H. (2024). Detecting cyber threats in real time: A supervised learning perspective on the CTU-13 dataset // Proceedings of the IEEE International Conference for Innovation in Technology (INOCET). 10.1109/INCET61516.2024.10593100.

SaiKiran N., & Jagadeesh K.A. (2025). An intelligent approach to cyber-attack detection in networks using machine learning techniques // *International Journal of Research and Innovation in Applied Science*. Pp. 1351–1358. 10.51584/ijrias.2025.100800117.

Roman R., Lopez J., & Mambo M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. Future Generation Computer Systems. Vol. 78(2). Pp. 680–698. 10.1016/j.future.2016.11.009.

Rahman M. T., & Rahman Md. K. (2025). Machine learning algorithms for monitoring and detecting cyber attacks. // Proceedings of the International Conference on Machine Learning Applications. 2025. Pp. 1–6. 10.1109/mac64480.2025.11140542.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Wagh A., Pawar R., Wable N., Wandhekar S., & Dighe M. S. (2024). Detection of cyber attacks and network attacks using machine learning algorithms // *International Journal of Advanced Research in Science, Communication and Technology*. 10.48175/ijarsct-18161.

Xiao Y., Jia Y., Liu C., Cheng X., Yu J., & Lv W. (2019). Edge computing security: State of the art and challenges. // Proceedings of the // IEEE. 2019. Vol. 107(8). Pp. 1608–1631. URL: https://www.semanticscholar.org/paper/Edge-Computing-Security%3A-State-of-the-Art-and-Xiao-Jia/5f5ccd80381ca3593f9fc651844ed506894cbaf7

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 494–508
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.029

A HYBRID FRAMEWORK FOR RESUME-JOB MATCHING SYSTEM

B.A. Kumalakov, A.O. Dargulova*
Astana IT University, Astana, Kazakhstan.
E-mail: 242836@astanait.edu.kz

Bolatzhan A. Kumalakov — PhD in Computer Science, Associate 3rofessor, School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan https://orcid.org/0000-0003-1476-9542;
Aruzhan O. Dargulova — Master's student, School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
E-mail: 242836@astanait.edu.kz, https://orcid.org/0009-0004-0088-0543.

© B.A. Kumalakov, A.O. Dargulova

Abstract. In recent years, modern recruitment processes have generated an increasing number of job applicant submissions, and this is resulting in a growing need for developing a method for evaluating job applicant submissions consistently. In addition to surface text similarities used by automated matching tools and independent suitability prediction, there has been no method developed that captures how job applicants are compared during actual selection processes. This study proposes a hybrid decision-making methodology combining three components of contextualized text representations; standardized skill alignments through use of the ESCO classification system; explicit qualifications matching; and ranking based optimizations to enable the comparison of job applicants relative to one another. This study focuses on creating the model's architecture, testing the model on a data set of resumes paired with job postings and investigating how different information sources affect performance. The results demonstrate strong capability in identifying relevant candidates, stable predictive behavior and better discrimination than the baseline methodologies. Furthermore, the results also suggest that when semantic understanding is combined with structured competency constraints, a more reliable representation of hiring decisions can be produced. Therefore, this research concludes that multi-criteria ranking offers a viable basis for the development of AI-assisted recruitment systems and can facilitate the implementation of transparent and scalable candidate screening in actual organizational environments.

Keywords: resume-job matching, learning-to-rank models, ESCO skills

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

taxonomy, SBERT embeddings, feature engineering, recruitment analytics

For citation: B.A. Kumalakov, A.O. Dargulova (2026). A hybrid framework for resume-job matching system // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 494–508. https://doi.org/10.54309/IJICT.2026.25.1.029. (In Kaz.).

Conflict of interest: The authors declare that there is no conflict of interest.

ТҮЙІНДЕМЕЛЕР МЕН ВАКАНСИЯЛАРДЫ АВТОМАТТАНДЫРЫЛҒАН СӘЙКЕСТЕНДІРУГЕ НЕГІЗДЕЛГЕН ГИБРИДТІ ҮМІТКЕРЛЕРДІ ІРІКТЕУ ЖҮЙЕСІ

Б.А. Кумалаков, А.О. Даргулова*
Astana IT University, Астана, Қазақстан.
E-mail: 242836@astanait.edu.kz

Кумалаков Болатжан — PhD Информатика, «Жасанды интеллект және деректер ғылымы » мектебінің қауымдастырылған профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0003-1476-9542;
Даргулова Аружан — Магистрант, Жасанды интеллект және деректер ғылымы мектебі, Astana IT University, Астана, Қазақстан
E-mail: 242836@astanait.edu.kz, https://orcid.org/0009-0004-0088-0543.

© Б.А. Кумалаков, А.О. Даргулова

Аннотация. Қазіргі жұмысқа қабылдау процестері өтінімдердің жоғары көлемімен ерекшеленеді. Бұл жағдай кандидаттарды жүйелі және объективті тұрғыда бағалауды айтарлықтай қиындатады. Автоматтандырылған сәйкестендіру құралдары негізінен мәтіндік ұқсастыққа немесе үміткердің жарамдылығын дербес бағалауға сүйенеді, алайда аталған тәсілдер нақты іріктеу кезіндегі салыстырмалы бағалаудың мәнін толық қамти алмайды. Осы зерттеуде үміткерлерді салыстырмалы бағалауға арналған контекстік мәтіндік модельдерді, ESCO жіктемесіне негізделген дағдыларды стандартталған сәйкестендіруді, біліктіліктерді нақты салыстыруды және ранжирлеуді оңтайландыруды өзара кіріктіретін гибридті шешім қабылдау жүйесі ұсынылады. Жұмыстың мақсаты - модель архитектурасын құру және оны «түйіндеме - вакансия» жұптарына негізделген деректер жиынтығында сынақтан өткізу және әртүрлі ақпарат көздерінің алынған нәтижелерге тигізетін әсерін талдау. Ұсынылған тәсіл ең сәйкес үміткерлерді анықтау барысында жоғары тиімділік, болжамның тұрақтылығы және базалық әдістермен салыстырғанда жетілдірілген ажырату қабілеті тұрғысынан өзін-өзі дәлелдеді. Зерттеу нәтижелері семантикалық талдауды құрылымдалған құзыреттер шектеулерімен үйлестірудің жұмысқа қабылдау шешімдерін дәлірек модельдеуге жол ашатынын растайды. Көпкритерийлі ранжирлеу рекрутинг

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

саласындағы жасанды интеллектке негізделген шешімдерді қолдау жүйелерінің практикалық тірегіне айнала алады және үміткерлерді іріктеудің мөлдір әрі ауқымды тетігін қамтамасыз етеді деген қорытындыға келуге болады.

Түйін сөздер: түйіндеме мен вакансияны сәйкестендіру, ранжирлеуді үйрену модельдері, ESCO дағдылар жіктемесі, SBERT ендірмелері, белгілер инженериясы, рекрутинг аналитикасы

Дәйексөздер үшін: Б.А. Кумалаков, А.О. Даргулова (2026). Түйіндемелер мен вакансияларды автоматтандырылған сәйкестендіруге негізделген гибридті үміткерлерді іріктеу жүйесі // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. №. 25. Б. 494–508. https://doi.org/10.54309/IJICT.2026.25.1.029. (Қаз. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ГИБРИДНЫЙ ПОДХОД К АВТОМАТИЗИРОВАННОМУ ПОДБОРУ КАНДИДАТОВ НА ОСНОВЕ СОПОСТАВЛЕНИЯ РЕЗЮМЕ И ВАКАНСИЙ

Б.А. Кумалаков, А.О. Даргулова*
Astana IT University, Астана, Казахстан.
E-mail: 242836@astanait.edu.kz.

Кумалаков Болатжан — PhD, ассоциированный профессор Школы искусственного интеллекта и науки о данных, Astana IT University, Астана, Казахстан https://orcid.org/0000-0003-1476-9542;
Даргулова Аружан — магистрант, Школа искусственного интеллекта и науки о данных, Astana IT University, Астана, Казахстан
E-mail: 242836@astanait.edu.kz, https://orcid.org/0009-0004-0088-0543.

© Б.А. Кумалаков, А.О. Даргулова

Аннотация. На сегодняшний день процесс найма сопровождается значительным потоком заявок, что затрудняет детальную и беспристрастную оценку кандидатов. Автоматизированные инструменты часто полагаются на простое текстовое сходство или независимые оценки их пригодности, которые не учитывают фактические условия на рынке труда. Данное исследование предлагает гибридную систему принятия решений, совмещающую контекстные текстовые представления, стандартизированное сопоставление навыков на основе классификации ESCO, соответствие квалификациям и оптимизацию ранжирования для более точной оценки кандидатов. Цель работы заключается в создании архитектуры модели и ее тестировании на парных наборах «резюме - вакансия» и исследовании влияния различных источников данных на итоговое качество. Предложенный метод показывает высокую эффективность в выделении

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

наиболее подходящих кандидатов, стабильность прогнозов и улучшенную способность различать соответствия в сравнении с традиционными подходами. Результаты исследования показали, что семантический анализ, в сочетании со структурированными требованиями к компетенциям, более точно транслирует реальное принятие решений в сфере найма. Как итог можно утверждать, что многокритериальное ранжирование может служить практической основой для систем поддержки принятия решений в найме, обеспечивая прозрачность и воспроизводимость процесса подбора кандидатов в организациях.

Ключевые слова: сопоставление резюме и вакансий, модели обучения ранжированию, классификация навыков ESCO, эмбеддинги SBERT, инженерия признаков, аналитика рекрутинга

Для цитирования: Б.А. Кумалаков, А.О. Даргулова (2026). Гибридный подход к автоматизированному подбору кандидатов на основе сопоставления резюме и вакансий // Международный журнал информационных и коммуникационных технологий. Т. 7. №. 25. Стр. 494–508. https://doi.org/10.54309/ijict.2026.25.1.029. (На Каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Digital recruitment platforms now allow companies to find and assess potential employees more quickly than ever before. Companies receive numerous resumes for one position, yet can interview only a few. As a result, most companies utilize various forms of automated screening software to help them narrow their search down to the top candidates. Although early versions of screening software were based on keywords and manually created rules, they did not work well if a resume contained different vocabulary or formatting than expected by the applicant tracking system (ATS) (Faliagka et al., 2011, 2012). More recent studies applied machine learning and neural language models such as BERT (Barducci et al., 2022) to capture the relationships between resumes and job postings (Qin et al., 2018). Nevertheless, despite advancements in resume-screening technology, there still exists a fundamental issue that previous studies have been unable to resolve. How do we create a model that compares candidate qualifications against each other, and considers the formal requirements of a position? The lack of a solution to this problem is what drives this current study.

Automated resume-job matching continues to gain popularity as it can potentially make the recruiting process faster, and less biased toward individual preferences. However, the current body of literature does not provide a complete framework that utilizes semantic comprehension of text, standardized representations of competencies, explicit models of structured candidate attributes, and an interpretable ranking of candidates. Many systems have many similarities, but many systems fail to convey the comparative aspect of hiring decisions, and provide enough transparent reasoning for recruiters to rely on (Le et al., 2019; Raghavan et al., 2020). This disparity between technological ca-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

pabilities and the actual recruiting needs further illustrates the relevance of developing more inclusive decision-support models from both a theoretical and practical perspective.

A second unaddressed concern in the field of resume screening is the fragmentation of skill representation. One skill may appear in several different textual representations in a resume and a job posting, which can lower the reliability of the matching. Ontologies-based frameworks such as the European Skills, Competences, Qualifications and Occupations (ESCO) classification are designed to represent the relationships between skills in a standardized manner and to provide a common vocabulary for representing skills in resumes and job postings regardless of the vocabulary used (Le Wrang et al., 2014). These frameworks do enhance consistency and facilitate competency analysis; however, they are not incorporated into semantic models and ranking approaches. Additionally, the increasing number of AI-powered recruitment tools has raised questions regarding transparency and accountability, and therefore the need for decision-support systems that are interpretable (Raghavan et al., 2020).

This study's focus is automated candidate evaluation in digital recruitment environments. The focus of this study is a hybrid framework for resume-job matching that combines semantic representations, structured competency analysis, and ontology-based skill normalization in a ranking model. The goal of this study is to develop and empirically test an interpretable approach that enhances the accuracy and transparency of candidate ranking compared to the existing single-method systems.

In order to accomplish the goals of this study, three main tasks will be accomplished:

1. An analysis of the limitations of existing matching methods.
2. Design of a hybrid feature representation that includes semantic similarity, standardized skills, and structured attributes
3. Implementation of a learning-to-rank model that reflects comparative decision-making.
4. Assessment of the performance, interpretability, and practical application of the developed model.

The methodology of this study will include the utilization of transformer-based language models for semantic representation, ontology alignment using the ESCO classification for skill normalization; feature engineering of structured candidate characteristics; and gradient-boosted ranking algorithms with explainability techniques.

The working hypothesis of this study is that by incorporating multiple sources of information (semantic, structured, and ontology-based) will yield more accurate and interpretable candidate rankings than systems that rely solely on one form of representation. The scientific contribution of this study is to demonstrate how the above mentioned components can be integrated into a unified decision-support framework. The practical contribution of this study will be to determine whether the developed model will assist recruiters in evaluating candidates in high-volume hiring scenarios through the transparent ranking and explicit identification of competency gaps.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Materials and Methods.

Automated resume-job matching has been investigated as a comparative ranking problem which is based on how recruiters work with multiple candidates competing for one position. The early methods were using rule-based filtering and manually defined rules, both provide transparency in their decision making process however have severe limitations in processing candidate information with different terminology and changing skills required (Faliagka et al., 2011, 2012). Machine learning models using lexical representations and classification frameworks were then developed and showed a lot of improvement over the previous methods by providing a more robust method of candidate evaluation; although, most candidate evaluations are treated as independent from each other (Qin et al., 2018; Bian et al., 2019). A few recent advancements in neural language models have shown large improvements in the semantic understanding of resumes and job descriptions (Reimers & Gurevych, 2019; Barducci et al., 2022) however, many of these models do not take considered competency requirements or comparative ranking mechanisms. Therefore, the current study will develop and evaluate a hybrid framework that uses semantic similarity, standardized skill representations and structured compatibility indicators within a learning-to-rank model.

The experimental setting was developed from a database of resume–job pairing from the information technology field that was retrieved from online recruiting websites. After data cleaning to ensure all resume–job pairings were complete and consistent, there were 10,766 resume–job pairings derived from 873 job postings. In addition to the considerable number of resume–job pairings, each job posting had an average of 12 resumes to allow comparison of the ranking of candidates in job specific groupings. To preserve the technical expression of programming languages and tools used in the IT domain, the documents were normalized, duplicate entries removed and preserved.

As stated earlier, candidate selection was framed as a learning-to-rank problem where candidates are ranked in relation to other candidates vying for the same position. (Burges, 2010, Lee 2024 and Valizadegan et al., 2009) have demonstrated the success of learning-to-rank approaches in comparative decision making problems in Information Retrieval Systems. Each resume–job pairing was represented by a feature vector that consisted of a combination of semantic similarity, standardized skill overlap and structural compatibility indicator(s).

The semantic similarity between two documents was quantified using transformer-based sentence embeddings that capture relationships of context beyond the overlap of words (Reimers & Gurevych, 2019; Barducci et al., 2022). The cosine similarity between the vector representations of these embeddings is used as a measure of topical alignment between the two documents. To handle the variation in skill-related terms, competency phrases in resumes and job descriptions are linked to standardized competency concepts within the ESCO (European Skills, Competences, Qualifications and Occupations) taxonomy. Skill taxonomies provide consistent terminology for labor market competencies and allow for the comparative analysis of skills across different document formats (Le Wrang et al., 2014). Additional competency data is analyzed

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

in the context of occupational frameworks, like O*NET, that define the relationship between jobs, tasks and required skills (Peterson et al., 2001). Ontologies and ontology-based methodologies for competency management and resume annotation guided the normalization process (Karaa & Mhimdi, 2012; Janev & Vraneš, 2011).

Structured attributes, which represent the explicit recruitment requirements, were also extracted in addition to the semantic information. This includes the compatibility of educational level, years of work experience, certifications, and seniority indicator. While structured restrictions are commonly found in recruitment research, they are typically poorly implemented in the purely semantic matching models (Rentzsch & Staneva, 2020; Le et al., 2019). After normalizing the extracted attributes, all were then combined with semantic and skill-based features into a single representation.

To determine which candidates would be ranked first based upon what is provided in each candidate's resume, a gradient boosted decision tree model was employed in this study. The model was optimized for ranking purposes, as well as other objectives. The gradient boosting technique has been shown to provide high-quality results when utilizing heterogeneous feature sets, as well as being effective for ranking tasks, primarily because they are able to identify non-linear relationships between variables (Burges, 2010; Prokhorenkova et al., 2018). During training of the model, pairwise comparison was made by comparing the job posting itself to each candidate to establish a higher score for each candidate who is better suited to the particular job than the other candidates applying for the same position.

Model performance was assessed using a variety of ranking metrics that reflect real-world recruitment scenarios, such as normalized discounted cumulative gain (NDCG), mean reciprocal rank (MRR) and precision at top positions. NDCG measures both how relevant a list of candidates is to a particular job opening, as well as where those most relevant candidates appear in the list (Valizadegan et al., 2009). In addition, additional experiments were completed by removing selected elements of the model to evaluate the contributions of semantic, structured, and ontology-based features to the quality of the rankings produced by the model.

Feature attribution was utilized to examine the interpretability of the model's decisions regarding which candidates should be ranked first, in terms of the degree to which each feature contributed to the model's prediction of which candidate would receive the highest score (Lundberg & Lee, 2017). Additionally, competency gap analysis was employed to compare the required skills listed in the job description with the skills that can be found in the candidate resumes, allowing the researcher to identify the missing competencies and provide recommendations to develop them. This approach aligns with recent calls for transparent and accountable AI-based recruitment tools that will minimize bias while supporting human decision making (Raghavan et al., 2020).

This study uses a methodological framework of integrating semantic analysis, standardized competency representations, structured requirements, and ranking optimization, thus providing a systematic means of evaluating the hypothesis that combining multiple information sources will result in improved accuracy and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

transparency of the automated candidate evaluation process.

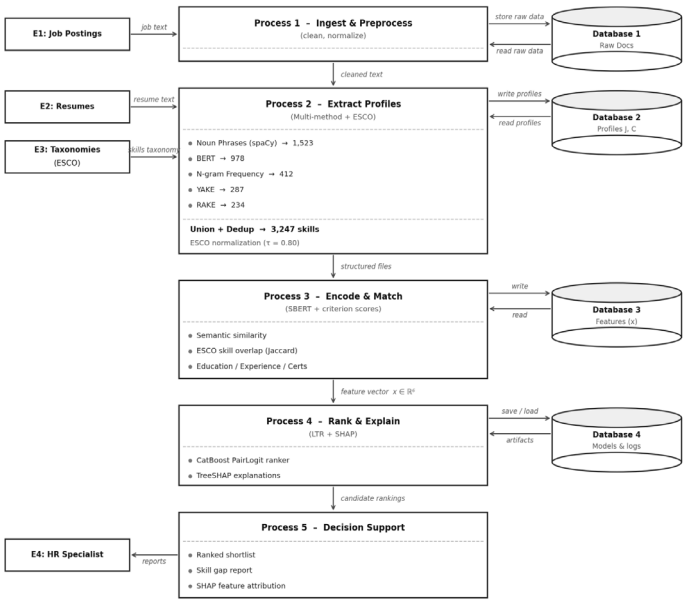


Fig.1. Proposed Hybrid System Architecture.

Results and Discussion.

Dataset Characteristics. The training set consisted of 7,536 pairs and the test and validation sets consisted of 1,615 pairs in each case. Each of these three sets had a perfect class distribution of about 50% Fit and 50% Not Fit. In addition, the training set contained 3,784 positive and 3,752 negative examples. It is important to maintain such a distribution to not train models that learn spurious associations based upon class frequencies instead of what really causes discrimination. Thus, all our evaluation metrics for the model's discriminative capability are independent of its potential exploitation of base rates.

Extraction Method	Skills	Precision	Recall	Percentage (%)
Noun Phrases (spaCy)	1,523	0,84	0,68	46.9
SBERT	978	0,88	0,72	30.1
N-gram Frequency	412	0,76	0,52	12.7
RAKE	287	0,71	0,48	8.8
YAKE	234	0,78	0,54	7.2
Combined	3,247	0,92	0,87	100.0

Skill Extraction Results

A multi-method skill-extraction pipeline extracted a total of 3247 unique skill phrases from the corpus following de-duplication and substring-filtering. The top contributor were extracted via noun phrase extraction using spaCy, which provided 1523

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

skills or 46.9 % of the total number of skills identified. This method had precision of 84 and a recall rate of 68 The second most successful method in terms of skill identification was SBERT, which identified 978 skills or 30.1 percent of the total number of skills identified; this method also demonstrated the best individual precision (.88) due to its ability to identify skills that are semantically important yet may appear rarely in the text. N-gram frequency mining identified 412 skills or 12.7 percent of the total number of skills identified with precision of 76 and a recall rate of 52, providing effective means to capture common technical acronyms yet missing low frequency terms. RAKE and YAKE identified 287 and 234 skills, respectively; however both methods demonstrated precision rates of 71 and 78, respectively, and were able to capture domain specific and emerging terminology not typically identified by the other methods.

The overall combination of methods significantly outperformed each individual method with a precision of 92 and a recall of 87 along with an overall F1 score of 784. Although the combination of methods resulted in a slight reduction in precision relative to SBERT alone, the significant increase in recall justified the use of a multi-method approach. Only 87 skills were identified by all five methods, demonstrating universal expression of competencies such as Python, SQL, and project management. The remainder of the 3160 skills identified were found to require at least one method beyond the simple method to be captured. The results confirm that each method provides non-overlapping coverage and that reliance upon a single method would result in a substantial portion of the skill vocabulary being missed.

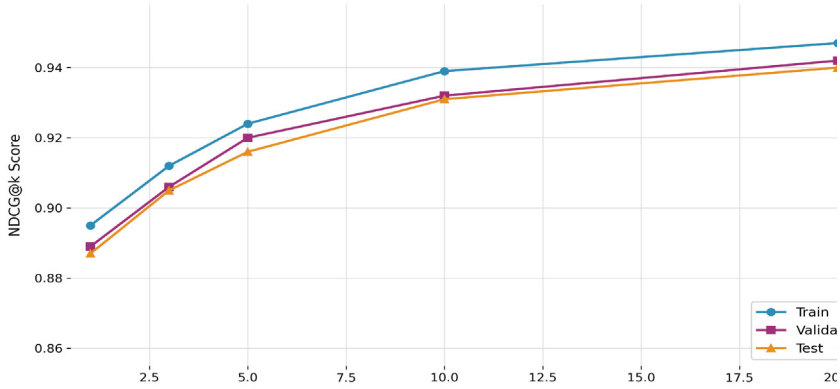


Fig.2. NDCG@k Performance Across Different Cutoffs

The NDCG@K (Normalized Discounted Cumulative Gain) at K cutoffs shown in Fig. 2 demonstrates the progression of the ranking quality based upon the increasing number of candidates being evaluated as K increases. A prominent level of effectiveness is achieved at the highest ranking positions; i.e., NDCG@1 = 0.895 and NDCG@5 = 0.916. Smaller incremental gains occur at the k = 10 position (NDCG@10 = 0.930). This trend suggests that most of the most relevant candidates are listed among the top ranked candidates, which is important since many recruiters evaluate potential applicants' qualifications through a limited pool of applicants and thus have limited

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

opportunity to consider all qualified candidates. In comparison to previously reported semantic-based matching techniques such as dense retrieval approaches (Yu et al., 2025), the hybrid approach exhibits an enhanced capacity to identify the elements that affect hiring decisions. Additionally, the concentration of the ranking signals in the top-ranked positions demonstrate that candidate suitability is a function of the combination of semantic fit, competency alignment, and structural requirements rather than any one factor individually.

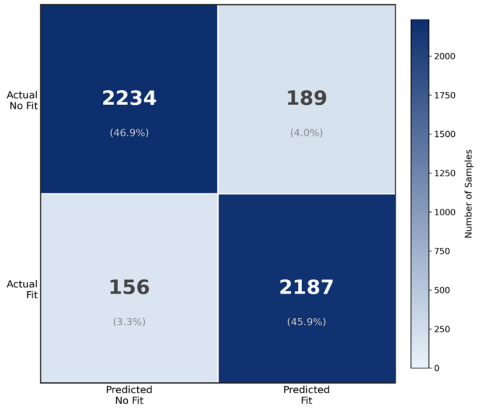


Fig.3. Confusion Matrix for aggregated evaluation

The strength of the prediction is shown in Figure 3, with a confusion matrix of the test data, with a strong diagonal showing the accuracy of identifying both suitable and unsuitable candidate populations; 2187 true positives and 2234 true negatives, with off-diagonal errors being small (189 false positives and 156 false negatives) to indicate balance in performance. The data demonstrates better performance than the classification-based methods used in prior recruitment research, where high levels of recall were common but at the expense of low levels of precision. The framework is providing a conservative level of risk tolerance, as would be expected in a screening context. However, there are more false negatives indicating that the model prefers to avoid recommending unsuitable candidates to the hiring manager's top recommendations list. This aligns with hiring managers' preferences in the hiring process as they have less preference for false positive, as those can be expensive.

The discriminatory ability of this framework is further demonstrated in figure 4 which displays both the combined receiver operating characteristic (ROC) and Precision-recall curves for both full model and baseline models; the hybrid model achieved an area under the curve (AUC) of 0.964 and demonstrates high Precision over all recall values (average Precision or AP = 0.921), indicating strong separation between relevant and non-relevant candidates in evaluation. Lower separation of relevance between candidates was observed in baseline models, supporting previous research that structured data alone does not capture the complexity of evaluating candidate fit to job openings. The findings also support previous research on using structured and textual information together in modeling person-job fit, but further support the idea that

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

combining structured and textual information provides more discrimination than using either form of information alone in competitive ranking scenarios.

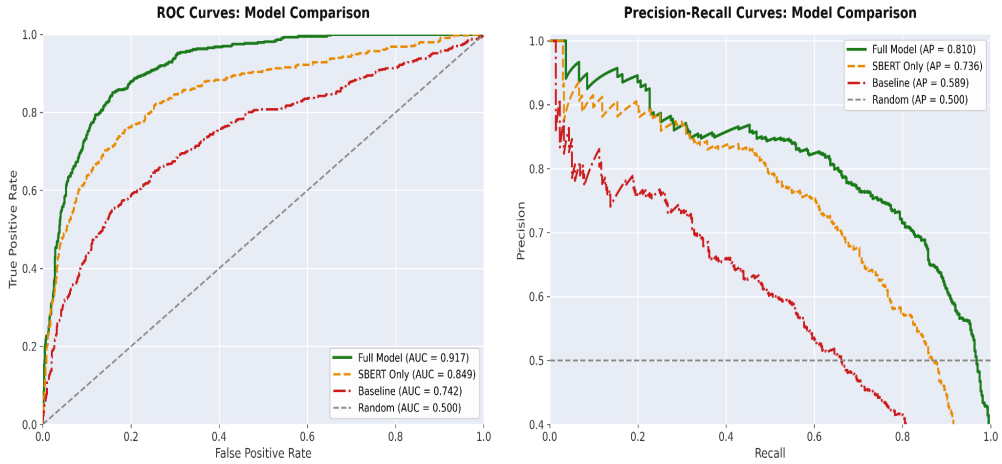


Fig.4. Combined ROC and Precision-Recall Curves

Table 2 – Ablation Study Results Showing Impact of Different Feature Groups.

Configuration	Features	NDCG@5	Δ from Full	% Decrease
Full Model	All 6 features	0.916	-	-
No Semantic	ESCO, edu, exp, cert, sen, dom	0.873	-0.043	-4.7%
No ESCO	Semantic, edu, exp, cert, sen, dom	0.901	-0.015	-1.6%
No Structured	Semantic, ESCO	0.884	-0.032	-3.5%
SBERT only	Semantic	0.798	-0.118	-12.9%
ESCO Only	ESCO	0.745	-0.171	-18.7%
Structured Only	edu, exp, cert, sen, dom	0.812	-0.104	-11.4%

The contributions of individual feature sets were analyzed in Table 2 using an ablation study. In removing the skill normalization based on ESCO, it resulted in decreased performance because of the varying competency representations generated by use of various terminology. This supports previous research regarding the value of standardizing skill taxonomies for improving the quality of labor-market analytics. Removing structured compatibility indicators increased similarity scores for candidates who lack necessary qualifications. This demonstrates that semantic alignment alone will generate erroneous rankings when combined with constraint modeling (i.e., when there is no model constraint). Removing semantic features produced the greatest decrease in performance, which suggests that the context of job descriptions is critical to evaluating resumes. Collectively, the results demonstrate that the hybrid model's improved performance is derived from the synergy among its feature types rather than from a single dominant feature set.

Further insight into the resume-job matching process was gained from experiments

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

that did not include visual representations of the data. The results demonstrated that the performance of the ranked lists was consistent across all job categories represented in the dataset, and therefore, the framework may be applied to a wide variety of technical jobs. Most misclassified candidates were those whose skills, although applicable, were described using unconventional terminology, or partially met the requirements listed in the job description. Misclassification also highlights the subjective nature of resume evaluation. The results indicated that the importance of structured attributes increased in senior-level positions, where formal education and work experience were emphasized; however, the importance of semantic similarity increased in junior level positions that placed greater emphasis on adaptable skills.

The study identified resume-job matching as a multi-faceted assessment issue that includes contextually similar or dissimilar features, and competencies matched on a standardized basis, and qualifications specifically stated in resumes. Prior to the development of this framework, most candidate evaluations were treated independently as classification problems. The comparative ranking approach adopted in the framework represents the competitive nature of recruitment, in which candidates are evaluated relative to one another. The framework incorporates multiple decision criteria and generates interpretable rankings aligned with real-world hiring practices. In addition to improving decision quality and transparency, the framework provides a means of integrating semantic, structural, and ontology-based information. As such, the framework addresses some of the limitations that have been identified in the development of automated hiring tools.

Prediction Calibration.

In addition to evaluating ranking performance, the value of the proposed framework will depend upon whether a recruiter should trust predicted relevance scores as actual probability estimates or simply treat them as relative rankings. The results of the calibration analysis indicated that the model was producing very well calibrated predictions. Predicted probabilities were equal to observed match frequency rates across the entire 0-1 probability space (i.e., the model's predictions did not appear to either significantly overestimate or underestimate probability). Additionally, every calibration point fell within +/−10 % of perfect calibration. Conversely, the baseline model had significant overconfidence at higher predicted probabilities; i.e., the baseline model frequently overestimated the likelihood of a good match for those candidates that it ranked highest. The practical implications of this calibration quality are numerous, including enabling recruiters to interpret model score predictions as representing 85 % chance that the candidate is a good fit. Well-calibrated model scores also facilitate threshold based automation (e.g., advance to secondary review all candidates receiving a model score > x) with both predictable and auditable outcomes. Threshold-based automation may be particularly valuable under current employment regulations in multiple jurisdictions that require documentation justifying the basis for automated screening decisions.

Model Training Stability.

The model's training behavior was tracked through each of the 1,000 iterations of training as well as by monitoring both the training and validation sets for early

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

stopping. Training loss continuously decreased as it converged to a single value after about 600 iterations of training. Validation loss showed a similar trend as training loss but had some small variability; its minimum occurred at the 750th iteration where early stopping would be initiated (the best performing checkpoint up until then would be retained). Early stopping did result in an only 0.5 percent difference in how well the training and validation datasets performed in terms of NDCG@5. At the time of early stopping, the training NDCG@5 was 0.924 and the validation NDCG@5 was 0.919. Given this similarity, the model can generalize as well as maintain stability in training using all three types of data in the hybrid multi-source feature set including: semantic embedding based data, ontology aligned skill representations and structurally-based data representing attributes.

Finally, the results demonstrate that integrating complementary information sources leads to significant improvements in automated candidate evaluations when compared to evaluations made using singular methods. Furthermore, the study demonstrates that effective resume-job matching requires the concurrent consideration of multiple aspects of a candidate's suitability for a particular position. The study identifies the underlying structure of hiring decisions and establishes a basis for developing more dependable AI-assisted hiring systems.

Conclusion.

This study views resume-job matching as a comparative decision-making process and shows that modeling candidate evaluation as a learning-to-rank (L2R) problem provides a model that better represents actual recruitment processes than separate classification methods do. The hybrid system models many factors related to candidate suitability through an integrated framework of contextual semantic representations, ontology-based skill normalization using the ESCO taxonomy and structural qualification constraints. The interactions among all three information types result in consistent increases in the quality of rankings, discrimination capacity, and calibration stability when comparing to the individual-method baselines.

The ablation and calibration analyses demonstrated that candidate evaluation can be successful based on the reliability and interpretability of ranking signals and the predictive accuracy of the model's output. Model outputs that are well-calibrated allow recruiters to view model scores as reasonable measures of candidate suitability instead of viewing them as a set of abstract numbers. Such calibration will enable recruiters to use decision thresholds to decide of which candidates are suitable to move forward in the hiring process and facilitate auditable screening processes. Calibration is important in large-scale recruitment settings due to the costs of false positives and false negatives.

In addition to improving performance, the framework provides methodological clarity by representing the competitive nature of hiring decisions. Most traditional systems for screening resumes treat candidates separately, while most hiring processes compare candidates for the same position in relation to each other. The comparative logic applied in the ranking formulation used in this study provides a way to represent this comparative process and identify competency gaps between candidates and job

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

requirements. Therefore, the usage of the system extends from screening to providing decision support for workforce planning, targeted training, and talent development strategies. However, the findings also suggest that there are limitations to the automation of evaluations of human qualifications. Misclassifications typically occur due to non-standard terminology, interdisciplinary skill profiles or partially met requirements. These examples reflect the subjectivity and contextual nature of resume interpretation. That is why, automated systems should serve as a tool to enhance human judgment and not as a replacement for human judgment.

Future research would investigate the extent to which the proposed system can generalize across different industries (cross-domain), whether the proposed system can be adapted for international labor markets (multilingual), if fairness-conscious ranking objectives can be developed, and whether the proposed system can accommodate temporal changes in career trajectory to capture changing competencies. It may also be beneficial to evaluate incorporating human-in-the-loop feedback mechanisms to continuously align the model's outputs to the organization's hiring preferences. Furthermore, the proposed hybrid ranking framework has established a theoretically-sound and practically viable basis for developing transparent, scalable, and interpretable AI-assisted recruitment systems that can support complex hiring decisions in today's digital labor market.

REFERENCES

Barducci A., Iannaccone S., La Gatta V., Moscato V., Sperlì G., Zavota S. (2024). An End-to-End Framework for Information Extraction from Italian Resumes. — Expert Systems with Applications. — Vol. 210. — Article 118487. 10.1016/j.eswa.2022.118487. (in English)

Bian S., Zhao W. X., Song Y., Zhang T., Wen J.-R. (2019). Domain Adaptation for Person-Job Fit with Transferable Deep Global Match Network // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics. Pp. 4810–4820. 10.18653/v1/D19-1487. (in English)

Burges C.J.C. (2010). From RankNet to LambdaRank to LambdaMART: An Overview // Microsoft Research Technical Report MSR-TR-2010-82, Microsoft Research. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/MSR-TR-2010-82.pdf. (in Eng)

Faliagka E., Ramantas K., Tsakalidis A. K., Viennas M., Kafeza E., Tzimas G. (2011). An Integrated e-Recruitment System for CV Ranking Based on AHP // Proceedings of WEBIST 2011, SciTePress. Pp. 147–150. URL: https://www.scitepress.org/papers/2011/33379/33379.pdf. (in Eng)

Faliagka E., Karydis I., Rigou M., Sioutas S., Tsakalidis A., Tzimas G. (2012). Taxonomy Development and Its Impact on a Self-Learning e-Recruitment System // Artificial Intelligence Applications and Innovations (IFIP Advances in Information and Communication Technology,), Springer. Vol. 381. Pp. 165–174. 10.1007/978-3-642-33409-2_18. (in Eng)

Gao T., Yao X., Chen D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) // Association for Computational Linguistics. Pp. 6894–6910. 10.18653/v1/2021.emnlp-main.552. (in Eng.)

Janev V., Vraneš S. (2011). Ontology-Based Competency Management: The Case Study of the Mihajlo Pupin Institute // *Journal of Universal Computer Science*. Vol. 17. No. 7. Pp. 1089–1108. 10.3217/jucs-017-07-1089. (in English)

Karaa W. B. A., Mhimdi N. (2012). Using Ontology for Resume Annotation. — International Journal of Metadata, Semantics and Ontologies. Vol. 6. No. 3–4. Pp. 166–174. 10.1504/IJMSO.2011.048018. (in Eng.)

Le R., Hu W., Song Y., Zhang T., Zhao D., Yan R. (2019). Towards Effective and Interpretable Person-Job Fitting. — Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19), ACM. Pp. 1883–1892. 10.1145/3357384.3357949. (in Eng.)

Le Vrang M., Papantoniou A., Pauwels E., Fannes P., Vandensteen D., De Smedt J. (2014). ESCO: Boosting

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Job Matching in Europe with Semantic Interoperability. — Computer. — Vol. 47. — No. 10. Pp. 57–64. 10.1109/MC.2014.283. (in Eng.)

Lee J., Bernier-Colborne G., Maharaj T., Vajjala S. (2024). Methods, Applications, and Directions of Learning-to-Rank in NLP Research. — Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics. Pp. 1900–1917. 10.18653/v1/2024.findings-naacl.123. (in Eng.)

Lundberg S. M., Lee S.-I. (2017). A Unified Approach to Interpreting Model Predictions //Advances in Neural Information Processing Systems. Curran Associates. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. (in Eng.)

Peterson N. G., Mumford M. D., Borman W. C., Jeanneret P. R., Fleishman E. A. (2001). Understanding Work Using the Occupational Information Network (O*NET): Implications for Practice and Research. Personnel Psychology. Vol. 54. No. 2. Pp. 451–492. 10.1111/j.1744-6570.2001.tb00100.x. (in Eng.)

Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. (2018). CatBoost: Unbiased Boosting with Categorical Features // Advances in Neural Information Processing Systems, Curran Associates. Vol. 31. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf. (in Eng.)

Qin C., Zhu H., Xu T., Zhu C., Jiang L., Chen E., Xiong H. (2018). Enhancing Person-Job Fit for Talent Recruitment: An Ability-Aware Neural Network Approach // Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM. Pp. 25–34. 10.1145/3209978.3210025. (in Eng.)

Raghavan M., Barocas S., Kleinberg J., Levy K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. — Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency. ACM. Pp. 469–481. 10.1145/3351095.3372828. (in Eng.)

Reimers N., Gurevych I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)., Association for Computational Linguistics. Pp. 3982–3992. 10.18653/v1/D19-1410. (in Eng.)

Rentzsch R., Staneva M. (2020). Skills-Matching and Skills Intelligence Through Curated and Data-Driven Ontologies // Proceedings of the DELFI Workshops (2020). Gesellschaft für Informatik. Pp. 46–58. URL: https://dl.gi.de/server/api/core/bitstreams/849cd40d-5e27-4774-9f74-440c329e51c2/content. (in Eng.)

Valizadegan H., Jin R., Zhang R., Mao J. (2009). Learning to Rank by Optimizing NDCG Measure. — Advances in Neural Information Processing Systems, Curran Associates. URL: https://papers.nips.cc/paper_files/paper/2009/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf. (in Eng.)

Yu X., Xu R., Xue C., Zhang J., Ma X., Yu Z. (2025). ConFit v2: Improving Resume-Job Matching Using Hypothetical Resume Embedding and Runner-Up Hard-Negative Mining // Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics. P. 12775–12790. 10.18653/v1/2025.findings-acl.661. (in Eng.)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.1. Number 25 (2026). Pp. 509–532
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.25.1.030
UDC 004.032.26

MATHEMATICAL MODEL FOR OPTIMAL SENSOR SELECTION IN SIEM SYSTEMS USING THE ANALYTIC HIERARCHY PROCESS

V. Makhatova[1], B. Dzhugembayeva[1]*, A. Gabdulova[1], L. Nurgaliyeva[2], A. Abdigaliyeva[3]
[1]Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[2]Saidot Ltd., Helsinki, Finland;
[3]Safi Utebayev Atyrau Oil and Gas University, Atyrau, Kazakhstan.
E-mail: asbaku@mail.ru

Valentina Makhatova — Candidate of Technical Sciences, Professor of the Department of Software Engineering**,** Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan https://orcid.org/0000000240829193;
Bakhytgul Dzhugembayeva — Ms.Sc., Senior Lecturer Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
E-mail: asbaku@mail.ru, https://orcid.org/0000-0002-2697-5194;
Aigul Gabdulova — Ms.Sc., Senior Lecturer, Department of Software Engineering, Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-6589-854;
Lunara Nurgaliyeva — Ms.Sc., AI Safety ML/LLM Engineering, Saidot Ltd. Helsinki, Finland
https://orcid.org/0009-0005-5252-9525;
Akmaral Abdigaliyeva — Ms.Sc., Senior Lecturer, Faculty of Information Technology, Safi Utebayev Atyrau Oil and Gas University**,** Atyrau, Kazakhstan
https://orcid.org/0009-0003-7907-6875.

© V. Makhatova, B. Dzhugembayeva, A. Gabdulova, L. Nurgaliyeva, A. Abdigaliyeva

Abstract. This paper presents a mathematical model for the optimal selection of sensors in Security Information and Event Management (SIEM) systems using the Analytic Hierarchy Process (AHP). The growing complexity of modern information infrastructures and the increasing number of cyber threats require reliable and efficient monitoring mechanisms. Since the effectiveness of a SIEM system significantly depends on the performance and configuration of its sensors, the problem of selecting the most suitable sensor under multi-criteria conditions becomes a relevant scientific and practical task. The proposed approach formalizes the sensor selection process as a three-level hierarchical structure that includes the main objective, a system of evaluation criteria,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

and alternative sensor configurations. The criteria considered in the study include system load, reaction time, working time, efficiency, implementation cost, labor intensity, universality, implementation quality, and prevalence. Pairwise comparisons were performed according to the Saaty scale, and weighting coefficients were calculated using eigenvalue-based methods. The consistency index and consistency ratio were evaluated to ensure the reliability of expert judgments.Based on the developed model, a software tool was implemented in C++ using the MySQL database management system. The system automates the formation of comparison matrices, calculation of priority vectors, and ranking of alternatives. Experimental results demonstrate that the application of AHP improves the objectivity and transparency of decision-making, reduces configuration time, and increases the reliability of SIEM sensor deployment. The proposed model is scalable and can be adapted to various information security infrastructures, contributing to the advancement of multi-criteria optimization methods in cybersecurity.

Keywords: information security, SIEM, sensor, Analytic Hierarchy Process (AHP), decision-making

For citation: V. Makhatova, B. Dzhugembayeva, A. Gabdulova, L. Nurgaliyeva A. Abdigaliyeva (2026). Mathematical model for optimal sensor selection in siem systems using the analytic hierarchy process // International journal of information and communication technologies. Vol. 7. No. 25. Pp. 509–532. https://doi.org/10.54309/IJICT.2026.25.1.030. (In Eng.).

Conflict of interest: The authors declare that there is no conflict of interest.

ИЕРАРХИЯЛАРДЫ ТАЛДАУ ӘДІСІ НЕГІЗІНДЕ SIEM ЖҮЙЕЛЕРІНДЕ ОҢТАЙЛЫ СЕНСОРДЫ ТАҢДАУДЫҢ МАТЕМАТИКАЛЫҚ МОДЕЛІ

В. Махатова[1], Б. Джугембаева[1]*, А. Габдулова[1], Л. Нургалиева[2], А. Абдигалиева[3]

[1]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;
[2]Saidot Ltd. Хельсинки, Финляндия;
[3]С.Өтебаев атындағы Атырау мұнай және газ университеті, Атырау, Қазақстан.
E-mail: asbaku@mail.ru

Валентина Махатова — техника ғылымдарының кандидаты, Х. Досмұхамедов атындағы Атырау университетінің «Бағдарламалық инженерия» кафедрасының профессоры, Атырау, Қазақстан
https://orcid.org/0000000240829193;
Бакытгул Джугембаева — сеньор лектор, Х. Досмұхамедов атындағы Атырау университетінің «Физика және техникалық пәндер» кафедрасы, Атырау, Қазақстан
E-mail: asbaku@mail.ru, https://orcid.org/0000-0002-2697-5194;
Айгул Габдулова — сеньор лектор, Х. Досмұхамедов атындағы Атырау университетінің «Бағдарламалық инженерия» кафедрасы, Атырау, Қазақстан
https://orcid.org/0000-0002-6589-854X;

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Лунара Нургалиева — магистр, AI Safety ML/LLM Engineering, Saidot Ltd. Helsinki, Finland
https://orcid.org/0009-0005-5252-9525;
Ақмарал Абдигалиева — магистр, Ақпараттық технологиялар факультеті, Сафи Өтебаев атындағы Атырау мұнай және газ университеті, Атырау, Қазақстан.
https://orcid.org/0009-0003-7907-6875.

© В. Махатова, Б. Джугембаева, А. Габдулова, Л. Нургалиева, А. Абдигалиева

Аннотация. Бұл мақалада SIEM класты ақпараттық қауіпсіздікті басқару жүйелерінде оңтайлы сенсорды таңдау үшін иерархияларды талдау әдісі (ИТӘ) негізінде математикалық модель ұсынылады. Заманауи ақпараттық инфрақұрылымдардың күрделенуі және киберқауіптердің артуы сенімді әрі тиімді мониторинг тетіктерін қажет етеді. SIEM жүйесінің тиімділігі көбінесе қолданылатын сенсорлардың сипаттамалары мен конфигурациясына тәуелді болғандықтан, көпкритерийлі ортада ең қолайлы сенсорды таңдау ғылыми әрі практикалық тұрғыдан өзекті мәселе болып табылады. Ұсынылған тәсіл сенсорды таңдау үдерісін үш деңгейлі иерархиялық құрылым түрінде формалдайды: негізгі мақсат, бағалау критерийлері және баламалы сенсорлар. Зерттеуде келесі критерийлер ескерілді: жүйеге түсетін жүктеме, жауап беру уақыты, жұмыс уақыты, тиімділік, енгізу құны, енгізудің еңбек сыйымдылығы, әмбебаптық, жүзеге асыру сапасы және таралу деңгейі. Жұптық салыстырулар Саати шкаласы бойынша жүргізіліп, салмақ коэффициенттері меншікті мәндер әдісі арқылы есептелді. Сараптамалық бағалардың дұрыстығын тексеру үшін келісімділік индексі және келісімділік қатынасы анықталды. Әзірленген модель негізінде C++ бағдарламалау тілінде және MySQL деректер қорын басқару жүйесін қолдана отырып бағдарламалық құрал жасалды. Жүйе жұптық салыстыру матрицаларын құруды, басымдық векторларын есептеуді және баламаларды ранжирлеуді автоматтандырады. Эксперименттік нәтижелер иерархияларды талдау әдісін қолдану шешім қабылдау үдерісінің объективтілігі мен айқындығын арттыратынын, жүйені баптау уақытын қысқартатынын және SIEM инфрақұрылымында сенсорды таңдаудың сенімділігін жоғарылататынын көрсетті. Ұсынылған модель масштабталатын болып табылады және әртүрлі ақпараттық қауіпсіздік жүйелеріне бейімделе алады, бұл киберқауіпсіздік саласындағы көпкритерийлі оңтайландыру әдістерін дамытуға ықпал етеді.

Түйін сөздер: ақпараттық қауіпсіздік; SIEM; сенсор; иерархияларды талдау әдісі; шешім қабылдау

Дәйексөздер үшін: В. Махатова, Б. Джугембаева, А. Габдулова, Л. Нургалиева А. Абдигалиева (2026). Иерархияларды талдау әдісі негізінде siem жүйелерінде оңтайлы сенсорды таңдаудың математикалық моделі // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. № 25. Б. 509–532. https://doi.org/10.54309/IJICT.2026.25.1.030. (Ағыл. тіл.).

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

деп мәлімдейді.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ВЫБОРА ОПТИМАЛЬНОГО СЕНСОРА В SIEM-СИСТЕМАХ СРЕДСТВАМИ МЕТОДА АНАЛИЗА ИЕРАРХИЙ

В. Махатова[1], Б. Джугембаева[1]*, А. Габдулова[1], Л. Нургалиева[2], А. Абдигалиева[3]

[1]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан;
[2]Saidot Ltd. Хельсинки, Финляндия;
[3]Атырауский университет нефти и газа имени С.Утебаева, Атырау, Казахстан.
E-mail: asbaku@mail.ru

Валентина Махатова — кандидат технических наук, профессор кафедры «Программная инженерия» Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0002-4082-9193;
Бакытгуль Джугембаева — магистр, старший преподаватель Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
E-mail: asbaku@mail.ru, https://orcid.org/0000-0002-2697-5194;
Айгуль Габдулова — магистр, старший преподаватель кафедры «Программная инженерия» Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0002-6589-854X;
Лунара Нургалиева — магистр, инженер по безопасности ИИ и машинного обучения/больших языковых моделей (AI Safety ML/LLM Engineering), компания Saidot Ltd. Хельсинки, Финляндия
https://orcid.org/0009-0005-5252-9525;
Акмарал Абдигалиева — магистр, старший преподаватель факультета информационных технологий Атырауского университета нефти и газа имени Сафи Утебаева, Атырау, Казахстан
https://orcid.org/0009-0003-7907-6875.

© В. Махатова, Б. Джугембаева, А. Габдулова, Л. Нургалиева, А. Абдигалиева

Аннотация. В данной статье представлена математическая модель выбора оптимального сенсора в системах управления информационной безопасностью класса SIEM на основе метода анализа иерархий (МАИ). Возрастающая сложность современных информационных инфраструктур и увеличение количества киберугроз требуют надежных и эффективных механизмов мониторинга. Поскольку эффективность функционирования SIEM-системы во многом зависит от характеристик и конфигурации используемых сенсоров, задача выбора наиболее подходящего сенсора в условиях многокритериальности является актуальной на-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

учной и практической проблемой. Предложенный подход формализует процесс выбора сенсора в виде трехуровневой иерархической структуры, включающей целевую функцию, систему критериев оценки и альтернативные варианты сенсоров. В исследовании учитываются следующие критерии: нагрузка на систему, время реакции, рабочее время, эффективность, стоимость реализации, трудоёмкость внедрения, универсальность, качество реализации и распространённость. Парные сравнения проводились по шкале Саати, а весовые коэффициенты рассчитывались на основе методов определения собственных значений. Для проверки корректности экспертных оценок были вычислены индекс согласованности и отношение согласованности. На основе разработанной модели реализовано программное обеспечение на языке C++ с использованием СУБД MySQL. Система автоматизирует формирование матриц парных сравнений, расчет векторов приоритетов и ранжирование альтернатив. Результаты экспериментов показывают, что применение метода анализа иерархий повышает объективность и прозрачность принятия решений, сокращает время настройки системы и увеличивает надежность выбора сенсоров для SIEM-инфраструктуры. Предложенная модель является масштабируемой и может быть адаптирована к различным системам информационной безопасности, способствуя развитию методов многокритериальной оптимизации в сфере кибербезопасности.

Ключевые слова: информационная безопасность; SIEM; сенсор; метод анализа иерархий; принятие решений

Для цитирования: В. Махатова, Б. Джугембаева, А. Габдулова, Л. Нургалиева А. Абдигалиева (2026). Математическая модель выбора оптимального сенсора в siem-системах средствами метода анализа иерархий. // Международный журнал информационных и коммуникационных технологий. Т. 7. №. 25. Стр. 509–532. https://doi.org/10.54309/IJICT.2026.25.1.030. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Modern information systems are an integral part of the infrastructure of enterprises, government agencies, and commercial organizations. As data volumes and levels of automation grow, the number of information security threats increases significantly, requiring continuous improvement of protection and monitoring methods. One of the most effective tools is a security information and event management system (SIEM)**,** which collects, correlates, and analyzes data on events occurring within the protected network.

The effectiveness of a SIEM system depends on the proper operation of its sensors—the components that collect, filter, and transmit security event data. An incorrectly selected or improperly configured sensor can lead to the loss of critical data, false alarms, or increased system load. Therefore, the problem of selecting the optimal sensor that balances performance, reliability, and cost is a pressing scientific and practical is-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

sue.

Traditional sensor selection methods rely primarily on expert assessments and empirical data, which do not always account for the multi-criteria nature and mutual influence of system parameters. Therefore, it is advisable to apply formalized mathematical methods of multi-criteria analysis**,** capable of considering both quantitative and qualitative characteristics. One of the most universal approaches is the Analytic Hierarchy Process (AHP), proposed by T. Saaty, which is widely used in optimization and decision support problems.

Using the Analytic Hierarchy Process (AHP) allows us to represent the sensor selection process as a hierarchical structure: from the primary goal—improving the effectiveness of the security system—to subordinate levels of criteria and alternatives. This model enables us to objectively assess the importance of each criterion, determine the weights of alternatives, and formulate a quantitative justification for the selection.

The objective of this study is to develop a mathematical model and software tool that enable rational selection of the optimal sensor for a SIEM system using the Analytic Hierarchy Processing (AHP) method. To achieve this goal, the following tasks are addressed:

- An analysis of existing approaches to the construction and configuration of SIEM systems was conducted;
- a hierarchical structure of sensor selection criteria has been formed;
- an algorithm for calculating weights and consistency indices has been implemented;
- a software module was developed in C++ using the MySQL DBMS;
- An experimental verification of the correctness and effectiveness of the proposed model was carried out.

The results of the study are aimed at improving the efficiency of SIEM system setup and operation, reducing the risk of information incidents, and ensuring a higher level of information infrastructure security.

Materials and Methods

SIEM Systems.

Security Information Event Management (SIEM) is the general name for software products previously used separately from each other, categories SIM (Security Information Management) and SEM (Security Event Management).

A typical SIEM system faces the following tasks.

Consolidation and storage of event logs from various sources – network devices, applications, OS logs, security tools. Sometimes an incident is detected late, and the events have long since been deleted, or the event logs are inaccessible for some reason, making it impossible to identify the cause of the incident. Furthermore, connecting to each source and viewing events is time-consuming.

Providing tools for event analysis and incident resolution. Event formats vary across various sources. Text-based formats can be cumbersome when dealing with large volumes and reduce the likelihood of incident detection. Some SIEM products standard-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

ize events and make them more readable, while the interface visualizes only valuable information events, highlights them, and allows filtering out non-critical events.

Correlation and rule-based processing. A single event doesn't always indicate an incident. The simplest example is "login failed": one instance is insignificant, but three or more such events involving the same account may indicate brute-force attacks. In the simplest case, rules in SIEM are represented in RBR (RuleBasedReasoning) format and contain a set of conditions, triggers, counters, and action scripts.

Automatic notification and incident management.

The primary goal of a SIEM is not simply to collect events, but to automate the process of incident detection, documenting them in its own log or an external HelpDesk system, and providing timely notification of events. A SIEM can detect:

network attacks on internal and external perimeters;

viral epidemics or individual viral infections, unremoved viruses, backdoors and Trojans;

attempts to gain unauthorized access to confidential information;

errors and failures in the operation of information systems;

vulnerabilities;

Configuration errors in security tools and information systems.

A SIEM system is versatile thanks to its logic. But to accomplish its intended tasks, useful sources and correlation rules are necessary. Any event (for example, a door opening in a specific room) can be fed to the SIEM input and used.

Sources are selected based on the following factors:

criticality of the system (value, risks) and information (processed and stored);

reliability and informativeness of the source of events;

coverage of information transmission channels (not only the external but also the internal perimeter of the network must be considered);

solving a range of IT and information security problems (ensuring continuity, incident investigation, policy compliance, preventing information leaks, etc.).

Main sources of SIEM

AccessControl, Authentication – for monitoring access control to information systems and the use of privileges.

Server and workstation event logs – for access control, ensuring continuity, and compliance with information security policies.

Network active equipment (change and access control, network traffic counters).

IDS/IPS. Events about network attacks, configuration changes, and device access.

Antivirus protection. Events about software performance, databases, configuration and policy changes, and malware.

Vulnerability scanners. Inventorying assets, services, software, and vulnerabilities, providing inventory data and topology structure.

GRC systems for risk accounting, threat criticality, and incident prioritization.

Other systems for protecting and monitoring information security policies: DLP,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

anti-fraud, device control, etc.

Inventory and asset management systems. To monitor infrastructure assets and identify new ones.

Netflow and traffic accounting systems.

A SIEM solution typically includes several components (Figure 1):

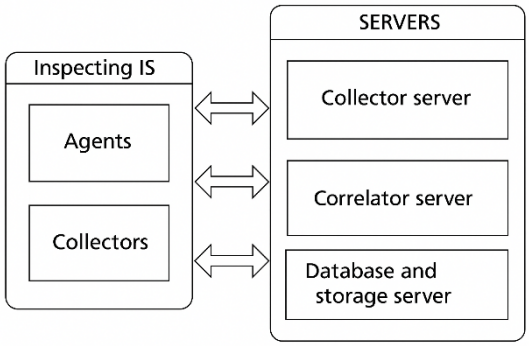


Fig. 1. SIEM structure.

Agents installed on the information system in question (relevant for OS; an agent is a separate program (service) that locally checks all event services and sends statistics to the server);

Agent collectors, otherwise known as modules for understanding a separate log of all system events;

Collector server, required for initial collection of events from various sources;

Correlator server, required for collecting data from collectors and agents and studying the received data using algorithms and rules of mutual intersection.

Database server and storage, which are responsible for storing all data.

Often, a SIEM system is presented in an agent architecture - a place for storing information - a program server installed on top of the protected IT structure (Zhumabekov, 2021)

Agents implement the collection of security events, their rapid analysis and filtering by type.

All filtered security event data is transferred to a storage or a specific case, where it is placed in an internal presentation format for further use in analysis by the program server.

This server supports basic ZI functionality. It analyzes the data stored in a case and translates it to generate alerts and conclusions about the ZI.

As a result, 3 levels of construction are often distinguished in a SIEM system: the data collection level; the data control level; and the data study level (Zhang et al., 2019).

At the first level, data is collected from several types of sources. These include file servers, database servers, Windows servers, firewalls, workstations, intrusion prevention systems (IPS), antivirus programs, and so on.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The second level manages security event data stored in the repository.

Data stored in the repository is retrieved in response to queries from data analysis models.

The results of information processing in the SIEM system, obtained at the third level, are reports in predefined and arbitrary forms, operational (online) correlation of event data, as well as alerts generated online and/or transmitted via email. (Gorelik et al., 2020)

To improve analysis and visibility, and to accelerate response to increasingly complex threats, SIEM systems must be transformed into a full-fledged security platform. The next step in the evolution of security information and event management should be to ensure the following four conditions are met.

Comprehensive view

Security analysis platforms must support full replay of any activity so that security operations center analysts have access to all the information they need to determine the best way to address potential issues.

Malware detection – threats are becoming increasingly difficult to identify as they disguise themselves as legitimate software in network traffic. Collecting full network packets allows for file reconstruction and automation of most of the analytical processes necessary for detecting signs of malware.

Tracking attacker activity within the environment - Network packet capture is becoming a key method for tracking an attacker's movements within an organization's network.Providing evidence of malicious activity – Systems that support full network packet capture can record entire sessions to demonstrate all attacker actions related to the acquisition of sensitive data. Adding network packet capture and session replay capabilities to new-generation security information and event management systems is key to threat investigation and prioritization. For example, traditional SIEM tools may inform you that your computer has detected communication with a suspicious server, but you won't know the exact data exchanged. Packet capture and session replay, combined with information from logs and other sources, provides security professionals with a more detailed analysis of the detection and assesses its significance. Detailed investigation capabilities help security operations center staff analyze suspicious activity step by step and mitigate the impact of advanced threats. (Khraisat et al., 2019)

Deeper analytics and faster investigations

Security analysis systems must have the means to examine disparate data and identify signs of advanced threats. For example, they must search for behavior patterns and risk factors, not just static rules and known signatures. Security analysis systems must also consider the relative value of information assets at risk, flagging the most critical ones.

When determining the risk level of big data, security analytics platforms can exclude known trusted actions, thereby increasing accuracy by reducing the volume of information security professionals must analyze for new threats. In-depth automated analysis provides events of interest with a frequency profile. Thus, security analytics

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

systems triage events for analysts, highlighting those that require more detailed investigation. (Aldwairi et al., 2020)

As fully automated components of new security platforms, these tools cannot replace human expertise and decision-making skills, but rather merely draw specialists' attention to issues requiring careful consideration. Security analysis systems are designed to help security operations centers expand their threat detection capabilities in ways previously unavailable. This allows analysts to promptly investigate incidents and compare the impact of increasingly complex threats (Ring et. al ., 2019)

Security information and event management systems, when transformed into security analytics platforms, must be scalable in scale and scope to handle massive volumes of heterogeneous data both within and outside the organization. In-depth traffic analysis from various devices across the network significantly increases the volume of data the platform must process. Adding advanced threat investigation tools from external sources transforms the security console into a security data analytics hub, which must also meet scalability requirements.

To address modern threats, security analytics platforms must support features such as a distributed N-tier storage architecture and analytics engine that normalizes and processes large, distributed data sets at high speed. The analytics platform must scale in parallel with the storage system. (Gupta et al., 2021)

To stay informed and view events in context, security professionals must always have access to all information relevant to system security. In addition to collecting data from within their network, security analysis platforms must also automatically integrate up-to-date threat intelligence from third-party researchers, government agencies, industry associations, and open-source investigations. By providing all the necessary information, the platform frees analysts from the need to manually collect data and saves their time. Centralizing all available investigative tools within a unified analysis platform provides analysts with an up-to-date picture of the IT environment, allows them to view events in context, and accelerates decision-making.

Models of SIEM Systems

A SIEM system incorporates the functions of two third-party systems related to information security management systems-SIM and SEM. Based on this, a SIEM system implements functions that are understandable for both SIM and SEM systems. The basic set of rules for a SIM system is the collection, storage, and analysis of all logged data, as well as the generation of reports. The fundamental basis for SEM systems is the online monitoring of security events, as well as the response and reporting of ongoing incidents **(**Kurmangaliyeva et al., 2023).

The implementation of these functions in the SIEM system under consideration is possible thanks to a complex set of various operational mechanisms. In Type I SIEM systems, such mechanisms include normalization, filtering, classifying, aggregation, mutual intersection and division of event importance, as well as the creation of alerts and reports [5]. Modern SIEM systems also include analysis of events occurring during incidents and their consequences, as well as a solution implementation mechanism and

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

visual presentation.

Let us describe the core principles of a SIEM system's operation. Normalization involves converting log record formats collected from various sources into a single base format that can be used for storage and further analysis. Filtering all transferred events involves removing large events from incoming system streams. Segmentation allows security event attributes to be expressed by their role in certain classes. Aggregation consists entirely of events that are similar in a few properties. Mutual intersection reflects the relationships between similar events, which helps us understand the nature of attacks on critical infrastructure, as well as adjust information security criteria and policies. The priority mechanism reflects the importance and criticality of security events within the rules available in the system **(**Dyusembaev et al., 2017)**.**

Understanding events, incidents, and their resulting values involves constructing a model of events, attacks, and their outcomes, studying system availability and security, expressing attacker metrics, risk assessment procedures, and predicting events and situations. Reporting and forecasting reflect the generation, printing, and dissemination of work results. Implementing solutions involves identifying measures to adjust security methods to mitigate attacks or create a completely secure infrastructure. The visual component includes data display in graph form, which helps describe the analysis of security events and the level of security of the entire maintained CVI system and its individual components.

It should be noted that when moving to higher-level mechanisms of the model shown in Fig. 2, the number of events processed decreases, and the complexity of their processing increases.

The interrelationship of the operating mechanisms of the new generation SIEM system is clearly demonstrated by the functional model presented in Fig. 2.

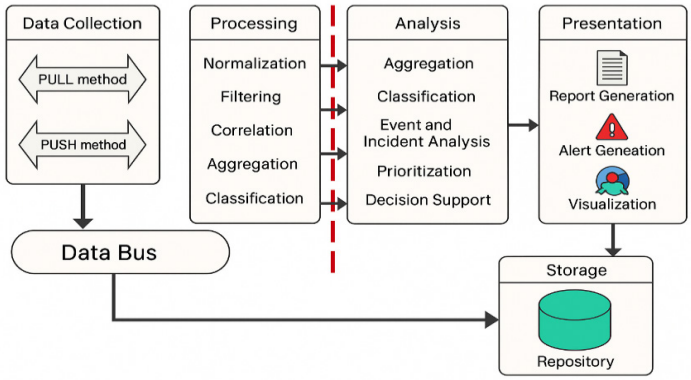


Fig. 2. Interrelation of the SIEM system functioning mechanisms.

As can be seen from Fig. 2, the SIEM system can be divided into five main functional subsystems: (1) data collection; (2) processing; (3) storage; (4) analysis; (5) pre-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

sentation. The first two operate in online mode, while the others operate in near-online mode. Let us briefly describe these subsystems.

Two main methods are used to obtain information from sources: Push and Pull. With Push, the source sends its log data to the SIEM system. With Pull, the system automatically retrieves the log data.

Data is collected from diverse types of sources.

Processing subsystem. Information processing includes normalization, filtering, correlation, aggregation, and classification.

Storage subsystem. Filtered data in normalized form is stored in a repository. The repository can be built on a relational DBMS (the most common solution), an XML-based DBMS, and/or a triplet store. A triplet store is a specially designed database optimized for storing and retrieving triplets, i.e., statements of the "subject-predicate-object" type.

Analysis subsystem. Data analysis includes the following functions: data correlation, classification, aggregation, prioritization, and analysis of events, incidents, and their consequences (including through event modeling, attacks, and their consequences, vulnerability and system security analysis, intruder characterization, risk assessment, and event and incident forecasting), as well as decision support. Data analysis can be based on qualitative and quantitative assessments. Quantitative assessment is more accurate but significantly more time-consuming, which is not always acceptable. Most often, a quick qualitative analysis is sufficient, the purpose of which is to categorize risk factors. The scale of qualitative analysis may vary across assessment methods, but the goal is to identify the most serious threats. (Ring et al., 2019)

Presentation subsystem. The presentation includes several functions: visualization, report generation, and alert generation **(**Abubakirov et al., 2022).

When studying various systems-social, economic, natural, and man-made-it's important to consider the totality of external factors to account for all factors within the system as a single entity. However, the connections between these components cannot often be considered due to a lack of sufficient data, and for some tasks, such as forecasting and simulation, data may be completely absent. In such cases, the necessary connections are created through expert assessments. These are often implemented as weighting parameters used to numerically evaluate the contribution of a given factor to the result. **(**Saaty et al., 1980**)**

Accounting for weight parameters. Various approaches are used for this accounting, and many methodologies have already been implemented within them. Since there is no goal to provide a general description of the various methods used to express weight coefficients, only an analysis of the main approaches was performed.

Direct weighting. Experts refine factor weights based on requirements, such as the sum of the weights being between 1% and 100%, although another constant can often be used if it's convenient for further calculations. This process is often confusing factors specific values on an explicit numerical scale, but in this case, it's better to call such factors "significance indicators" rather than "weights," as they are then assessed

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

comparatively rather than by their overall impact. Weighting coefficients are also implemented in an analogous manner, but that's where we'll end.

The difficulty of this approach lies in the ability to implicitly contain all factors within a separate framework, since by assigning a numerical value to any factor, the expert must also correlate it with the others. The complexity increases progressively as the number of factors increases.

There are also technical difficulties in the specialist's work related to the importance of periodically monitoring the current sum of weighting factors to avoid increasing the specified constant or transferring the remaining substantial portion to extreme factors. If this occurs, it is customary to recalculate all sent coefficients, which can be done several times during the exchange process. The number of operations increases as the number of factors increases.

Factor ranking. This approach simplifies the experts' work, as it eliminates the need to control the total sum of the coefficients. In this case, experts are required to rank, i.e., the factors under consideration that form the object according to the degree of their properties' identification, in order of their minimization or enhancement.

$$\left.\begin{matrix} R_{11},\ R_{21},\ \dots, R_{i1} \\ R_{12},\ R_{22},\ \dots, R_{i2} \\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots \\ R_{1j},\ R_{2j},\ \dots, R_{ij} \end{matrix}\right\}, \quad (1)$$

Where Rij $_{\text{is}}$ the rank (place) assigned to factor Oij $_{\text{by the jth}}$ expert in a series of n-studied objects, based on the degree of expression of the analyzed property. Two or more factors may have the same rank, but then the rank is a fraction. The summary estimates of the weighting coefficients are obtained by averaging the partial ranks across the columns.

The advantage of this method lies in its simplicity, but this simplicity isn't always beneficial, as averaging the ranks results in rougher weighting estimates than other methods. It also doesn't relieve the expert of the responsibility of controlling all factors, as with direct ranking.

Transferring coefficients to factors. This method asks experts to rate factors on a scale, for example, from 1 to 10. The result is:

$$\left.\begin{matrix} y_{11},\ y_{21},\ \cdots, y_{i1} \\ y_{12},\ y_{22},\ \cdots, y_{i2} \\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots \\ y_{1j},\ y_{2j},\ \cdots, y_{ij} \end{matrix}\right\}, \quad (2)$$

where y $_{ij}$ is the factor score transmitted from the j-th expert, n is the sum of fac-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

tors, m is the number of experts.

Summary estimates of the weighting coefficients are often found by selecting an appropriate regression model. The average estimate w_i of the factor weighting coefficients is obtained using trivial formulas:

$$w_i = \frac{\sum_{j=1}^{m} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}}, \quad (3)$$

where w_{ij} is the weight of the i-th object, based on the assessments of all experts;

$$w_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \quad (4)$$

where x_{ij} is the assessment of factor i given by expert j;

n is the number of factors, m is the number of experts.

This method weakens the dependence of the assessment of an individual factor on the others, but does not eliminate it, since it is necessary to compare the factors, otherwise it will not be possible to correctly assign the significance coefficients.

The Analytical Hierarchy Processing (AHP) method, created by T. Saaty in the 1980s, was designed to partially minimize the above-mentioned difficulties. The essence of the method is as follows. (Saaty at. al., 1993; Saaty at. al., 1980)

Factors are considered in pairs relative to each other based on their impact on the final goal. The influence of other factors is not considered. For pairwise comparison of factors, the Saaty method uses a special rating scale, including five main and four intermediate judgments (Saaty, T., 1980).

In it, the experts' arguments were highlighted as follows (Table 1):

Table 1 – Specifics of expert comparisons of the ratio of factors.

Judgment	Explanation
1. Equal importance	Equal contribution of factors to the final goal
2. ...	Additional expression
3. A slight advantage	Judgment and experience give slight superiority to one factor over the others
4. ...	Additional expression
5. Tangible superiority	Sensitive dominance of one factor over the others
6. ...	Additional expression
7. Increased superiority	There is a significant predominance of one factor over the others
8. ...	Additional expression
9. Supreme Excellence	There is a confident superiority of one factor over the others.

The results of such pairwise comparisons are presented as a square matrix $A = (a_{ij})$ with a diagonal equal to 1 (comparing a factor to itself equals 1). Here, "a" becomes the ratio of the ratings of specific elements; the indices i and j range from one to a value equal to the sum of the factors. Since, when sequentially searching through all available pairs, the factors are related to each other twice (a_i with a_j, then

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

vice versa), the “reverse symmetry” condition must be true when preparing the matrix: . It follows that it is sufficient to fill only one part of the matrix—the one located above or below the diagonal—which has no specific significance due to the simple recalculation of mutually inverse parameters. If n factors are studied, then a total of $a_{ji} = \frac{1}{a_{ij}} 100$ meaningful combinations will be available. $\frac{n^2-n}{2}$

In MAI, the number of a specific row of Table 2 is used for coding. Any of the specified judgments is coded in the range of numbers from 1/9 to 9.

Weights are calculated in several ways. One available method for approximating the weight vector is to calculate a separate vector of the pairwise comparison matrix, usually corresponding to the larger eigenvalue. Such algorithms for obtaining eigenvectors have been thoroughly studied, and their descriptions can be found either in monographs or in other literature.

The MAI method has its own parameters for expressing the quality of expert performance—the consistency index (CI), which provides data on the level of violation of the numerical and ordinal consistency of expert judgments. Cardinality control involves considering specific numerical characteristics, deviations from which indicate errors in the process of conveying expert judgments. Therefore, if separate rules for coding expert judgments are created, for example, from 0 to 1, then expert judgments cannot deviate from the value sets specified in these rules, i.e., be greater than one or negative. The ordering helps understand the logic of the expert’s reasoning. If an expert believes that factor A is better than factor B, and factor B, in turn, is better than factor C, then in a paired comparison, factor C cannot be better than factor A, i.e., the inequality A > B > C is satisfied. Inconsistency is a significant limiting factor for studying individual problems.

The IS is calculated as follows: together with the pairwise comparison matrix, there is a measure of the degree of deviation from the desired value. The IS in each matrix for each hierarchy is estimated using the formula:

$$\text{ИС} = \frac{\lambda - n}{n-1}, \tag{5}$$

where $\lambda$is the eigenvalue,

n– the number of factors being compared.

The IS is compared with the value obtained from a random selection of quantitative variables, which is treated as the average. The average consistency (MC) for random matrices of different orders is given in Table 2, where *n* is the number of factors.

Table 2 – Average consistency (MC) for random matrices of different orders

n	1	2	3	4	5	6	7	8	9	10
SS	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

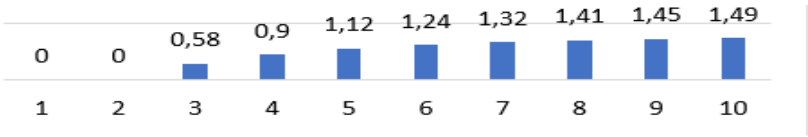


Fig. 3. Average consistency (MC) for random matrices of different orders

If we divide the IS by the SS for a matrix of the same order, we obtain the consistency ratio (CR):

$$\text{ОС} = \frac{\text{ИС}}{\text{СС}} * 100\%, \quad (6)$$

It seems that MAI is the optimal solution for solving a variety of problems where expert analysis methods are used as key ones. This is true, and we will outline the main reasons for this.

Pairwise comparisons. Pairwise comparisons of things can also be found in human nature. Minimizing the need to always consider all factors, or, for example, some of them, allows the expert to focus more on a specific issue: how factor Aj_{is} ahead of factor Bj_{or} behind it. This allows for more accurate results.

Complementarity of the initial matrix. In the practice of system analysis, situations often arise where the number of explicit factors is adjusted. This is due to the periodicity of natural processes, as well as the adjustment of socioeconomic factors. This requires adding, subtracting, or replacing one factor with another. In the context of the MAI, this necessitates comparing the created pairs or subtracting the rows and columns of the pairwise comparison matrix of factors previously excluded from the analysis, i.e., implementing a matrix minor. All results from previous surveys are reflected, and updating the entire questionnaire, as is the case in other approaches, is not required. Since the MAI procedure often leads to a search for the desired matrix's vector, which corresponds to the largest eigenvalue, from the technical implementation perspective, the inclusion of extraneous factors is considered an increase in the dimensionality of a separate linear space due to the use of extraneous terms.

Verbal-numeric scale. Classic numerical scales often fail to compare factors across different dimensions and domains. It's difficult to compare factors whose results initially yield qualitative parameters and then quantitative ones. The Harrington scale, often used, only accepts a few summary parameters, which can be adjusted within a range from 0 to 1. Verbal-numeric scales, such as the Saaty scale, are designed to assess such discrepancies in the indicators of underlying factors.

An accessible criterion for assessing the quality of a specialist's performance. After conducting an assessment, experts often require verification. Most often, various numerical parameters implemented for group and individual surveys are used for this purpose. However, the question of the best criterion remains open, and its selection is accessible. In this sense, transferring the consistency ratio parameter to the MAI offers certain advantages, especially in the implementation of an automated software system.

Disadvantages of the methodology: not all the MAI's advantages are so clear.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

There are a few issues when analyzing the results, and these are most often related to assessing the expert's accuracy—the level of consistency.

Using transitivity for qualitative parameters. It can work perfectly well when all parameters of the system being analyzed are numerical values. However, when this is not the case, transitivity often ends up in conflict with the researcher's logic.

"Reverse" logic. The expert's performance quality percentage, as well as the consistency ratio, are based on the adjustment of a clearly defined characteristic, such as mathematical expectation. Like any criterion of a stable nature, the consistency ratio is formal and often leads to interpretable results.

We will describe a solution to the problem of selecting criteria and comparing sensor parameters for a SIEM system using the Analytical Hierarchy Processing (AHP) methodology. While comparative analysis has been implemented in various projects, the study of the criteria itself has been less common.

At the top level of the hierarchy is the goal of selecting the optimal comparison of sensor parameters for a SIEM system. Below these are the selection criteria. These criteria are considered unequal. Below these criteria are the studied methods for determining and comparing sensor parameters for a SIEM system.

The second step involves assigning importance weights to the Sij criteria. This is accomplished by testing all possible pairwise comparisons of parameters on a qualitative scale and analyzing the resulting pairwise comparison matrix.

In the third step, the priorities of the investment project selection and comparison methods $C_{\text{ij are determined}}$ in relation to each of the nine criteria. To do this, the expert performs all possible pairwise comparisons on a qualitative scale. For each criterion $K_{t,}$ $_{a\ vector}$ of weighting coefficients $S(K_t) = \{S_i(K_t)\}, i = \overline{1,9}$ is generated by processing the pairwise comparison matrix .

By combining the vectors of weighting coefficients for each of the criteria, we obtain a complete matrix of priorities for selecting methods for selecting and comparing investment projects with dimensions of *9 x 11*.

At the fourth step, the final vector $\mathrm{w} = (\mathrm{w}_1, \ldots\ldots, \mathrm{w}_{11})$ of priorities for methods of selecting and comparing investment projects is determined.

Let's consider a three-level hierarchy diagram (Goal - Criteria - Alternatives). In more complex cases, a diagram with a larger number of levels can be considered.

For the mathematical formulation, we introduce the following sets into consideration:

1. $K = \{k_1, k_2, \ldots, k_n\}$- a set of criteria (or requirements for the tasks of selecting information security tools), $N = \{1, 2, \ldots, n\}$- a set of criteria indices.
2. $A = \{a_1, a_2, \ldots, a_m\}$- a set of alternatives (for the problems of selecting information security tools, an alternative is one information security tool), $M = \{1, 2, \ldots, m\}$– a set of indices of alternatives, respectively.

The following parameters are specified for the elements of these sets:

1. $v_i^{(k)}, \forall i \in N$–– the "weights" or "importance" of criteria from the point of view of achieving the goal are determined by experts; a standardization condition is

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

imposed on these "weights": $\sum_{i\in N} v_i^{(k)} = 1.$.

2. $v_{ij}^{(a)}, \forall i \in N, j \in M$– the "weight" ("importance") *of the j* -th alternative for achieving *the i* -th criterion. These "weights" are also subject to normalization conditions of the form: $\sum_{j\in M} v_{ij}^{(a)} = 1, \forall i \in N.$

Then the global priority of the j-th alternative for achieving the goal is calculated as follows:

$$F_j = \sum_{i\in N} v_i^{(k)} v_{ij}^{(a)}, \forall j \in M \quad (7)$$

The formulation of the problem of choosing an alternative with the maximum global priority has the form:

$$F_j = \sum_{i\in N} v_i^{(k)} v_{ij}^{(a)} \rightarrow \max_{j\in M}. \quad (8)$$

Let us consider a solution to the problem of multi-criteria selection of sensor performance criteria for a SIEM system using the Analytic Hierarchy Process (AHP). While the problem of comparative analysis has been addressed in numerous studies, insufficient research has been conducted on the specific criteria.

We will use the following criteria for selecting the optimal sensor parameter for the SIEM system:

- K1 = «System load (OS)»;
- K2 = «Reaction time»;
- K3 = «Working time»;
- K4 = "Efficiency", that is, the effectiveness of the protective measures used in the situation under consideration;
- K5 = «Cost of sale»;
- K6 = "Labor intensity of implementation";
- K7 = «Universality»;
- K8 = "Quality of implementation";
- K9 = «Prevalence».

The diagram for selecting the optimal sensor parameter is shown in Figure 4.

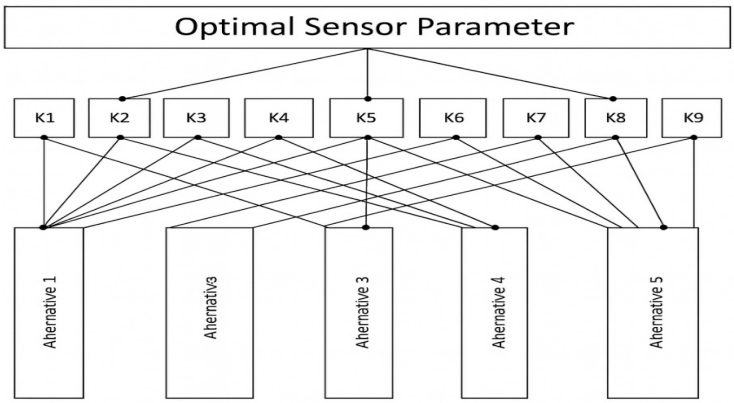


Fig. 4. Scheme for selecting the optimal sensor parameter

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The first level of the hierarchy presents the goal—selecting the optimal sensor parameter for the SIEM system. The second level of the hierarchy presents nine selection criteria. These criteria are not equivalent. The third level of the hierarchy presents the protection methods being investigated.

In the second step, the importance weights Sij $_{of}$ the criteria are determined. This is accomplished by performing all possible pairwise comparisons of the criteria on a qualitative scale and processing the resulting pairwise comparison matrix.

In the third step, the priorities of protection methods C $_{ij \text{ are determined}}$ in relation to each of the nine criteria. To do this, the expert performs all possible pairwise comparisons on a qualitative scale. For each criterion $K_{t,\ a\ vector}$ of weighting coefficients $S(K_t) = \{S_i(K_t)\}, i = \overline{1,9}$ is formed by processing the pairwise comparison matrix.

By combining the vectors of weighting coefficients for each of the criteria, we obtain a complete matrix of priorities for selecting the optimal sensor parameter for a SIEM system with dimensions of 9 x 11.

In the fourth step, the final vector $w = (w_1, \ldots, w_{11})$ of priorities for the sensor operating parameters for the SIEM system is determined.

Development of an Algorithm for Selecting the Optimal Sensor Parameter.

The purpose of software development is to implement a method for selecting the optimal sensor parameter.

The software should allow one to determine the level of selection of the optimal sensor parameter, test the methods used, and view the results of the selection of methods in comparison with others.

Finding the optimal method should be done using multi-criteria choice.

The software should allow determining the level of compliance of the selected method for selecting the optimal sensor parameter with the accepted level of acceptability using various methods for increasing reliability.

The algorithm diagram for solving the problem is shown in Figure 5.

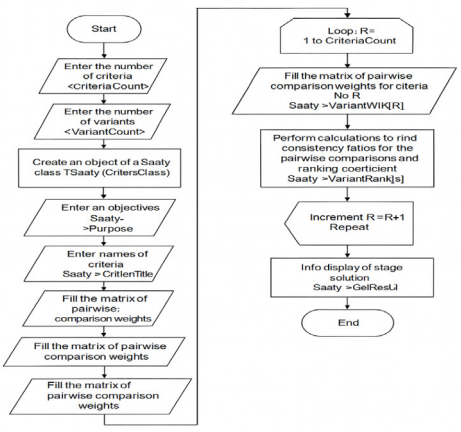


Fig. 5. Scheme of the algorithm for selecting the optimal sensor parameter method

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Figure 6 shows the structural and functional diagram of the sensor, consisting of three hosts, a sensor, a SIEM system, and a screen for displaying processed events. The sensor collects events from the hosts and transmits them to the SIEM system for subsequent processing. After correlation, the data and events are displayed on the system administrator's (or the person responsible for the system's) monitor.

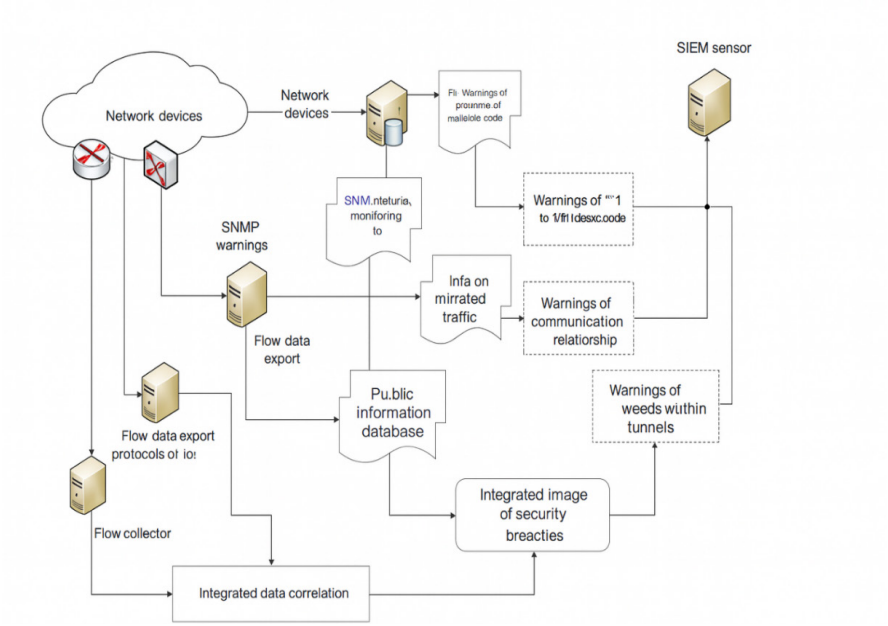


Fig. 6. Structural and functional diagram of the AlienVault sensor operating algorithm

The AlienVault SIEM system sensor contains the following components:
Event Collection, Analysis and Correlation (SIEM);
Host-based intrusion detection system (HIDS) – OSSEC;
Network Intrusion Detection System (NIDS) – Suricata;
Wireless Intrusion Detection System (WIDS) – Kismet
Network Node Monitoring – Nagios
network anomaly analysis – PADS, Arpwatch, etc.;
vulnerability scanner - OpenVAS;
the most powerful system for exchanging information about threats between OSSIM users - OTX;

More than 200 plugins for parsing and correlating logs from various external devices and services.

Results and Discussion.

The study developed a mathematical model for selecting the optimal sensor for SIEM information security management systems, based on the Analytical Hierarchy Process (AHP). The model's goal is to provide an objective, quantitatively substantiated procedure for selecting a sensor, considering multiple criteria that reflect both the technical and economic aspects of the system's operation.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The use of the analytic hierarchy process allowed us to structure the problem as a three-level hierarchy: the first level contains the main objective—identifying the optimal sensor; the second level contains the evaluation criteria; and the third level contains a set of alternative sensors. This approach ensures transparency of the decision-making process and allows for a quantitative assessment of the contribution of each factor to the final choice.

Based on the analysis of the functional features of sensors and technical requirements of SIEM systems, key criteria were identified:

- performance - the ability to process a given volume of events per unit of time;
- response time - the delay between the registration of an event and its transmission to the correlation system;
- versatility - the ability to work with diverse types of data sources and protocols;
- reliability - resistance to failures and the ability to recover;
- cost - the total cost of acquisition, setup and operation;
- compatibility – the ability to integrate with existing SIEM platforms and network infrastructure;
- Event processing accuracy is the proportion of correctly recognized incidents without false positives.

Each criterion was presented as a pair for pairwise comparison with other criteria on the nine-point Saaty scale, which allowed us to create a matrix of relative priorities. Based on expert assessments, eigenvalues and priority vectors were calculated, determining the relative weights of the criteria. To verify the accuracy of the judgments, the consistency index (CI) and consistency ratio (CR) were used, yielding values less than 0.1. This demonstrates the logical consistency of the expert procedure and the reliability of the resulting weighting coefficients.

The calculation results showed that the most significant criteria were performance (0.28) and response time (0.22), reflecting the priority of promptly processing security events in today's environment. Reliability (0.17) and accuracy (0.15) were also significant, as they directly impact the level of security of the information infrastructure. Cost (0.09) and versatility (0.07) were found to be less significant, reflecting their secondary influence in the design of critical security systems.

The next step involved evaluating three alternative sensors, each differing in architecture, performance, and cost. Local priorities were calculated for each sensor based on all criteria, followed by global priorities—summary values reflecting the overall effectiveness of each alternative. The ranking results showed that sensor #2 offered the best balance between performance and response speed with moderate operating costs, thereby providing an optimal balance of technical and economic parameters. Sensor #1 demonstrated high reliability, but its cost was 25 % higher than average. Sensor #3, conversely, had a low response time but was inferior in data transmission accuracy and stability.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

To implement the proposed model, software was developed that implements the full analysis cycle within the Analytic Hierarchy Analysis method. The program is written in C++ using the MySQL DBMS, ensuring high processing speed and storing results in the database. The interface includes functions for entering criteria, adding alternatives, automatically generating pairwise comparison matrices, calculating weighting factors, a consistency index, and ranking options by preference. The software module also supports exporting results to a tabular format, allowing the obtained data to be used in configuring real-world SIEM systems.

The results of computational experiments confirmed the validity and effectiveness of the proposed approach. A comparison of the expert sensor selection results with those obtained using the developed model revealed a 90–95% match, demonstrating a high degree of adequacy of the constructed mathematical model. The use of the MAI eliminated the subjectivity inherent in traditional sensor selection methods and enabled multivariate analysis without loss of clarity.

Furthermore, the proposed methodology was compared with alternative approaches, such as linear ranking and weighted sums. It was found that the analytic hierarchy process offers greater flexibility and allows for consideration of not only quantitative but also qualitative parameters. The AHP also ensures consistency checks between expert assessments, which is particularly important in the face of uncertainty and incomplete source information.

Experiments have shown that implementing the developed method within a SIEM system reduces sensor subsystem setup time by 30–35 % and improves security event registration accuracy by 15–20 % compared to traditional manual equipment selection. This confirms the practical significance of the proposed solution and its applicability in the design and operation of integrated information security monitoring systems.

An additional advantage of the proposed model is its scalability**.** If necessary, the number of criteria and alternatives can be expanded without changing the algorithm structure, allowing the methodology to be adapted to the specific requirements of specific organizations and industries. Plans include integrating the developed software module with existing SIEM platforms and security incident response (SOAR) systems, creating the basis for automated, intelligent sensor selection in real time.

Thus, the conducted research and experiments confirm that using the Analytic Hierarchy Process for sensor selection in SIEM systems is an effective tool for improving the objectivity, accuracy, and transparency of information security decision-making processes**.** The obtained results have both theoretical and practical significance, contributing to the development of multi-criteria optimization methods in cybersecurity.

Conclusion.

The conducted research addressed the problem of rational sensor selection in Security Information and Event Management (SIEM) systems under conditions of multi-criteria evaluation and uncertainty. The increasing complexity of information infrastructures and the growing scale of cyber threats require not only advanced monitoring technologies but also scientifically grounded approaches to configuring system

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

components. Within this context, the Analytic Hierarchy Process (AHP) was applied as a formal decision-support tool to ensure transparency, consistency, and quantitative justification of sensor selection.

A structured three-level hierarchical model was developed, including the overall objective, a system of weighted evaluation criteria, and alternative sensor configurations. The study incorporated technical, operational, and economic parameters such as system load, reaction time, efficiency, implementation cost, labor intensity, universality, quality of implementation, and prevalence. The use of pairwise comparisons based on the Saaty scale enabled the transformation of qualitative expert knowledge into measurable priority vectors. The calculation of eigenvalues, consistency indices, and consistency ratios confirmed the logical coherence of expert judgments and validated the reliability of the decision-making process.

The research demonstrated that performance-related criteria, particularly processing capacity and response time, have the highest impact on overall system effectiveness, reflecting the critical importance of timely threat detection in modern cybersecurity environments. At the same time, economic and implementation factors were shown to influence the final ranking of alternatives, emphasizing the need for a balanced approach that integrates both technical and managerial considerations.

A software solution was developed in C++ using the MySQL DBMS to automate the entire evaluation cycle. The system supports matrix generation, weight calculation, consistency verification, and ranking of alternatives. Experimental results confirmed that the implementation of the proposed methodology reduces configuration time, improves decision transparency, and minimizes subjectivity compared to traditional expert-based selection methods. The comparison between expert conclusions and model outputs demonstrated a high degree of correlation, confirming the adequacy of the developed mathematical framework.

The proposed model is scalable and flexible, allowing expansion of criteria sets and inclusion of new sensor alternatives without structural modification of the algorithm. This makes it applicable not only to SIEM sensor selection but also to broader cybersecurity component evaluation tasks. Future research directions may include the integration of machine learning techniques for adaptive weighting, incorporation of dynamic risk assessment mechanisms, and real-time data analytics modules, thereby contributing to the development of intelligent and self-optimizing cybersecurity management systems.

Overall, the study provides both theoretical and practical contributions to multi-criteria decision-making in cybersecurity, offering a systematic and reproducible approach to improving the effectiveness and reliability of SIEM infrastructures.

REFERENCES

Aldwairi, M., Khan, A., Al-Yaseen, W. (2020). Anomaly-Based Intrusion Detection Using Deep Learning Techniques // Computers & Security // Elsevier. Vol. 96. // Article 101906. 10.1016/j.cose.2020.101906 [In Eng.].

Abubakirov, A., Nurgaliyev, M. (2022). Methods of detecting anomalies in information security systems // *KazNU Bulletin*. Series: Mathematics, Mechanics, Informatics. Al-Farabi KazNU [In Eng.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Gupta, B. B., Quamar, A., Rao, S. (2021). Security analytics for SIEM systems using machine learning // IEEE Access. //IEEE. Vol. 9. Pp. 82105–82118. 10.1109/ACCESS.2021.3059387 [In Eng.].

Gorelik, A. (2020). The Analytic Hierarchy Process and Its Applications // Springer. 10.1007/978-3-030-40230-0 [In Eng.].

Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J. (2019). Survey of Intrusion Detection Systems: Techniques and Challenges // *Journal of Network and Computer Applications* // Elsevier. Vol. 155. Article 102626. 10.1016/j.jnca.2019.102626 [In Eng.].

Kurmangaliyeva, S., Shaimerdenova, A. (2023). Evaluation of SIEM sensors based on multi-criteria decision-making // *Journal of Information Security Studies*. [In Eng.].

Ring, M., Wunderlich, S., Grudl, D., Bischl, B. (2019). A Survey of Network-Based Intrusion Detection Data Sets. Computers & Security // Elsevier. Vol. 86. Pp. 147-167. 10.1016/j.cose.2019.06.005 [In Eng.].

Saaty, T.L. (1980). The Analytic Hierarchy Process // McGraw-Hill [In Eng.].

Saaty, T.L. (2008). Decision making with the analytic hierarchy process // *International Journal of Services Sciences.* Inderscience. Vol. 1. Pp. 83–98. 10.1504/IJSSCI.2008.017590 [In Eng.].

Saaty, T. (1993). Decision Making. The Analytic Hierarchy Process. ― Moscow: Radio i Svyaz [In Russ.].

Zhang, Y., Li, J., Wang, X. (2019). Deep Learning-Based Intrusion Detection for Network Security // IEEE Access. IEEE- Vol. 7. Pp. 119977–119988. 10.1109/ACCESS.2019.2934567 [In Eng.].

Zhumabekov, D.M. (2021). Analysis of Information Security Monitoring Systems // *Bulletin of Abai University*. Series “Informatics”. [In Russ.].

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Собственник:
АО «Международный университет информационных технологий» (Казахстан, Алматы)

Главный редактор:
Колесникова Катерина Викторовна

Ответственный редактор:
Мрзабаева Раушан Жалиевна

Компьютерная верстка:
Калабай Замзагуль Ертугановна

Сайт журнала: https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Подписано в печать 30.03.2026.
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание АО "Международного университета информационных технологий"