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. 2 (26). 2026
April–June

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

2020 жылдан бері шығарылады
Том 7. 2 (26). 2026
Сәуір-Маусым

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Издается с 2020 г.
Том 7. 2 (26). 2026
Апрель-Июнь

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 SOCIO-ECONOMIC SYSTEMS

D. Abzhanova, A. Biloshchytskyi
A MODEL AND METHOD FOR MANAGING DATA ON EMISSIONS FROM STATIONARY SOURCES OF POLLUTION IN AN INTELLIGENT ENVIRONMENTAL MONITORING SYSTEM 11
A. Slanbekova, M. Rakhimzhanova, A. Zhanibekova, A. Alimagambetova, M. Xudoyberganov
EARLY DETECTION OF HYDROLOGICAL HAZARDS BASED ON SPATIOTEMPORAL ANALYSIS 27
A. Sakhipov, R. Seiitbek, D. Yedilkhan
REAL-TIME PASSENGER FLOW PREDICTION USING AUTOMATED DATA PREPARATION 38
K. Tokayev, A. Moldagulova
A REVIEW OF PREDICTIVE MODELS APPROACHES TO STOCK MARKET FORECASTING IN KAZAKHSTAN 60

INFORMATION TECHNOLOGY

F.N. Abdraimova, A.A. Kereibayeva, D.S. Dyussenova, D.A. Aliyeva, T.Zh. Toktarova
AI TECHNOLOGIES IN LANGUAGE EDUCATION: PRACTICAL ASPECTS AND CHALLENGES OF STUDENT USAGE 83
G. Azieva, M. Yessenova, A. Abzhapparova, G. Abdikerimova, P. Schmidt
HYBRID STACKING FRAMEWORK FOR CROP CLASSIFICATION USING UAV DATA 97
A.K. Aitim
JOINT MORPHOLOGICAL DISAMBIGUATION AND POS TAGGING FOR AGGLUTINATIVE LANGUAGES 109
A. Auyezova, Aigerim Aitim, B. Sinchev, W. Wójcik
NP-COMPLETE ALGORITHMS IN LEGALTECH: HYBRID ARCHITECTURE FOR SEMANTIC RETRIEVAL IN KAZAKHSTAN'S LEGAL ACTS 125
Y. Hozak, S. Paliy, T. Honcharenko
EVALUATING OPTIMAL CONFIGURATIONS OF SHALLOW CNNS ON LOW-COMPLEXITY DATASETS 138
S.A. Yesniyazova, S.T. Kaimov
PREDICTIVE MAINTENANCE OF HEAVY-DUTY TRUCKS USING EXPLAINABLE MACHINE LEARNING 153
T. Imanbekova, Zh. Ibrayeva, G. Jakanova, G. Askanbay
DATA COMPRESSION ALGORITHM BASED ON WAVELET TRANSFORMER; ANALYSIS AND IMPLEMENTATION IN MATLAB 167
B.Z. Kenzhegulov, Zh.T. Bilyalova, K.N. Uteuliyeva, L. Nurgaliyeva, Sh.S. Nurzhanova
A MATHEMATICAL AND ALGORITHMIC APPROACH TO THE DEVELOPMENT OF AN INTELLIGENT TEXT-TO-SQL SYSTEM BASED ON LARGE LANGUAGE MODELS 185
M. Kabdullin, A. Kairbekov, A. Kabdullin, L. Naizabayeva, A. Mukhit
DESIGN AND IMPLEMENTATION OF A WEB-BASED PROTOTYPE FOR CARDIOLOGICAL IMAGE MANAGEMENT 206
L.Kurmangaziyeva, Sh. Yelezhanova, S. Iskakova, Sh. Kodanova, H.İ. Bülbül
APPROACHES TO BUILDING A HYBRID INTELLIGENT DECISION SUPPORT SYSTEM BASED ON FUZZY LOGIC, XAI, AND LARGE LANGUAGE MODELS 219
Z. Momynkulov, B. Omarov
DDPG FOR TRAJECTORY GENERATION 236
F. Myrzabekov, N. Mustafin, M. Mukushev
WHY IS THIS NEWS NEGATIVE?: AN ATTENTION-BASED VISUALIZATION TECHNIQUE FOR EXPLAINABLE SENTIMENT ANALYSIS. AN ANALYTICAL SURVEY AND CONCEPTUAL FRAMEWORK 248
N.Sh. Maxutova, J.A. Tussupov, A.A. Shekerbek, Zh.E. Kenzhebayeva, Q.O. Rakhimov
MACHINE LEARNING FOR COMPREHENSIVE EVALUATION OF CARDIOVASCULAR DISEASE RISK AND BIOCHEMICAL ALTERATIONS: FOCUS ON ASPARTATE AMINOTRANSFERASE 260
L. Naizabayeva, G. Turken, R.A. Druba, A.Sh. Yawar
LEVERAGING ENTERPRISE DATA WAREHOUSING TO IMPROVE WEB-BASED SALES

OPERATIONS 275
S. Nurgaliyeva, N. Naiman, K. Makhmetova, M. Kali, N. Zhakiyev
COMPARATIVE ANALYSIS OF MACHINE LEARNING METHODS FOR RAPID DETECTION OF VEGETABLE DISEASE USING COMPUTER VISION 291
O.S. Salykova, V.A. Madin, B.R. Salykov, D.N. Komarov, N.V. Manuilov
INTEGRATION OF MEMS ACCELEROMETER SENSOR MODULES IN INDUSTRIAL MONITORING SYSTEMS 309
R. Taberkhan, M.A. Sambetbayeva, G. Kalman
KAZCAUSAL: THE FIRST CORPUS-BASED ANNOTATION OF CAUSAL RELATIONSHIPS IN THE KAZAKH LANGUAGE 323
S.Tynymbayev, S.E. Mamanova, R. Berdybayev, Zh.E. Temirbekova, T. Chinibayeva
DIVIDING DEVICES WITH PRELIMINARY PREPARATION OF MULTIPLES OF THE DIVISOR 335
K.N. Uteuliyeva, B.Z. Kenzhegulov, T.A. Karazhigitova, H.İ. Bülbül, Z.Zh. Zhanuzakova
MATHEMATICAL AND ALGORITHMIC APPROACHES TO THE DEVELOPMENT OF A COLLABORATIVE FILTERING-BASED RECOMMENDER SYSTEM 351
S. Sharmukhanbet, G. Turmukhanova, O. Findik, V. Makhatova, L. Kurmangaziyeva
HIGH-PRECISION ROBOTIC ASSEMBLY UNDER VARIABLE ILLUMINATION: A ROBUST MECHATRONIC ARCHITECTURE FOR VISUAL SERVOING 372

INFORMATION SECURITY AND COMMUNICATIONTECHNOLOGIES

A. Amirbay, Z. Amanbaikyzy, K. Maxutova, A. Mukhanova, M. Kassim
MACHINE LEARNING ALGORITHM FOR EARLY DETECTION OF AUTISM SPECTRUM DISORDERS IN CHILDREN BASED ON MULTIMODAL ANALYSIS OF EYE MOVEMENTS AND FACIAL EXPRESSIONS 390
K. Baisylbayeva, Sh. Mussiraliyeva, Zh. Yeltay
DETECTION OF EXTREMIST IDEOLOGY IN THE KAZAKH LANGUAGE: ANNOTATION CHALLENGES AND DEEP LEARNING APPROACHES 405
M.A. Bolatbek, A.M.Usmanova, K.B. Bagitova, G.B. Baispay
DEVELOPMENT AND RESEARCH OF A METHOD FOR ANALYZING NETWORK TRAFFIC TO IDENTIFY A CYBER THREAT 424
D.I. Prokopovych-Tkachenko, N.K. Zhumagalieva, D.N. Shchytov, N.F. Mormul, D.A. Cherkaskyi
FUZZY MODEL FOR EVALUATING INFORMATION SECURITY PARAMETERS OF INFORMATION SYSTEMS UNDER INCOMPLETE AND QUALITATIVE DATA: CONSTRUCTION METHODOLOGY, RULE BASE TUNING, AND DEMONSTRATION CASE FOR ORGANIZATIONS 442
E.A. Pustovoy, O.A. Pustovaya, A.N. Raushanova, I.S. Zaurbekov
EVALUATION OF THE EFFECTIVENESS OF SYNTHESIS OF STOCHASTIC MODELS WITH CONTROLLED PROPERTIES 468
Y. Serzhan, T. Umarov, A. Abilbayeva
FRAUD DETECTION IN CREDIT CARD TRANSACTIONS USING MACHINE LEARNING: A COMPARATIVE ANALYSIS 484
T. Babenko, Ye. Bakhtiyarova, D. Yeskendirova, A. Sysoev, Sh. Makilenov
FLEXIBLE IDPS BASED ON MACHINE LEARNING ALGORITHMS FOR DIPTYCHS AND IOT EDGE DEVICE PROTECTION 496

МАЗМҰНЫ
ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР

Д.Е. Абжанов, А.А. Белощицкий
ЭКОЛОГИЯЛЫҚ МОНИТОРИНГТІҢ ЗИЯТКЕРЛІК ЖҮЙЕСІНДЕГІ СТАЦИОНАРЛЫҚ ЛАСТАНУ КӨЗДЕРІНІҢ ШЫҒАРЫНДЫЛАРЫ ТУРАЛЫ ДЕРЕКТЕРДІ БАСҚАРУДЫҢ МОДЕЛІ МЕН ӘДІСІ 11
А.Е. Сланбекова, М.Б. Рахимжанова, А.И. Жанибекова, А.З. Алимагамбетова, М. Худойберганов
КЕҢІСТІКТІК-УАҚЫТТЫҚ (SPATIOTEMPORAL) ТАЛДАУ НЕГІЗІНДЕ ГИДРОЛОГИЯЛЫҚ ҚАУІП-ҚАТЕРДІ ЕРТЕ АНЫҚТАУ 27
А. Сахипов, Р. Сеитбек, Д. Едильхан
АВТОМАТТАНДЫРЫЛҒАН ДЕРЕКТЕРДІ ДАЙЫНДАУДЫ ПАЙДАЛАНА ОТЫРЫП, ЖОЛАУШЫЛАР

АҒЫНЫН НАҚТЫ УАҚЫТ РЕЖИМІНДЕ БОЛЖАУ 38
К. Токаев, А. Молдагулова
ҚАЗАҚСТАНДАҒЫ ҚОР НАРЫҒЫН БОЛЖАУДЫҢ БОЛЖАМДЫҚ МОДЕЛЬДЕРІНЕ ШОЛУ 60

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР

Ф.Н. Абдраимова, А.А. Керейбаева, Д.С. Дюсенова, Д.А. Алиева, Т.Ж. Токтарова
ТІЛ БІЛІМІНДЕ ЖАСАНДЫ ИНТЕЛЛЕКТ ТЕХНОЛОГИЯЛАРЫ: СТУДЕНТТЕР ҚОЛДАНУЫНЫҢ ПРАКТИКАЛЫҚ АСПЕКТІЛЕРІ МЕН МӘСЕЛЕЛЕРІ 83
Г.Т. Азиева, М.Б. Есенова, А.К. Абжаппарова, Г.Б. Абдикеримова, P. Schmidt
UAV ДЕРЕКТЕРІ НЕГІЗІНДЕ АУЫЛ ШАРУАШЫЛЫҒЫ ДАҚЫЛДАРЫН ЖІКТЕУГЕ АРНАЛҒАН ГИБРИДТІ СТЕКИНГ МОДЕЛІ 97
Ә.Қ. Әйтім
АГГЛЮТИНАТИВТІ ТІЛДЕРГЕ АРНАЛҒАН МОРФОЛОГИЯЛЫҚ ДИЗАМБИГУАЦИЯ МЕН POS-ТАҢБАЛАУДЫ БІРЛЕСІП МОДЕЛЬДЕУ 109
А. Әуезова, Ә. Әйтім, Б. Синчев, В. Вуйчик
ҚАЗАҚСТАННЫҢ ҚҰҚЫҚТЫҚ АКТІЛЕРІНДЕГІ NP-ТОЛЫҚ АЛГОРИТМДЕР: SEMANTIC ІЗДЕУГЕ АРНАЛҒАН ГИБРИДТІ АРХИТЕКТУРА 125
Я. Гозак, С. Палий, Т. Гончаренко
КҮРДЕЛІЛІГІ ТӨМЕН ДЕРЕКТЕР ЖИЫНТЫҚТАРЫНДА ТЕРЕҢ ЕМЕС ОРАМДЫ НЕЙРОНДЫҚ ЖЕЛІЛЕРДІҢ (CNN) ОҢТАЙЛЫ КОНФИГУРАЦИЯЛАРЫН БАҒАЛАУ 138
С.А. Есниязова, С.Т. Каимов
ТҮСІНДІРІЛЕТІН МАШИНАЛЫҚ ОҚЫТУДЫ ҚОЛДАНА ОТЫРЫП АУЫР ЖҮК КӨЛІКТЕРІНЕ БОЛЖАМДЫ ТЕХНИКАЛЫҚ ҚЫЗМЕТ КӨРСЕТУ 153
Т.Д. Иманбекова, Ж.Б. Ибраева, Г.Т. Джаканова, Г.Т. Асқанбай
МӘЛІМЕТТЕРДІ ВЕЙВЛЕТ-ТҮРЛЕНДІРГІШТІҢ НЕГІЗІНДЕ ҚЫСУ АЛГОРИТМІ; MATLAB ОРТАСЫНДА ТАЛДАУ ЖӘНЕ ІСКЕ АСЫРУ 167
Б.З. Кенжегулов, Ж.Т. Билялова, К.Н. Утеулиева, Л. Нургалиева, Ш.С. Нуржанова
ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР НЕГІЗІНДЕ ИНТЕЛЛЕКТУАЛДЫ TEXT-TO-SQL ЖҮЙЕСІН ӘЗІРЛЕУДІҢ МАТЕМАТИКАЛЫҚ-АЛГОРИТМДІК ТӘСІЛІ 185
М. Кабдуллин, А. Каирбеков, А. Кабдуллин, Л. Найзабаева, А. Мухит
КАРДИОЛОГИЯЛЫҚ КЕСКІНДЕРДІ БАСҚАРУҒА АРНАЛҒАН ВЕБ-НЕГІЗДЕЛГЕН ПРОТОТИПТІ ӘЗІРЛЕУ ЖӘНЕ ІСКЕ АСЫРУ 206
Л. Курмангазиева, Ш. Ележанова, С. Искакова, Ш. Коданова, Х. И. Бұлбұл
АЙҚЫН ЕМЕС ЛОГИКА, ХАІ ЖӘНЕ ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР НЕГІЗІНДЕ ГИБРИДТІ ЗИЯТКЕРЛІК ШЕШІМ ҚАБЫЛДАУДЫ ҚОЛДАУ ЖҮЙЕСІН ҚҰРУ ТӘСІЛДЕРІ 219
З. Момынкулов, Б. Омаров
ТРАЕКТОРИЯЛАРДЫ ГЕНЕРАЦИЯЛАУ ҮШІН DDPG 236
Ф. Мырзабеков, Н. Мустафин, М. Мукушев
БҰЛ ЖАҢАЛЫҚ НЕЛІКТЕН ТЕРІС?: ТҮСІНДІРІЛЕТІН СЕНТИМЕНТТІК ТАЛДАУҒА АРНАЛҒАН НАЗАР АУДАРУ МЕХАНИЗМІНЕ НЕГІЗДЕЛГЕН ВИЗУАЛИЗАЦИЯ ӘДІСІ. ТАЛДАУЛЫҚ ШОЛУ ЖӘНЕ ТҰЖЫРЫМДАМАЛЫҚ НЕГІЗ 248
Н.Ш. Максутова, Ж.А. Тусупов, А.Ә. Шекербек, Ж.Е. Кенжебаева, К.О. Рахимов
ЖҮРЕК-ҚАН ТАМЫРЛАРЫ АУРУЛАРЫНЫҢ ҚАУІП-ҚАТЕРІН ЖӘНЕ БИОХИМИЯЛЫҚ ӨЗГЕРІСТЕРДІ КЕШЕНДІ БАҒАЛАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ: АСПАРТАТАМИНОТРАНСФЕРАЗАҒА ЕРЕКШЕ НАЗАР 260
Л. Найзабаева, Г. Түркен, Р.А. Друба, А.Ш. Явар
ВЕБ НЕГІЗІНДЕГІ САТУ ОПЕРАЦИЯЛАРЫН ЖЕТІЛДІРУ ҮШІН КӘСІПОРЫНДЫҚ ДЕРЕКТЕР ҚОЙМАСЫН ПАЙДАЛАНУ 275
С. Нургалиева, Н. Найман, К. Махметова, М. Кали, Н. Жакиев
КОМПЬЮТЕРЛІК КӨРУ ТЕХНОЛОГИЯСЫН ПАЙДАЛАНА ОТЫРЫП, КӨКӨНІС ДАҚЫЛДАРЫНЫҢ АУРУЛАРЫН ЖЕДЕЛ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ 291
О.С. Салыкова, В.А. Мадин, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов
ӨНЕРКӘСІПТІК МОНИТОРИНГ ЖҮЙЕЛЕРІНДЕГІ MEMS-АКСЕЛЕРОМЕТРЛЕРДІҢ СЕНСОРЛЫҚ МОДУЛЬДЕРІН ИНТЕГРАЦИЯЛАУ 309
Р. Таберхан, М.А. Самбетбаева, Г. Қалман

KAZCAUSAL: ҚАЗАҚ ТІЛІНДЕГІ СЕБЕП-САЛДАРЛЫҚ ҚАТЫНАСТАРДЫҢ АЛҒАШҚЫ КОРПУСТЫҚ АННОТАЦИЯСЫ 323
С. Тынымбаев, С.Е. Маманова, Р. Бердібаев, Ж.Е.Темірбекова, Т. Чинибаева
БӨЛГІШТІҢ ЕСЕЛІ МӘНДЕРІН АЛДЫН АЛА ДАЙЫНДАУМЕН ЖҮЗЕГЕ АСЫРЫЛАТЫН БӨЛУ ҚҰРЫЛҒЫЛАРЫ 335
К.Н. Утеулиева, Б.З. Кенжегулов, Т.А. Каражигитова, Х. Булбул, З.Ж. Жанузакова
КОЛЛАБОРАТИВТІК СҮЗГІЛЕУ НЕГІЗІНДЕГІ ҰСЫНЫМДЫҚ ЖҮЙЕНІ ӘЗІРЛЕУДІҢ МАТЕМАТИКАЛЫҚ-АЛГОРИТМДІК ТӘСІЛДЕРІ 351
С. Шармұханбет, Г. Тұрмұханова, О. Финдик, В. Махатова, Л. Курмангазиева
АЙНЫМАЛЫ ЖАРЫҚ ЖАҒДАЙЫНДАҒЫ ЖОҒАРЫ ДӘЛДІКТІ РОБОТТЫҚ ҚҰРАСТЫРУ: ВИЗУАЛДЫ СЕРВОТЕЖЕУДІҢ ТӨЗІМДІ МЕХАТРОНИКАЛЫҚ АРХИТЕКТУРАСЫ 372

АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАРҒА АРНАЛҒАН

А. Амирбай, З. Аманбайқызы, К. Максутова, А. Муханова, M. Kassim
КӨЗ ҚОЗҒАЛЫСТАРЫ МЕН БЕТ МИМИКА БЕЛГІЛЕРІН МУЛЬТИМОДАЛЬДЫ ТАЛДАУҒА НЕГІЗДЕЛГЕН БАЛАЛАРДАҒЫ АУТИЗМ СПЕКТРІНІҢ БҰЗЫЛЫСТАРЫН ЕРТЕ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМІ 390
К.Д. Байсылбаева, Ш.Ж. Мусиралиева, Ж. Елтай
ҚАЗАҚ ТІЛІНДЕГІ ЭКСТРЕМИСТІК ИДЕОЛОГИЯНЫ АНЫҚТАУ: АННОТАЦИЯЛАУ МӘСЕЛЕЛЕРІ ЖӘНЕ ТЕРЕҢ ОҚЫТУ ТӘСІЛДЕРІ 405
М.А. Болатбек, А.М. Усманова, Қ.Б. Багитова, Г.Б. Байспай
КИБЕР ҚАУІПТІ АНЫҚТАУ ҮШІН ЖЕЛІЛІК ТРАФИКТІ ТАЛДАУ ӘДІСІН ӘЗІРЛЕУ ЖӘНЕ ЗЕРТТЕУ 424
Д.И. Прокопович-Ткаченко, Н.К. Жумагалиева, Д.Н. Щитов, Н.Ф. Мормуль, Д.А. Черкасский
ТОЛЫҚ ЕМЕС ЖӘНЕ САПАЛЫҚ ДЕРЕКТЕР ЖАҒДАЙЫНДА АҚПАРАТТЫҚ ЖҮЙЕЛЕРДІҢ АҚПАРАТТЫҚ ҚАУІПСІЗДІК ПАРАМЕТРЛЕРІН БАҒАЛАУДЫҢ БҰЛЫҢҒЫР МОДЕЛІ: ҚҰРУ ӘДІСТЕМЕСІ, ЕРЕЖЕЛЕР БАЗАСЫН БАПТАУ ЖӘНЕ ҰЙЫМДАРҒА АРНАЛҒАН ДЕМОНСТРАЦИЯЛЫҚ КЕЙС 442
Е.А. Пустовой, О.А. Пустовая, А.Н. Раушанова, И.С. Заурбеков
БАСҚАРЫЛАТЫН ҚАСИЕТТЕРІ БАР СТОХАСТИКАЛЫҚ МОДЕЛЬДЕРДІ СИНТЕЗДЕУДІҢ ТИІМДІЛІГІН БАҒАЛАУ 468
Е. Сержан, Т. Умаров, А. Әбілбаева
МАШИНАЛЫҚ ОҚУ ӘДІСІ АРҚЫЛЫ КРЕДИТ КАРТА ОПЕРАЦИЯЛАРЫНДАҒЫ АЛАЯҚТЫҚТЫ АНЫҚТАУ: САЛЫСТЫРМАЛЫ ТАЛДАУ 484
Т.В. Бабенко, Е.А. Бахтиярова, Д.М. Ескендирова, А. Сысоев, Ш.Н. Макиленов
ДИПТИХ ЖӘНЕ IOT ШЕТКІ ҚҰРЫЛҒЫЛАРЫН ҚОРҒАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІНЕ НЕГІЗДЕЛГЕН ИКЕМДІ IDPS 496

СОДЕРЖАНИЕ
ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИСОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

Д.Е. Абжанова, А.А. Белощицкий
МОДЕЛЬ И МЕТОД УПРАВЛЕНИЯ ДАННЫМИ О ВЫБРОСАХ СТАЦИОНАРНЫХ ИСТОЧНИКОВ ЗАГРЯЗНЕНИЯ В ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЕ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА 11
А.Е. Сланбекова, М.Б. Рахимжанова, А.И. Жанибекова, А.З. Алимагамбетова, М. Худойберганов
РАННЕЕ ВЫЯВЛЕНИЕ ГИДРОЛОГИЧЕСКИХ ОПАСНОСТЕЙ НА ОСНОВЕ ПРОСТРАНСТВЕННО-ВРЕМЕННОГО (SPATIOTEMPORAL) АНАЛИЗА 27
А. Сахипов, Р. Сеитбек, Д. Едильхан
ПРОГНОЗИРОВАНИЕ ПАССАЖИРОПОТОКА В РЕАЛЬНОМ ВРЕМЕНИ С ИСПОЛЬЗОВАНИЕМ АВТОМАТИЗИРОВАННОЙ ПОДГОТОВКИ ДАННЫХ 38
К. Токаев, А. Молдагулова
ОБЗОР ПРОГНОСТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ФОНДОВОГО РЫНКА В КАЗАХСТАНЕ 60

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Ф.Н. Абдраимова, А.А. Керейбаева, Д.С. Дюсенова, Д.А. Алиева, Т.Ж. Токтарова
ТЕХНОЛОГИИ ИИ В ЯЗЫКОВОМ ОБРАЗОВАНИИ: ПРАКТИЧЕСКИЕ АСПЕКТЫ И ПРОБЛЕМЫ ПРИМЕНЕНИЯ СТУДЕНТАМИ 83
Г.Т. Азиева, М.Б. Есенова, А.К. Абжаппарова, Г.Б. Абдикеримова, P. Schmidt
ГИБРИДНАЯ МОДЕЛЬ СТЕКИНГА ДЛЯ КЛАССИФИКАЦИИ СЕЛЬСКОХОЗЯЙСТВЕННЫХ КУЛЬТУР ПО ДАННЫМ UAV 97
Ә.Қ. Әйтім
СОВМЕСТНАЯ МОРФОЛОГИЧЕСКАЯ ДИЗАМБИГУАЦИЯ И POS-РАЗМЕТКА ДЛЯ АГГЛЮТИНАТИВНЫХ ЯЗЫКОВ 109
А. Ауезова, А. Айтим, Б. Синчев, В. Вуйчик
NP-ПОЛНЫЕ АЛГОРИТМЫ В LEGALTECH: ГИБРИДНАЯ АРХИТЕКТУРА ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА В ПРАВОВЫХ АКТАХ КАЗАХСТАНА 125
Я. Гозак, С. Палий, Т. Гончаренко
ОЦЕНКА ОПТИМАЛЬНЫХ КОНФИГУРАЦИЙ НЕГЛУБОКИХ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ НА ДАТАСЕТАХ НИЗКОЙ СЛОЖНОСТИ 138
С.А. Есниязова, С.Т. Каимов
ПРЕДИКТИВНОЕ ТЕХНИЧЕСКОЕ ОБСЛУЖИВАНИЕ ТЯЖЁЛЫХ ГРУЗОВИКОВ С ИСПОЛЬЗОВАНИЕМ ОБЪЯСНИМОГО МАШИННОГО ОБУЧЕНИЯ 153
Т.Д. Иманбекова, Ж.Б. Ибраева, Г.Т. Джаканова, Г.Т. Асқанбай
АЛГОРИТМ СЖАТИЯ ДАННЫХ НА ОСНОВЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАТЕЛЯ: АНАЛИЗ И РЕАЛИЗАЦИЯ В MATLAB 167
Б.З. Кенжегулов, Ж.Т. Билялова, К.Н. Утеулиева, Л. Нургалиева, Ш.С. Нуржанова
МАТЕМАТИКО-АЛГОРИТМИЧЕСКИЙ ПОДХОД К РАЗРАБОТКЕ ИНТЕЛЛЕКТУАЛЬНОЙ TEXT-TO-SQL СИСТЕМЫ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ 185
М. Кабдуллин, А. Каирбеков, А. Кабдуллин, Л. Найзабаева, А. Мухит
РАЗРАБОТКА И РЕАЛИЗАЦИЯ ВЕБ-ПРОТОТИПА ДЛЯ УПРАВЛЕНИЯ КАРДИОЛОГИЧЕСКИМИ ИЗОБРАЖЕНИЯМИ 206
Л. Курмангазиева, Ш. Ележанова, С. Искакова, Ш. Коданова, Х. И. Бюльбюль
ПОДХОДЫ К ПОСТРОЕНИЮ ГИБРИДНОЙ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ, XAI И БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ 219
З. Момынкулов, Б. Омаров
DDPG ДЛЯ ГЕНЕРАЦИИ ТРАЕКТОРИЙ 236
Ф. Мырзабеков, Н. Мустафин, М. Мукушев
ПОЧЕМУ ЭТА НОВОСТЬ НЕГАТИВНА?: МЕТОД ВИЗУАЛИЗАЦИИ НА ОСНОВЕ МЕХАНИЗМА ВНИМАНИЯ ДЛЯ ОБЪЯСНИМОГО АНАЛИЗА ТОНАЛЬНОСТИ. АНАЛИТИЧЕСКИЙ ОБЗОР И КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ 248
Н.Ш. Максутова, Д.А. Тусупов, А.А. Шекербек, Ж.Е. Кенжебаева, К.О. Рахимов
МАШИННОЕ ОБУЧЕНИЕ ДЛЯ КОМПЛЕКСНОЙ ОЦЕНКИ РИСКА СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ И БИОХИМИЧЕСКИХ ИЗМЕНЕНИЙ: АКЦЕНТ НА АСПАРТАТАМИНОТРАНСФЕРАЗЕ 260
Л. Найзабаева, Г. Туркен, Р.А. Друба, А.Ш. Явар
ИСПОЛЬЗОВАНИЕ КОРПОРАТИВНОГО ХРАНИЛИЩА ДАННЫХ ДЛЯ УЛУЧШЕНИЯ ВЕБ-ОРИЕНТИРОВАННЫХ ОПЕРАЦИЙ ПРОДАЖ 275
С. Нургалиева, Н. Найман, К. Махметова, М. Кали, Н. Жакиев
СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ БЫСТРОГО ОБНАРУЖЕНИЯ ЗАБОЛЕВАНИЙ ОВОЩНЫХ КУЛЬТУР С ИСПОЛЬЗОВАНИЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ 291
О.С. Салыкова, В.А. Мадин, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов
ИНТЕГРАЦИЯ СЕНСОРНЫХ МОДУЛЕЙ MEMS-АКСЕЛЕРОМЕТРОВ В СИСТЕМАХ ПРОМЫШЛЕННОГО МОНИТОРИНГА 309
Р. Таберхан, М.А. Самбетбаева, Г. Калман
KAZCAUSAL: ПЕРВАЯ КОРПУСНАЯ АННОТАЦИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ НА КАЗАХСКОМ ЯЗЫКЕ 323
С. Тынымбаев, С.Е. Маманова, Р. Бердибаев, Ж.Е.Темирбекова, Т. Чинибаева

УСТРОЙСТВА ДЕЛЕНИЯ ЧИСЕЛ С ПРЕДВАРИТЕЛЬНОЙ ПОДГОТОВКОЙ КРАТНЫХ ДЕЛИТЕЛЮ 335
К.Н. Утеулиева, Б.З. Кенжегулов, Т.А. Каражигитова, Х.Бюльбюль, З.Ж. Жанузакова
МАТЕМАТИКО-АЛГОРИТМИЧЕСКИЕ ПОДХОДЫ К РАЗРАБОТКЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ 351
С. Шармуханбет, Г. Турмуханова, О.Финдик, В.Махатова, Л. Курмангазиева
ВЫСОКОТОЧНАЯ РОБОТИЗИРОВАННАЯ СБОРКА ПРИ ПЕРЕМЕННОЙ ОСВЕЩЁННОСТИ: РОБАСТНАЯ МЕХАТРОННАЯ АРХИТЕКТУРА ВИЗУАЛЬНОГО СЕРВОУПРАВЛЕНИЯ 372

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

А. Амирбай, З. Аманбайкызы, К. Максутова, А. Муханова, M. Kassim
АЛГОРИТМ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАННЕГО ВЫЯВЛЕНИЯ РАССТРОЙСТВ АУТИСТИЧЕСКОГО СПЕКТРА У ДЕТЕЙ НА ОСНОВЕ МУЛЬТМОДАЛЬНОГО АНАЛИЗА ДАННЫХ ДВИЖЕНИЯ ГЛАЗ И МИМИЧЕСКИХ СИГНАЛОВ 390
К.Д. Байсылбаева, Ш.Ж. Мусиралиева, Ж.Елтай
ОБНАРУЖЕНИЕ ЭКСТРЕМИСТСКОЙ ИДЕОЛОГИИ НА КАЗАХСКОМ ЯЗЫКЕ: ПРОБЛЕМЫ АННОТИРОВАНИЯ И МЕТОДЫ ГЛУБОКОГО ОБУЧЕНИЯ 405
М.А. Болатбек, А.М. Усманова, К.Б. Багитова, Г.Б. Байспай
РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДА АНАЛИЗА СЕТЕВОГО ТРАФИКА ДЛЯ ВЫЯВЛЕНИЯ КИБЕРУГРОЗЫ 424
Д.И. Прокопович-Ткаченко, Н.К. Жумагалиева, Д.Н. Щитов, Н.Ф. Мормуль, Д.А. Черкасский
НЕЧЕТКАЯ МОДЕЛЬ ОЦЕНИВАНИЯ ПАРАМЕТРОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ИНФОРМАЦИОННЫХ СИСТЕМ В УСЛОВИЯХ НЕПОЛНЫХ И КАЧЕСТВЕННЫХ ДАННЫХ: МЕТОДИКА ПОСТРОЕНИЯ, НАСТРОЙКА БАЗЫ ПРАВИЛ И ДЕМОНСТРАЦИОННЫЙ КЕЙС ДЛЯ ОРГАНИЗАЦИЙ 442
Е.А. Пустовой, О.А. Пустовая, А.Н. Раушанова, И.С. Заурбеков
ОЦЕНКА ЭФФЕКТИВНОСТИ СИНТЕЗА СТОХАСТИЧЕСКИХ МОДЕЛЕЙ С УПРАВЛЯЕМЫМИ СВОЙСТВАМИ 468
Е. Сержан, Т. Умаров, А. Абильбаева
ВЫЯВЛЕНИЕ МОШЕННИЧЕСТВА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ ПРИ ОПЕРАЦИЯХ С КРЕДИТНЫМИ КАРТАМИ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ 484
Т. Бабенко, Е. Бахтиярова, Д. Ескендирова, А. Сысоев, Ш. Макиленов
ГИБКАЯ IDPS НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ МОНИТОРИНГА И ЗАЩИТЫ ПЕРИФЕРИЙНЫХ IoT-УСТРОЙСТВ 496

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.2. Number 26 (2026). Pp. 11–26
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.001
IRSTI / ҒТАХР / МРНТИ / 20.15.05

A MODEL AND METHOD FOR MANAGING DATA ON EMISSIONS FROM STATIONARY SOURCES OF POLLUTION IN AN INTELLIGENT ENVIRONMENTAL MONITORING SYSTEM

D. Abzhanova*, A. Biloshchytskyi
Astana IT University, Astana, Kazakhstan.
E-mail: dilara.abzhanova@astanait.edu.kz

Dilara Abzhanova — PhD doctoral student, Director of the Center of Competence and Excellence, Astana IT University, Astana, Kazakhstan
E-mail: dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Andrii Biloshchytskyi — Doctor of Technical Sciences, Vice-Rector for Science and Innovations, Astana IT University, Astana, Kazakhstan
Department of Information Technology, Kyiv National University of Construction and Architecture, Kyiv, Ukraine
https://orcid.org/0000-0001-9548-1959.

© D. Abzhanova, A. Biloshchytskyi

Abstract. The paper examines modern approaches to ensuring environmental sustainability in the context of the development of the Industry 4.0 concept. Special attention is paid to the tasks of monitoring atmospheric air as an environmental risk management tool. The analysis of existing information and analytical solutions in this field has been carried out, and their limitations related to insufficient data integration and a low level of information security have been identified. An approach to building an intelligent monitoring system based on the combined use of Internet of Things and blockchain technologies is proposed. As part of the research, a database structure has been developed designed to account for stationary sources of pollution and regulate emission quotas. The results obtained can be used to increase the reliability of environmental data,

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

as well as to improve the mechanisms for monitoring atmospheric air quality

Keywords: emission of harmful substances, regulation of emission quotas, modern data structure, regulatory and control mechanisms, blockchain technologies, smart contracts, high level of security, intelligent approach, environmental problems

For citation: D. Abzhanova, A. Biloshchytskyi (2026). A model and method for managing data on emissions from stationary sources of pollution in an intelligent environmental monitoring system // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 11–26. https://doi.org/10.54309/IJICT.2026.26.2.001. (In Russ.).

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

ЭКОЛОГИЯЛЫҚ МОНИТОРИНГТІҢ ЗИЯТКЕРЛІК ЖҮЙЕСІНДЕГІ СТАЦИОНАРЛЫҚ ЛАСТАНУ КӨЗДЕРІНІҢ ШЫҒАРЫНДЫЛАРЫ ТУРАЛЫ ДЕРЕКТЕРДІ БАСҚАРУДЫҢ МОДЕЛІ МЕН ӘДІСІ

Д.А. Абжанова*, А.А. Белощицкий
Astana IT University, Астана, Қазақстан.
E-mail: dilara.abzhanova@astanait.edu.kz

Абжанова Дилара Ерлановна — PhD докторант, құзыреттілік және жетілдіру орталығының директоры, Astana IT University, Астана, Қазақстан
E-mail: dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Белощицкий Андрей Александрович — техника ғылымдарының докторы, ғылым және инновациялар жөніндегі проректор, Astana IT University, Астана, Қазақстан; Ақпараттық технологиялар кафедрасы, Киев ұлттық құрылыс және сәулет университеті, Киев, Украина
https://orcid.org/0000-0001-9548-1959.

© Д.Е. Абжанова, А.А. Белощицкий

Аннотация. Жұмыста Индустрия 4.0 тұжырымдамасын дамыту жағдайында экологиялық тұрақтылықты қамтамасыз етудің заманауи тәсілдері қарастырылады. Экологиялық тәуекелдерді басқару құралы ретінде атмосфералық ауаны бақылау міндеттеріне ерекше назар аударылады. Осы саладағы қолданыстағы ақпараттық-талдамалық шешімдерге талдау жүргізілді, олардың деректерді жеткіліксіз интеграциялаумен және ақпараттың қорғалу деңгейінің төмендігімен байланысты шектеулері анықталды. IoT және Blockchain технологияларын бөлісуге негізделген интеллектуалды бақылау жүйесін құру тәсілі ұсынылды. Зерттеу аясында стационарлық ластану көздерін есепке алуға және эмиссиялық квоталарды реттеуге арналған мәліметтер базасының құрылымы жасалды. Алынған нәтижелер экологиялық деректердің дұрыстығын арттыру үшін, сондай-ақ атмосфералық ауаның сапасын бақылау тетіктерін жетілдіру үшін пайдаланылуы мүмкін.

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

Түйінді сөздер: зиянды заттардың эмиссиясы, эмиссия бойынша квоталарды регламенттеу, деректердің заманауи құрылымы, регламенттеу және бақылау тетіктері, блокчейн-технологиялар, смарт-келісімшарттар, қауіпсіздіктің жоғары деңгейі, зияткерлік тәсіл, экологиялық проблемалар

Дәйексөздер үшін: Д.Е. Абжанова, А.А. Белощицкий (2026). Экологиялық мониторингтің зияткерлік жүйесіндегі стационарлық ластану көздерінің шығарындылары туралы деректерді басқарудың моделі мен әдісі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 11–26. https://doi.org/10.54309/IJICT.2026.26.2.001. (Орыс. тіл.).

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

МОДЕЛЬ И МЕТОД УПРАВЛЕНИЯ ДАННЫМИ О ВЫБРОСАХ СТАЦИОНАРНЫХ ИСТОЧНИКОВ ЗАГРЯЗНЕНИЯ В ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЕ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА

Д.А. Абжанова*, А.А. Белощицкий
Astana IT University, Астана, Казахстан.
E-mail: dilara.abzhanova@astanait.edu.kz

Абжанова Дилара Ерлановна — PhD докторант, Директор Центра компетенций и совершенства, Astana IT University, Астана, Казахстан
E-mail: dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Белощицкий Андрей Александрович — доктор технических наук, проректор по науке и инновациям, Astana IT University, Астана, Казакстан
Факультет информационных технологий, Киевский национальный университет строительства и архитектуры, Киев, Украина
https://orcid.org/0000-0001-9548-1959.

© Д.Е. Абжанова, А.А. Белощицкий

Аннотация. В работе рассматриваются современные подходы к обеспечению экологической устойчивости в условиях развития концепции Индустрии 4.0. Особое внимание уделено задачам мониторинга атмосферного воздуха как инструменту управления экологическими рисками. Проведен анализ существующих информационно-аналитических решений в данной области, выявлены их ограничения, связанные с недостаточной интеграцией данных и низким уровнем защищенности информации. Предложен подход к построению интеллектуальной системы мониторинга, основанный на совместном использовании технологий интернета вещей и блокчейна. В рамках исследования разработана структура базы данных, предназначенной для учета стационарных источников загрязнения и регулирования эмиссионных квот. Полученные результаты могут быть использованы для повышения достоверности экологических данных, а также для совершенствования механизмов

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

контроля качества атмосферного воздуха.

Ключевые слова: эмиссия вредных веществ, регламентирование квот по эмиссии, современная структура данных, механизмы регламентации и контроля, блокчейн-технологии, смарт-контракты; высокий уровень безопасности; интеллектуальный подход, экологические проблемы

Для цитирования: Д.Е. Абжанова, А.А. Белощицкий (2026). Модель и метод управления данными о выбросах стационарных источников загрязнения в интеллектуальной системе экологического мониторинга // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 11–26. https://doi.org/10.54309/IJICT.2026.26.2.001. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение.

В условиях цифровой трансформации экономики существенно возрастает значение интеллектуальных информационных систем в обеспечении экологической безопасности. Развитие цифровых технологий и переход к концепции Индустрии 4.0 формируют новые требования к системам мониторинга окружающей среды, включая необходимость непрерывного сбора, обработки и анализа данных о выбросах загрязняющих веществ (Абдрахманова и др., 2021; Постановление Правительства Республики Казахстан, 2022; Постановление Правительства Республики Казахстан, 2023).

Современные промышленные предприятия являются значимыми источниками антропогенного воздействия на окружающую среду, что обусловливает необходимость разработки и внедрения эффективных инструментов контроля и управления экологическими параметрами. Применение систем непрерывного мониторинга выбросов позволяет обеспечивать получение достоверной информации о состоянии окружающей среды в режиме реального времени, что повышает обоснованность принимаемых управленческих решений (Göttel, 2019).

Вместе с тем увеличение объема данных, поступающих от сенсорных устройств, приводит к возникновению ряда проблем, связанных с их обработкой, хранением и обеспечением целостности. Современные исследования в области цифровой экологии демонстрируют, что интеграция распределенных информационных систем и интеллектуальных методов анализа данных способствует повышению эффективности экологического мониторинга (Green et al., 2020).

Однако большинство существующих решений ориентированы на локальную обработку данных и недостаточно учитывают стохастическую природу информационных потоков, что ограничивает возможности прогнозирования и адаптивного управления экологическими процессами (Мутанов и др., 2009).

В этой связи актуальной научной задачей является разработка моделей и методов управления данными о выбросах стационарных источников, обеспечивающих интеллектуальную обработку, прогнозирование и надежное хранение экологической информации в условиях неопределенности и высокой динамики данных.

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

Материалы и методы исследования.

В рамках настоящего исследования рассматривается задача разработки интеллектуальной системы мониторинга выбросов загрязняющих веществ стационарными источниками в условиях высокой неопределенности и неоднородности входных данных.

Объектом исследования являются процессы формирования, передачи и анализа данных экологического мониторинга, поступающих от распределенной сети сенсорных устройств. Предметом исследования выступают модели и методы интеллектуальной обработки экологических данных, обеспечивающие повышение точности оценки и прогнозирования состояния окружающей среды.

Цель исследования заключается в разработке модели и метода управления потоками экологических данных, обеспечивающих их интеграцию, обработку, прогнозирование и надежное хранение в условиях стохастичности и неполноты информации.

Методологическую основу исследования составляют методы математического моделирования, теории вероятностей и математической статистики, методы анализа временных рядов и машинного обучения.

Исходные данные. Эмпирическая часть исследования выполнена на выборке данных мониторинга выбросов загрязняющих веществ стационарными источниками. В исследование включены временные ряды наблюдений за 2015–2024 гг. и данные сенсорных измерений по основным контролируемым показателям выбросов, включая CO_2, NO_x и SO_2. Совокупный объем выборки составил более 120 000 наблюдений. Для каждого наблюдения фиксировались временная метка, идентификатор источника, тип загрязняющего вещества, измеренное значение и сопутствующие технологические параметры. Исходные данные были сформированы на основе результатов мониторинга стационарных источников загрязнения, включающих архивные экологические наблюдения, результаты автоматизированных измерений сенсорных устройств и производственные данные объектов промышленной инфраструктуры. При формировании экспериментальной выборки персональные и коммерчески чувствительные сведения были исключены.

Предварительная обработка данных. Перед построением моделей была проведена предварительная обработка исходных данных. Она включала удаление аномальных значений методом межквартильного размаха, восстановление пропусков методом линейной интерполяции и нормализацию признаков. Доля пропусков в исходной выборке не превышала 8%, что соответствует реальным условиям функционирования систем экологического мониторинга. Все операции предобработки оценивались на обучающей части данных и затем применялись к валидационной и тестовой выборкам, что исключало утечку информации между этапами эксперимента.

Алгоритмы и параметры моделирования. Для сравнительного анализа использовались следующие базовые методы прогнозирования: линейная регрессия, ARIMA и экспоненциальное сглаживание. Предлагаемый авторский метод объеди-

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

няет стохастическую модель накопления и обработки данных, предметно-ориентированную структуру базы данных и интеллектуальный модуль прогнозирования экологических показателей. Параметры стохастической модели λ и μ интерпретируются как интенсивности поступления и обработки данных соответственно и оценивались по эмпирическим характеристикам потока записей в системе за анализируемый период. Для модели ARIMA использовалась спецификация ARIMA(p,d,q), параметры которой подбирались по минимуму информационного критерия AIC. Для всех моделей были зафиксированы одинаковые входные данные и единые метрики качества.

Экспериментальный протокол и валидация. Поскольку исследование основано на временных рядах, оценка качества выполнялась на хронологически разделенных данных: обучающая выборка формировалась из раннего временного интервала, а тестовая-из более позднего. Для настройки параметров модели использовалась валидационная схема на основе последовательных временных разбиений. Качество прогнозирования оценивалось по метрикам MAE и RMSE. Для проверки устойчивости результатов эксперимент повторялся 100 раз с одинаковым протоколом, после чего вычислялись средние значения метрик и стандартные отклонения. Повторение эксперимента выполнялось для оценки устойчивости модели к случайным изменениям обучающей выборки и проверки воспроизводимости результатов.

Архитектура предлагаемой системы основана на интеграции нескольких функциональных модулей: модуля сбора данных с IoT-датчиков, модуля предварительной обработки и фильтрации данных, аналитического модуля, реализующего алгоритмы интеллектуального анализа, а также модуля хранения данных с использованием распределенного реестра для обеспечения целостности информации.

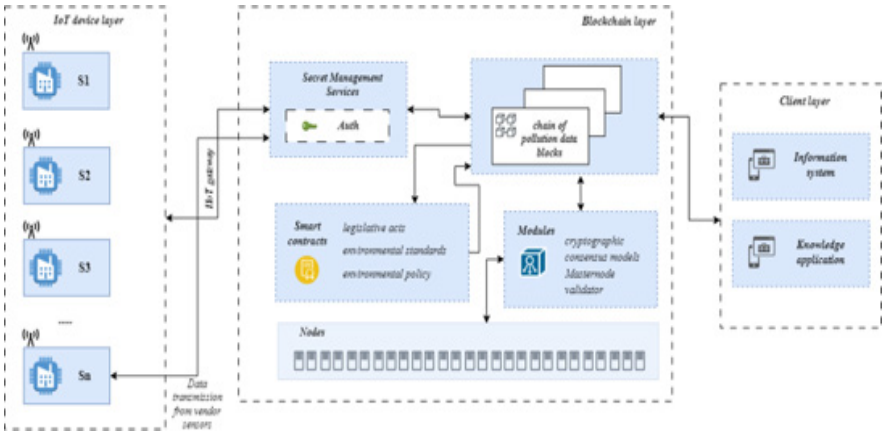


Рис. 1. Архитектура интеллектуальной системы экологического мониторинга выбросов

Как показано на рисунке 1, система включает три уровня: уровень сбора данных, уровень хранения и уровень аналитики. Это обеспечивает непрерывность мониторинга и достоверность данных.

Для систематизации информации в системе разработана структура базы дан-

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

ных, включающая сущности, описывающие источники выбросов, параметры загрязнения, временные ряды наблюдений и результаты анализа.

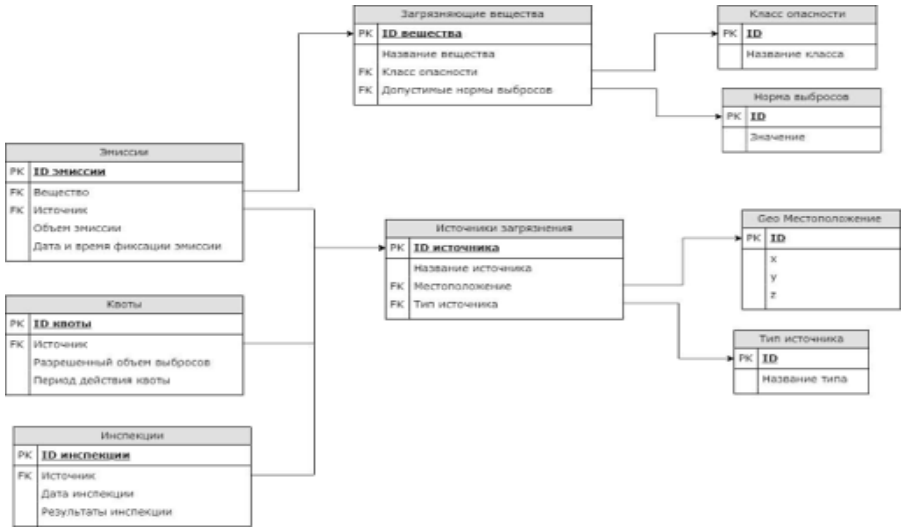


Рис. 2. Структура базы данных системы мониторинга выбросов

Представленная структура обеспечивает целостность данных и возможность их последующей обработки и анализа.

Для описания процесса накопления и обработки данных используется стохастическая модель с дискретными состояниями.

Пусть X(t) —случайная функция, описывающая количество накопленной информации в момент времени t. Поток поступления данных имеет интенсивность λ, поток обработки-интенсивность μ.

Динамика системы описывается системой уравнений:

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t)$$
$$\frac{dP_k(t)}{dt} = -(\lambda + k\mu)P_k(t) + \lambda P_{k-1}(t) + (k+1)\mu P_{k+1}(t)$$

где $P_k(t)$— вероятность нахождения системы в состоянии k.

Математическое ожидание определяется как:

$$m_x(t) = \sum_{k=0}^{\infty} k\, P_k(t)$$

При стационарных параметрах:

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

$$m_x(t) = \frac{\lambda}{\mu}(1-e^{-\mu t})$$

Дисперсия процесса определяется выражением:

$$D_x(t) = m_x(t)$$

Для ограниченной емкости системы используется:

$$\frac{dm_x(t)}{dt} = \lambda - \mu m_x(t) - \lambda P_n(t)$$

Таким образом, предложенная модель позволяет оценивать динамику накопления данных и устойчивость системы.

Обзор литературы

Современные исследования в области экологического мониторинга демонстрируют устойчивую тенденцию к цифровизации процессов сбора, обработки и анализа данных. В работах отечественных и зарубежных авторов подчеркивается необходимость перехода от традиционных методов контроля к интеллектуальным системам, способным функционировать в условиях высокой неопределенности и динамичности окружающей среды (Абдрахманова и др., 2021).

Значительное внимание в научной литературе уделяется применению геоинформационных систем для анализа пространственного распределения загрязнений и оценки экологических рисков. Геоинформационные подходы позволяют интегрировать разнородные источники данных и обеспечивать визуализацию экологических показателей, однако их использование в большинстве случаев ограничивается задачами статического анализа и не предполагает полноценной интеграции с интеллектуальными методами прогнозирования (Мутанов и др., 2009).

В ряде исследований рассматриваются системы непрерывного мониторинга выбросов, основанные на использовании сенсорных технологий и автоматизированных средств контроля. Такие системы обеспечивают высокую точность измерений и оперативность получения данных, однако сталкиваются с проблемами масштабируемости, а также с ограниченными возможностями обработки больших объемов информации в реальном времени (Göttel, 2019).

С развитием технологий Интернета вещей наблюдается активное внедрение распределенных сенсорных сетей, позволяющих формировать потоки данных высокой частоты. В работах зарубежных исследователей отмечается, что интеграция IoT-технологий с облачными платформами и аналитическими инструментами способствует повышению эффективности мониторинга окружающей среды, однако

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

при этом возрастает сложность обеспечения достоверности и целостности данных (Green et al., 2020).

Отдельное направление исследований связано с применением методов машинного обучения и анализа временных рядов для прогнозирования экологических показателей. Использование таких методов позволяет выявлять скрытые закономерности в данных и повышать точность прогнозов, однако их эффективность существенно зависит от качества входных данных и полноты обучающих выборок.

Несмотря на значительное количество исследований, следует отметить ряд нерешенных проблем. Во-первых, большинство существующих решений ориентированы на обработку данных в рамках отдельных подсистем и не обеспечивают их комплексную интеграцию. Во-вторых, недостаточно учитывается стохастическая природа потоков экологических данных, что снижает достоверность прогнозных моделей. В-третьих, вопросы обеспечения целостности и защищенности данных в распределенных системах мониторинга остаются недостаточно проработанными.

Таким образом, анализ существующих исследований показывает, что актуальной остается задача разработки интегрированной интеллектуальной системы экологического мониторинга, объединяющей методы машинного обучения, анализ временных рядов и технологии распределенного хранения данных. Предлагаемое в настоящей работе решение направлено на устранение указанных ограничений за счет комплексного подхода к обработке и анализу экологической информации. Научная новизна исследования.

В отличие от существующих работ, ориентированных либо на локальный сбор экологических показателей, либо на изолированное применение методов прогнозирования, в настоящем исследовании предложены модель и метод управления данными о выбросах стационарных источников загрязнения, объединяющие в единой архитектуре: а) многоуровневый сбор и интеграцию данных от распределенных сенсорных устройств; б) предметно-ориентированную структуру базы данных для учета источников загрязнения, загрязняющих веществ, квот и результатов инспекций; в) стохастическое описание процессов накопления и обработки данных; г) интеллектуальный модуль прогнозирования экологических показателей. Новизна работы состоит не в отдельном использовании IoT, блокчейна или временных рядов, а в их совместном применении как единого механизма управления экологическими данными.

Практическая значимость результатов заключается в том, что разработанный подход может быть использован при создании и модернизации интеллектуальных систем экологического мониторинга на промышленных предприятиях и в государственных системах контроля окружающей среды. Предлагаемая архитектура обеспечивает не только мониторинг показателей выбросов, но и повышение прослеживаемости данных, поддержку контроля эмиссионных квот и формирование достоверной информационной основы для принятия управленческих решений.

Результаты и обсуждение.

Для проверки эффективности предложенной модели и метода был проведен

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

вычислительный эксперимент на данных мониторинга выбросов загрязняющих веществ стационарными источниками за 2015–2024 гг. Базовую эмпирическую выборку составили более 120 000 наблюдений по показателям CO_2, NO_x и SO_2. Для анализа применялась единая схема предобработки данных, включающая удаление аномальных значений методом межквартильного размаха, восстановление пропусков методом линейной интерполяции и нормализацию признаков. Доля пропусков в исходных данных составляла до 8%.

Валидация моделей выполнялась на хронологически разделенных данных, что соответствует природе задачи прогнозирования временных рядов. Обучение проводилось на раннем временном интервале, настройка параметров-на валидационной части, а итоговая оценка качества-на тестовом интервале. Для повышения надежности результатов эксперимент повторялся 100 раз, после чего рассчитывались средние значения метрик и стандартные отклонения.

В качестве базовых методов сравнения использовались линейная регрессия, ARIMA и экспоненциальное сглаживание. Все модели обучались на одинаковом наборе признаков и оценивались по единым метрикам MAE и RMSE. Метрика MAE характеризует среднюю абсолютную ошибку прогноза, тогда как RMSE дополнительно усиливает штраф за крупные отклонения и измеряется в тех же единицах, что и прогнозируемая величина.

Процесс поступления и обработки экологических данных в системе моделируется как стохастический процесс с непрерывным временем и дискретными состояниями. Состояние системы определяется количеством накопленных записей в базе данных.

Пусть X(t)— случайная функция, характеризующая объем накопленной информации в момент времени t. Интенсивность поступления данных обозначим через λ, а интенсивность обработки-через μ.

Динамика вероятностей состояний системы описывается системой дифференциальных уравнений:

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t),$$

$$\frac{dP_k(t)}{dt} = -(\lambda + k\mu)P_k(t) + \lambda P_{k-1}(t) + (k+1)\mu P_{k+1}(t), k = 1,2,\dots,n.$$

где $P_k(t) = P(X(t) = k)$— вероятность того, что система находится в состоянии.

Математическое ожидание объема накопленных данных определяется следующим выражением:

$$m_x(t) = \sum_{k=0}^{\infty} k\, P_k(t).$$

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

Для стационарного случая при постоянных λ и μ решение имеет вид:

$$m_x(t) = \frac{\lambda}{\mu}(1-e^{-\mu t}) + m_{x_0}e^{-\mu t}.$$

При нулевом начальном состоянии $m_{x_0} = 0$:

$$m_x(t) = \frac{\lambda}{\mu}(1-e^{-\mu t}).$$

Дисперсия случайного процесса определяется как:

$$D_x(t) = \mathbb{E}[X^2(t)] - (m_x(t))^2.$$

Для рассматриваемой модели получено:

$$D_x(t) = \frac{\lambda}{\mu}(1-e^{-\mu t}),$$

что свидетельствует о том, что дисперсия пропорциональна математическому ожиданию, а значит процесс обладает свойствами пуассоновского типа.

В случае ограничения емкости базы данных значением n, динамика математического ожидания определяется уравнением:

$$\frac{dm_x(t)}{dt} = \lambda - \mu m_x(t) - \lambda P_n(t),$$

где $P_n(t)$— вероятность заполнения системы.

Соответствующее уравнение для дисперсии имеет вид:

$$\frac{dD_x(t)}{dt} = -2\mu D_x(t) + \mu m_x(t) + \lambda - \lambda P_n(t)[1 + 2n - 2m_x(t)].$$

Данные соотношения позволяют оценить вероятность перегрузки системы и определить критические режимы функционирования.

Граф состояний системы представлен на рисунке 3.

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



Рис. 3. Граф состояний процесса накопления информации в системе

Как видно из рисунка, переходы между состояниями определяются интенсивностями потоков поступления и обработки данных, что позволяет описывать систему как марковский процесс.

Реализация случайного процесса *X(t)* представлена на рисунке 4.

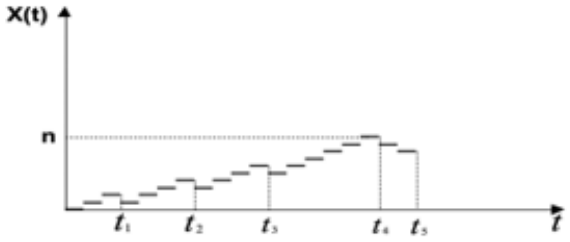


Рис. 4. Реализация процесса накопления данных

График демонстрирует стохастический характер изменения объема данных и подтверждает адекватность выбранной модели.

Для оценки устойчивости системы был рассмотрен случай, при котором интенсивность обработки зависит от текущего объема данных: μ_k=kμ.

Соответствующая модель представлена на рисунке 5.

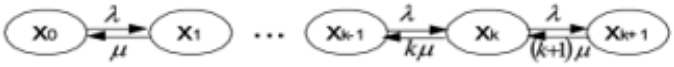


Рис. 5. Граф состояний системы при

Анализ показал, что при увеличении объема данных возрастает интенсивность их обработки, что предотвращает переполнение системы и обеспечивает ее устойчивость.

Для проверки эффективности предложенного метода была проведена сравнительная оценка с базовыми моделями прогнозирования.

Качество прогнозирования оценивалось с использованием метрик:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |,$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2},$$

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

где y_i — фактические значения, $\hat{y}_i$ — прогнозные значения.

Результаты представлены в таблице 1.

Таблица 1 – Сравнение точности моделей прогнозирования

Модель	MAE	RMSE
Линейная регрессия	12.8	18.4
ARIMA	10.2	14.7
Экспоненциальное сглаживание	9.6	13.9
Предложенная модель	6.1	8.3

Для оценки статистической устойчивости полученных результатов были рассчитаны средние значения метрик по повторным прогонам и их стандартные отклонения. Предлагаемая модель сохраняла наименьшие значения MAE и RMSE во всех сериях экспериментов, а вариативность результатов оставалась в допустимых пределах. Это позволяет сделать вывод не только о более высокой точности, но и о стабильности предложенного подхода.

Анализ результатов показывает, что предложенная модель продемонстрировала наименьшие значения MAE и RMSE среди всех сравниваемых методов. По сравнению с линейной регрессией ошибка RMSE снизилась более чем на 54%, относительно модели ARIMA-на 44%, а по сравнению с экспоненциальным сглаживанием-на 40%. Полученные результаты свидетельствуют о более высокой способности предложенного подхода учитывать стохастическую природу экологических данных.

Проведенные исследования подтвердили, что предложенная модель позволяет эффективно описывать процессы накопления и обработки экологических данных, а также обеспечивает высокую точность прогнозирования.

Интеграция методов стохастического моделирования, анализа временных рядов и интеллектуальной обработки данных позволяет повысить надежность системы и обеспечить ее устойчивость к неопределенности входной информации.

Таким образом, полученные результаты свидетельствуют о целесообразности применения разработанного подхода при создании современных систем экологического мониторинга.

Заключение.

В ходе проведенного исследования была решена актуальная научно-прикладная задача разработки модели и метода интеллектуального экологического мониторинга выбросов загрязняющих веществ стационарными источниками в условиях неопределенности и высокой динамики данных.

Реализация поставленной цели обеспечена за счет применения комплексного подхода, включающего методы математического моделирования, анализа временных рядов и машинного обучения, а также интеграцию распределенных технологий хранения данных. Предложенная архитектура системы обеспечивает непрерывный сбор, обработку и анализ данных, поступающих от сенсорных устройств, что позволяет формировать достоверную информационную основу для принятия управленческих решений.

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

Полученные результаты экспериментальных исследований подтвердили эффективность разработанного метода. Установлено, что предложенная модель обеспечивает повышение точности прогнозирования экологических показателей по сравнению с базовыми методами, а также демонстрирует устойчивость к наличию шумов и пропусков в данных. Это свидетельствует о возможности ее применения в реальных условиях функционирования систем экологического мониторинга.

Научная новизна исследования заключается в разработке интегрированной модели обработки экологических данных, учитывающей стохастическую природу информационных потоков, а также в предложении метода, обеспечивающего совместное использование интеллектуальных алгоритмов анализа и распределенных механизмов хранения данных для повышения надежности и достоверности информации.

Практическая значимость результатов определяется возможностью их использования при создании и модернизации систем экологического мониторинга на промышленных предприятиях, а также в государственных системах контроля окружающей среды. Разработанный программный прототип может быть внедрен в существующие информационные системы с целью повышения эффективности экологического контроля и снижения экологических рисков.

Перспективы дальнейших исследований связаны с расширением функциональных возможностей системы за счет внедрения адаптивных алгоритмов самообучения, интеграции дополнительных источников данных, включая спутниковый мониторинг, а также разработкой методов интеллектуальной поддержки принятия решений в области экологического управления.

Таким образом, результаты исследования вносят вклад в развитие методов интеллектуального экологического мониторинга и создают основу для дальнейшего совершенствования информационно-аналитических систем в условиях цифровой трансформации.

ЛИТЕРАТУРЫ

Абдрахманова, Г.И., Быховский, К.Б., Веселитская, Н.Н. и др. (2021). Цифровая трансформация отраслей: стартовые условия и приоритеты // Материалы 22-й международной научной конференции по проблемам развития экономики и общества, С. 11–16. [На русс.].

Айдосов, А.А., Заурбеков, Н.С. (2000). Математическое моделирование процесса распространения вредных примесей в атмосфере при переменном профиле скорости // Материалы конференции, — А. С. 232–275. [На русс.].

Айдосов, А., Турганбаева, К.С. (2002). Моделирование распространения вредных выбросов в пограничном слое атмосферы. Гидрометеорология и экология. № 2. С. 21–33. [На русс.].

Берлянд, М.Е., Генихович, Е.Л. (1979). Атмосферная диффузия и структура воздушного потока над неоднородной подстилающей поверхностью. В кн.: Метеорологические аспекты загрязнения атмосферы. Л.: Гидрометеоиздат, С. 49–69. [На русс.].

Бююль, А., Цефель, П. (2001). SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. — М.: ДиаСофт. С. 608. [На русс.].

Biloshchytskyi, A., Omirbayev, S., Mukhatayev, A., Faizullin, A., Toxanov, S., Kassenov, K. (2020). Research on the formation level of methodological competence of IT disciplines teachers. 2020 IEEE ATIT, С. 242–245. DOI: https://doi.org/10.1109/ATIT50783.2020.9349337 [На англ.].

Chupryna, I., Tormosov, R., Abzhanova, D., Ryzhakov, D., Gonchar, V., Plys, N. (2022). Scientific and methodological approaches to risk management of clean energy projects implemented in Ukraine on the terms of public-private partnership. SIST 2022. DOI: https://doi.org/10.1109/SIST54437.2022.9945809 [На англ.].

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

Davis, A. (2019). Environment Management. In: Mastering Salesforce DevOps. Berkeley: Apress, C. 141–201. [На англ.].

Göttel, H. (2019). Application of continuous emissions measurement systems for the determination of CO_2 emissions. German Emissions Trading Authority (DEHSt), C. 34. [На англ.].

Green, D.G., Klomp, N.I., Rimmington, G. (2020). Digital ecology: New technologies are revolutionizing ecology. Cham: Springer. — Т. 22. — С. 197–224. [На англ.].

Киянский, В. (2000). Новые системы управления охраной окружающей среды. Промышленность Казахстана. № 2. С. 78–80. [На русс.].

Жумагулов, Б.Т., Абдибеков, У.С., Ахметов, О.А. (2006). Турбулентный перенос примеси в искривленных каналах // Вестник Национальной инженерной академии РК, № 1, С. 24–27. [На русс.].

Kulzhanov, M., Rechel, B. (2007). Kazakhstan: Health system review. Health Systems in Transition. — Т. 9. — No. 7. С. 1–158. [На англ.].

Мутанов, Г.М., Швец, О.Я. (2009). Математическое обеспечение трансферта информационных технологий для задач экологического мониторинга. Материалы конференции. — Усть-Каменогорск, С. 113–117. [На русс.].

Постановление Правительства Республики Казахстан. (2022). Приоритеты цифровой трансформации Казахстана. Доступно по ссылке: https://adilet.zan.kz/rus/docs/P2300000248/info [На русс.].

Постановление Правительства Республики Казахстан. (2023). Концепция цифровой трансформации, развития отрасли информационно-коммуникационных технологий и кибербезопасности на 2023–2029 годы. Доступно по ссылке: https://adilet.zan.kz/rus/docs/P2300000269 [На русс].

Парк, Дж., Маккей, С. (2006). Сбор данных в системах контроля и управления. — М.: Группа ИДТ. С. 505.

Sultangazin, U.M. (2001). Mathematical problems connected with construction of algorithms for atmosphere correction in remote sensing // Journal of Inverse and Ill-Posed Problems. —Т. 9. — No. 6. — С. 655–668. [На англ.].

Темирбеков, Н.М., Мадияров, М.Н., Абдолдина, Ф.Н. и др. (2006). Численное моделирование атмосферных процессов на ограниченной территории // Вычислительные технологии. — Т. 11. — No. 2. С. 41–45. [На русс.].

REFERENCES

Abdrakhmanova, G.I., Bykhovsky, K.B., Veselitskaya, N.N., et al. (2021). Digital Transformation of Industries: Initial Conditions and Priorities. Proceedings of the 22nd International Scientific Conference on Economic and Social Development Issues. Pp. 11–16. [in Rus.].

Aidosov, A.A., Zaurbekov, N.S. (2000). Mathematical Modeling of the Atmospheric Dispersion of Harmful Pollutants under a Variable Wind Velocity Profile. Conference Proceedings. — A. Pp. 232–275. [in Rus.].

Aidosov, A., Turganbayeva, K.S. (2002). Modeling the Dispersion of Harmful Emissions in the Atmospheric Boundary Layer. Hydrometeorology and Ecology. No. 2. Pp. 21–33. [in Rus.].

Berlyand, M.E., Genikhovich, E.L. (1979). Atmospheric Diffusion and the Structure of Airflow over a Heterogeneous Underlying Surface. In: Meteorological Aspects of Atmospheric Pollution. Leningrad: Gidrometeoizdat. Pp. 49–69. [in Rus.].

Bühl, A., Zöfel, P. (2001). SPSS: The Art of Information Processing. Statistical Data Analysis and the Discovery of Hidden Patterns. — M.: DiaSoft. 608 p. [in Rus.].

Biloshchytskyi, A., Omirbayev, S., Mukhatayev, A., Faizullin, A., Toxanov, S., Kassenov, K. (2020). Research on the Formation Level of Methodological Competence of IT Disciplines Teachers. 2020 IEEE ATIT. Pp. 242–245. DOI: https://doi.org/10.1109/ATIT50783.2020.9349337. [in Eng.].

Chupryna, I., Tormosov, R., Abzhanova, D., Ryzhakov, D., Gonchar, V., Plys, N. (2022). Scientific and Methodological Approaches to Risk Management of Clean Energy Projects Implemented in Ukraine on the Terms of Public-Private Partnership. SIST 2022. DOI: https://doi.org/10.1109/SIST54437.2022.9945809. [in Eng.].

Davis, A. (2019). Environment Management. In: Mastering Salesforce DevOps. Berkeley: Apress. Pp. 141–201. [in Eng.].

Göttel, H. (2019). Application of Continuous Emissions Measurement Systems for the Determination of CO_2 Emissions. German Emissions Trading Authority (DEHSt). 34 p. [in Eng.].

Green, D.G., Klomp, N.I., Rimmington, G. (2020). Digital Ecology: New Technologies Are Revolutionizing Ecology. Cham: Springer. — Vol. 22. — Pp. 197–224. [in Eng.].

Kiyansky, V. (2000). New Environmental Management Systems. Industry of Kazakhstan. No. 2. Pp. 78–80. [in Rus.].

Zhumagulov, B.T., Abdibekov, U.S., Akhmetov, O.A. (2006). Turbulent Transport of Impurities in Curved

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

Channels // Bulletin of the National Engineering Academy of the Republic of Kazakhstan. No. 1. Pp. 24–27. [in Rus.].

Kulzhanov, M., Rechel, B. (2007). Kazakhstan: Health System Review. Health Systems in Transition. — Vol. 9. — No. 7. Pp. 1–158. [На англ.].

Mutanov, G.M., Shvets, O.Ya. (2009). Mathematical Support for Information Technology Transfer in Environmental Monitoring Tasks // Conference Proceedings. — Ust-Kamenogorsk. Pp. 113–117. [in Rus.].

Government of the Republic of Kazakhstan. (2022). Priorities of Kazakhstan's Digital Transformation. Available at: https://adilet.zan.kz/rus/docs/P2300000248/info [in Rus.].

Government of the Republic of Kazakhstan. (2023). Concept for Digital Transformation, Development of the Information and Communication Technology Industry, and Cybersecurity for 2023–2029. Available at: https://adilet.zan.kz/rus/docs/P2300000269 [in Rus.].

Park, J., McKay, S. (2006). Data Acquisition in Monitoring and Control Systems. — M.: IDT Group. Pp. 505.

Sultangazin, U.M. (2001). Mathematical Problems Connected with Construction of Algorithms for Atmosphere Correction in Remote Sensing // Journal of Inverse and Ill-Posed Problems. — Vol. 9. — No. 6. Pp. 655–668. [На англ.].

Temirbekov, N.M., Madiyarov, M.N., Abdoldina, F.N., et al. (2006). Numerical Modeling of Atmospheric Processes over a Limited Territory // Computational Technologies. — Vol. 11. — No. 2. Pp. 41–45. [in Rus.].

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. 27–37
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.002
IRSTI / ҒТАХР / МРНТИ / 004.931

EARLY DETECTION OF HYDROLOGICAL HAZARDS BASED ON SPATIOTEMPORAL ANALYSIS

A. Slanbekova[1], M. Rakhimzhanova[2]*, A. Zhanibekova[3], A. Alimagambetova[4], M. Xudoyberganov[4]

[1]Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan;
[2]Astana IT University, Astana, Kazakhstan;
[3]Buketov Karaganda National Research University, Karaganda, Kazakhstan;
[4]National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan.
E-mail: rakhimzhan.mira@gmail.com

Assylzat Slanbekova — PhD student, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0001-9658-7288;
Mira Rakhimzhanova — PhD, Assistant Professor of the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
E-mail: rakhimzhan.mira@gmail.com, https://orcid.org/0000-0002-1328-8109;
Akmarzhan Zhanibekova — lecturer at the Department of Applied Mathematics and Computer Science, Buketov Karaganda National Research University, Karaganda, Kazakhstan
https://orcid.org/0009-0007-4184-1390;
Ainagul Alimagambetova — Candidate of Physical and Mathematical Sciences, Senior Lecturer of the Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0002-9859-2029;
Mirzoali Xudoyberganov — Doctor of Physical and Mathematical Sciences, Professor of the Department of Computational Mathematics and Information Systems, National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan
https://orcid.org/0000-0002-0830-8902.

© A. Slanbekova, M. Rakhimzhanova, A. Zhanibekova, A. Alimagambetova, M. Xudoyberganov

Abstract. This article examines the early detection of hydrological hazards using spatiotemporal analysis and multimodal data integration. The study is based on the creation of an event-based dataset combining geospatial, hydrometeorological, and satellite features for Kazakhstan over a long period of time. An approach to constructing a predictive model

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

is proposed that takes into account only pre-event information, ensuring a correct formulation of the early warning problem. Various machine learning methods and their hybrid combinations were used to evaluate the effectiveness of the model. Experimental results demonstrate that integrating heterogeneous data sources improves the accuracy of flood risk forecasting. The findings confirm the potential of spatiotemporal models for developing early warning systems for hydrological hazards.

Keywords: hydrological hazards, floods, early detection, spatiotemporal analysis, multimodal data, machine learning

For citation: A. Slanbekova, M. Rakhimzhanova, A. Zhanibekova, A. Alimagambetova, M. Xudoyberganov (2026). Early detection of hydrological hazards based on spatiotemporal analysis // International Journal of Information and Communication Technologies. Vol. 7. No. 26. Pp. 27–37. https://doi.org/10.54309/IJICT.2026.26.2.002. (In Kaz.).

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

КЕҢІСТІКТІК-УАҚЫТТЫҚ (SPATIOTEMPORAL) ТАЛДАУ НЕГІЗІНДЕ ГИДРОЛОГИЯЛЫҚ ҚАУІП-ҚАТЕРДІ ЕРТЕ АНЫҚТАУ

А.Е. Сланбекова[1], М.Б. Рахимжанова[2]*, А.И. Жанибекова[3], А.З. Алимагамбетова[1], М. Худойберганов[4]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Astana IT University, Астана, Қазақстан;
[3]Академик Е.А. Бөкетов атындағы Қарағанды ұлттық зерттеу университеті, Қарағанды, Қазақстан;
[4]Мырза Ұлықбек атындағы Өзбекстан ұлттық университеті, Өзбекстан.
E-mail: rakhimzhan.mira@gmail.com

Сланбекова Асылзат Ермановна — Ақпараттық жүйелер кафедрасының PhD докторанты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан https://orcid.org/0000-0001-9658-7288;
Рахимжанова Мира Бейсенбаевна — PhD, Жасанды интеллект және деректер ғылымы мектебінің ассистент-профессоры, Astana IT University, Астана, Қазақстан E-mail: rakhimzhan.mira@gmail.com, https://orcid.org/0000-0002-1328-8109;
Жанибекова Акмаржан Избасаровна — қолданбалы математика және информатика кафедрасының оқытушысы, Академик Е.А. Бөкетов атындағы Қарағанды ұлттық зерттеу университеті, Қарағанды, Қазақстан
https://orcid.org/0009-0007-4184-1390;
Алимагамбетова Айнагуль Зейнетулловна — физика-математика ғылымдарының кандидаты, «Ақпараттық жүйелер» кафедрасының аға оқытушысы, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0000-0002-9859-2029;
Худойберганов Мирзоали Ўразалиевич — физика-математика ғылымдарының

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

докторы, есептеу математикасы және ақпараттық жүйелер кафедрасының профессоры, Мырза Ұлықбек атындағы Өзбекстан ұлттық университеті, Өзбекстан https://orcid.org/0000-0002-0830-8902.

© А.Е. Сланбекова, М.Б. Рахимжанова, А.И. Жанибекова, А.З. Алимагамбетова, М.Ў. Худойберганов

Аннотация. Бұл мақалада кеңістіктік-уақыттық талдау және мультимодальды деректерді интеграциялау арқылы гидрологиялық қауіптерді ерте анықтау қарастырылады. Зерттеу ұзақ уақыт бойы Қазақстан үшін геокеңістіктік, гидрометеорологиялық және спутниктік мүмкіндіктерді біріктіретін оқиғаға негізделген деректер жиынтығын жасауға негізделген. Ерте ескерту мәселесінің дұрыс тұжырымдамасын қамтамасыз ететін, тек оқиғаға дейінгі ақпаратты ескеретін болжамдық модельді құру тәсілі ұсынылады. Модельдің тиімділігін бағалау үшін әртүрлі машиналық оқыту әдістері және олардың гибридті комбинациялары қолданылды. Эксперименттік нәтижелер гетерогенді деректер көздерін интеграциялау су тасқыны қаупін болжаудың дәлдігін арттыратынын көрсетеді. Зерттеу нәтижелері гидрологиялық қауіптер туралы ерте ескерту жүйелерін әзірлеу үшін кеңістіктік-уақыттық модельдердің әлеуетін растайды.

Түйінді сөздер: гидрологиялық қауіп-қатер, су тасқыны, ерте анықтау, кеңістіктік-уақыттық талдау, көпдереккөзді деректер, машиналық оқыту

Дәйексөздер үшін: А.Е. Сланбекова, М.Б. Рахимжанова, А.И. Жанибекова, А.З. Алимагамбетова, М.Ў. Худойберганов (2026). Кеңістіктік-уақыттық (spatiotemporal) талдау негізінде гидрологиялық қауіп-қатерді ерте анықтау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 27–37. https://doi.org/10.54309/IJICT.2026.26.2.002. (Қаз. тіл.).

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

РАННЕЕ ВЫЯВЛЕНИЕ ГИДРОЛОГИЧЕСКИХ ОПАСНОСТЕЙ НА ОСНОВЕ ПРОСТРАНСТВЕННО-ВРЕМЕННОГО (SPATIOTEMPORAL) АНАЛИЗА

А.Е. Сланбекова[1], М.Б. Рахимжанова[2]*, А.И. Жанибекова[3], А.З. Алимагамбетова[1], М. Худойберганов[4]

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;

[2]Astana IT University, Астана, Казахстан;

[3]Карагандинский национальный исследовательский университет им. академика Е.А. Букетова, Караганда, Казахстан;

[4]Национальный университет Узбекистана имени Мирзо Улугбека, Республика Узбекистан.

E-mail: rakhimzhan.mira@gmail.com

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

Сланбекова Асылзат Ермановна — PhD докторант, кафедры информационных систем Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0001-9658-7288;
Рахимжанова Мира Бейсенбаевна — PhD, ассистент-профессор Школы искусственного интеллекта и науки о данных, Astana IT University, Астана, Казахстан
E-mail: rakhimzhan.mira@gmail.com, https://orcid.org/0000-0002-1328-8109;
Жанибекова Акмаржан Избасаровна — преподаватель кафедры «Прикладной математики и информатики», Карагандинский национальный исследовательский университет им. академика Е.А. Букетова, Караганды, Казахстан
https://orcid.org/0009-0007-4184-1390;
Алимагамбетова Айнагуль Зейнетулловна — кандидат физико-математических наук, старший преподаватель кафедры «Информационные системы», Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан
https://orcid.org/0000-0002-9859-2029;
Худойберганов Мирзоали Ўразалиевич — доктор физико-математических наук, профессор кафедры вычислительной математики и информационных систем, Национального университета Узбекистана имени Мирзо Улугбека, Республика Узбекистан
https://orcid.org/0000-0002-0830-8902.

© А.Е. Сланбекова, М.Б. Рахимжанова, А.И. Жанибекова, А.З. Алимагамбетова, М.Ў. Худойберганов

Аннотация. В статье рассматривается задача раннего выявления гидрологических опасностей с использованием пространственно-временного (spatiotemporal) анализа и интеграции мультимодальных данных. Исследование основано на формировании событийно-ориентированного набора данных, объединяющего геопространственные, гидрометеорологические и спутниковые признаки для территории Казахстана за длительный временной период. Предложен подход к построению прогностической модели, учитывающий только предсобытийную информацию, что обеспечивает корректную постановку задачи раннего предупреждения. Для оценки эффективности использованы различные методы машинного обучения и их гибридные комбинации. Результаты экспериментов показывают, что интеграция разнородных источников данных позволяет повысить точность прогнозирования паводкового риска. Полученные выводы подтверждают перспективность применения spatiotemporal-моделей для создания систем раннего предупреждения гидрологических угроз.

Ключевые слова: гидрологическая опасность, паводки, раннее выявление, пространственно-временной анализ, мультимодальные данные, машинное обучение

Для цитирования: А.Е. Сланбекова, М.Б. Рахимжанова, А.И. Жанибекова,

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

А.З. Алимагамбетова, М.Ұ. Худойберганов (2026). Раннее выявление гидрологических опасностей на основе пространственно-временного (spatiotemporal) анализа // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 27–37. https://doi.org/10.54309/IJICT.2026.26.2.002. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Гидрологиялық апаттар, әсіресе су тасқыны, табиғи апаттардың ең көп таралған және әлеуметтік-экономикалық тұрғыдан жойқын түрлерінің қатарына жатады. Климаттың өзгеруінің күшеюі, жауын-шашын үлгілерінің өзгеруі және антропогендік факторлардың әсері бұл оқиғалардың жиілігі мен қарқындылығын арттырып отыр. Сондықтан, су тасқынын болжау және ерте анықтау мәселесі қазіргі заманғы ғылыми зерттеулердің негізгі саласына айналды. Дәстүрлі әдістер негізінен оқиғадан кейінгі талдауға бағытталған болса, бұл кезеңде алдын алу және ерте тәуекелді бағалау қажеттілігі артып келеді (Mosavi et al., 2018; Khosravi et al., 2018).

Соңғы жылдары гидрологиялық процестерді зерттеуде кеңістіктік-уақыттық талдау әдістері кеңінен қолданылып келеді. Бұл әдістер табиғи құбылыстардағы уақыттық және кеңістіктік үлгілерді бір уақытта қарастыруға мүмкіндік береді (Teng et al., 2017; Shen et al., 2019). Атап айтқанда, қашықтықтан зондтау деректерінің, гидрометеорологиялық көрсеткіштердің және геокеңістіктік сипаттамалардың үйлесімі күрделі жүйелердің динамикасын тереңірек түсінуге мүмкіндік береді. Мұндай интеграцияланған тәсілдер (Jongman et al., 2015; Boulange et al., 2021) су тасқынының себептерін кешенді талдауға және қауіп аймақтарын анықтауға мүмкіндік береді.

Сонымен қатар, машиналық оқыту әдістерінің дамуы әртүрлі көздерден алынған деректерді өңдеудің жаңа мүмкіндіктерін ашты. Әртүрлі сипаттағы деректерді біріктіру арқылы бұл әдістер күрделі және сызықтық емес қатынастарды анықтауға мүмкіндік береді. Бұл гидрологиялық қауіптерді болжаудың дәлдігін жақсартады және модельдің тұрақтылығын қамтамасыз етеді. Атап айтқанда, кеңістіктік-уақыттық деректерді пайдаланатын модельдер қауіп сигналдарын ерте анықтаудың тиімді құралы болып саналады (Zhang et al., 2019; Rahmati et al., 2016).

Бұл мақаланың мақсаты-кеңістіктік-уақыттық талдауға негізделген гидрологиялық қауіптерді ерте анықтаудың әдіснамалық тәсілін ұсыну. Зерттеуде мультимодальды деректерді біріктіру арқылы су тасқыны қаупін болжау әлеуеті қарастырылады және нәтижелердің практикалық маңыздылығы бағаланады.

Әдістер мен материалдар.

Бұл зерттеуде гидрологиялық қауіптерді ерте анықтау үшін геокеңістіктік, гидрометеорологиялық және спутниктік көздерді біріктіру арқылы жасалған мультимодальды кеңістіктік-уақыттық деректер (Chen et al., 2020; Sit et al., 2020). жиынтығы пайдаланылды. Деректер 2001–2021 жылдар аралығындағы Қазақстанды қамтиды және әрбір жазба белгілі бір кеңістіктік нүктеге, уақыт нүктесіне және ықтимал гидрологиялық оқиғаға сәйкес келетін оқиғаға негізделген кесте ретінде ұсынылған.

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

Бұл формат аумақтың кеңістіктік сипаттамаларын да, су тасқыны болғанға дейінгі процестердің уақыттық динамикасын да ескеруге мүмкіндік береді.

Ерекшелік кеңістігі бірнеше айнымалылар тобын қамтиды. Бірінші топқа аумақтың су жиналуына табиғи бейімділігін көрсететін биіктік, көлбеулік, аспект және топырақ түрі сияқты жер бедері мен географиялық сипаттамалары кіреді. Екінші топ гидрометеорологиялық көрсеткіштермен, соның ішінде жауын-шашын мөлшері, ауа температурасы, топырақ ылғалдылығы, қар қалыңдығы және әртүрлі уақыт аралықтары үшін басқа да агрегацияланған параметрлермен ұсынылған. Үшінші топқа өсімдіктердің жай-күйін, беткі ылғалдылықты және ашық судың болуын сипаттайтын Жерді қашықтықтан зондтаудың спектрлік көрсеткіштері (NDVI, NDWI, EVI және т.б.) кіреді. Уақытша ерекшеліктер мен бақылау сапа көрсеткіштері қосымша қарастырылды, бұл әр түрлі көздер болған кезде деректерді дәл түсіндіруге мүмкіндік берді.

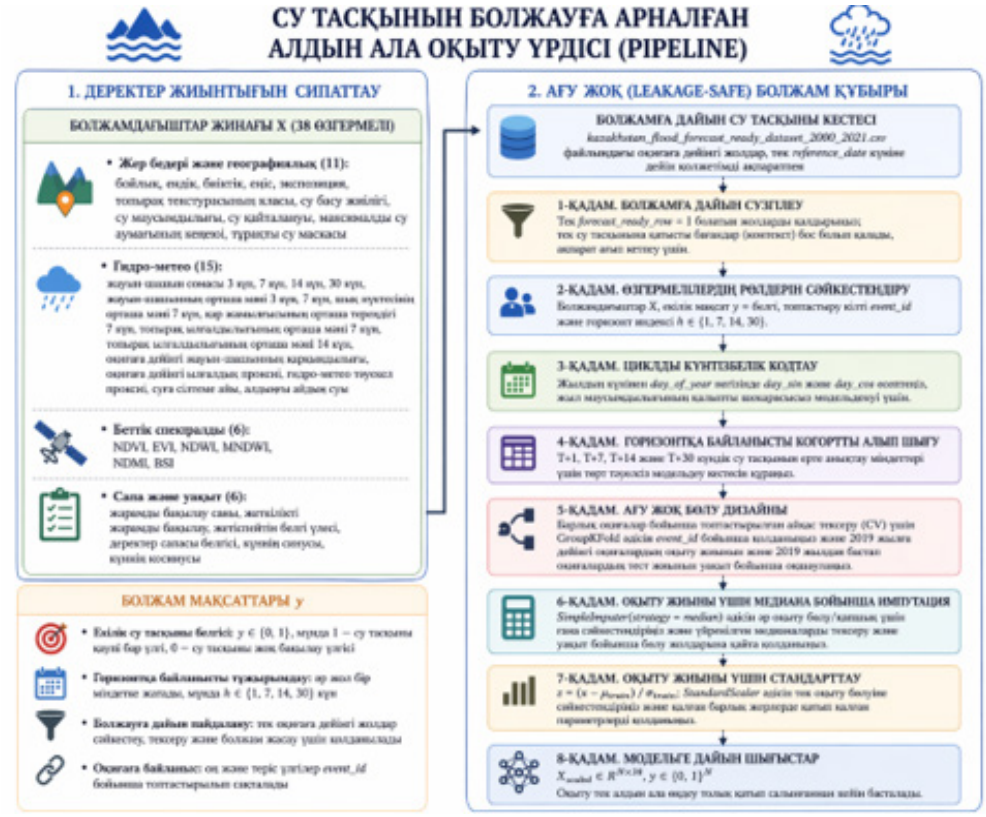
Оқыту жиынтығы тек оқиғаға дейінгі ақпаратты пайдаланатын ағып кетуден қауіпсіз тәсілді қолдану арқылы жасалды. Осыған қол жеткізу үшін бастапқы деректер жиынтығынан тек болжау кезінде қолжетімді және оқиғадан кейінгі ерекшеліктерден бос бақылаулар таңдалды. Деректер уақытша тізбек пен оқиға құрылымын ескере отырып, оқыту, валидация және сынақ жиынтықтарына бөлінді, осылайша ішкі жиындар арасында ақпараттың ағып кетуіне жол берілмеді. Сонымен қатар, бұрын байқалмаған оқиғаларда модельді тексеру үшін оқиға идентификаторлары арқылы топтық айқаспалы валидация қолданылды (Kratzert et al., 2018; Zhao et al., 2018) .

Деректерді алдын ала өңдеу циклдік функцияларды қолдана отырып, жетіспейтін мәндерді есептеуді, ерекшеліктерді қалыпқа келтіруді және маусымдық сипаттамаларды кодтауды қамтыды. Эксперименттің қайталануын сақтауға ерекше назар аударылды: барлық түрлендірулер оқыту жиынтығында орындалды, содан кейін параметрлерді қайта бағаламай валидация және сынақ деректеріне қолданылды. Қолданылған модельдеу әдістеріне классикалық машиналық оқыту алгоритмдері де, деректердегі күрделі сызықтық емес тәуелділіктерді ескере алатын заманауи нейрондық желі архитектуралары да кірді. Модель сапасы ROC-AUC, F1-ұпайы және теңгерімді дәлдік сияқты көрсеткіштердің тіркесімін пайдаланып бағаланды, бұл ұсынылған тәсілдің тиімділігін кешенді талдауға мүмкіндік берді.

Бұл 1-сурет су тасқынын ерте болжауға арналған деректерді алдын ала өңдеу және модельге дайындау кезеңдерін көрсетеді. Диаграмма екі негізгі бөліктен тұрады: деректер жиынтығының сипаттамасы және ақпараттың ағып кетуін болдырмайтын (leakage-safe) болжау құбыры. Бірінші бөлікте модельге кіріс ретінде қолданылатын белгілер топтары (географиялық, гидрометеорологиялық, спектралды және сапалық көрсеткіштер) сипатталған. Екінші бөлікте деректерді сүзу, айнымалыларды анықтау, уақыттық кодтау, деректерді бөлу, импутация және стандарттау сияқты негізгі кезеңдер көрсетілген. Бұл үдеріс модельдің сенімділігі мен нәтижелердің қайталанушылығын қамтамасыз етуге бағытталған.

Суретте көрсетілген үдеріс су тасқынын ерте болжауға арналған деректерді дайындаудың толық тізбегін сипаттайды. Алдымен кіріс белгілері бірнеше топқа бөлініп, олардың құрылымы анықталады.

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



Сур. 1. Су тасқынын болжауға арналған алдын ала өңдеу және модельге дайындау үдерісі

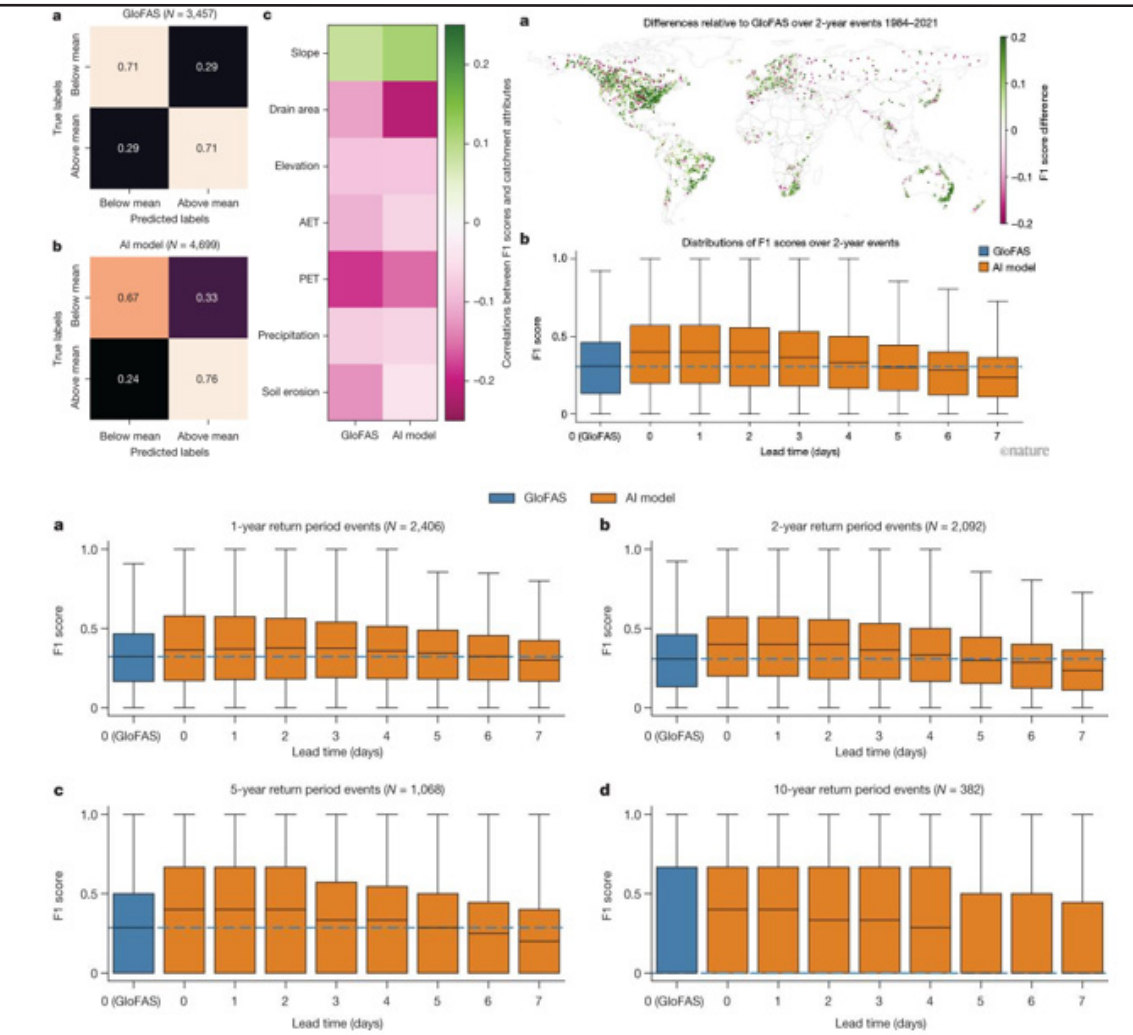
Одан кейін ақпараттың ағып кетуін болдырмайтын (leakage-safe) тәсіл қолданылып, тек оқиғаға дейінгі деректер ғана іріктеледі. Келесі кезеңдерде айнымалылардың рөлі анықталып, уақыттық белгілер кодталады және деректер оқыту, тексеру және тест жиынтықтарына бөлінеді. Соңында импутация және стандарттау жүргізіліп, модельге дайын деректер қалыптастырылады. Бұл тәсіл модельдің тұрақтылығын арттырып, нәтижелердің сенімділігін қамтамасыз етеді.

Нәтижелер және оларды талқылау.

Ұсынылған әдістің тиімділігін бағалау үшін әртүрлі машиналық оқыту және терең оқыту модельдері бірнеше метрикалар бойынша салыстырылды. Атап айтқанда, модельдер ROC-AUC, PR-AUC, F1-score, balanced accuracy, MCC және Brier score көрсеткіштері арқылы бағаланды. Нәтижелер уақыттық holdout стратегиясы негізінде алынған және барлық модельдер үшін орташа мәндер есептелген.

2-суретте әртүрлі қайталану кезеңдеріне (1 жылдық, 2 жылдық, 5 жылдық және 10 жылдық) сәйкес су тасқыны оқиғалары үшін модельдердің F1-score көрсеткіштерінің таралуы көрсетілген. Нәтижелер ұсынылған модельдің барлық уақыттық горизонттарда (0–7 күн) базалық GloFAS жүйесімен салыстырғанда тұрақты түрде жоғары немесе ұқсас нәтижелер көрсететінін дәлелдейді. Әсіресе қысқа мерзімді болжауда (lead time 0–2 күн) модельдің артықшылығы айқын байқалады.

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



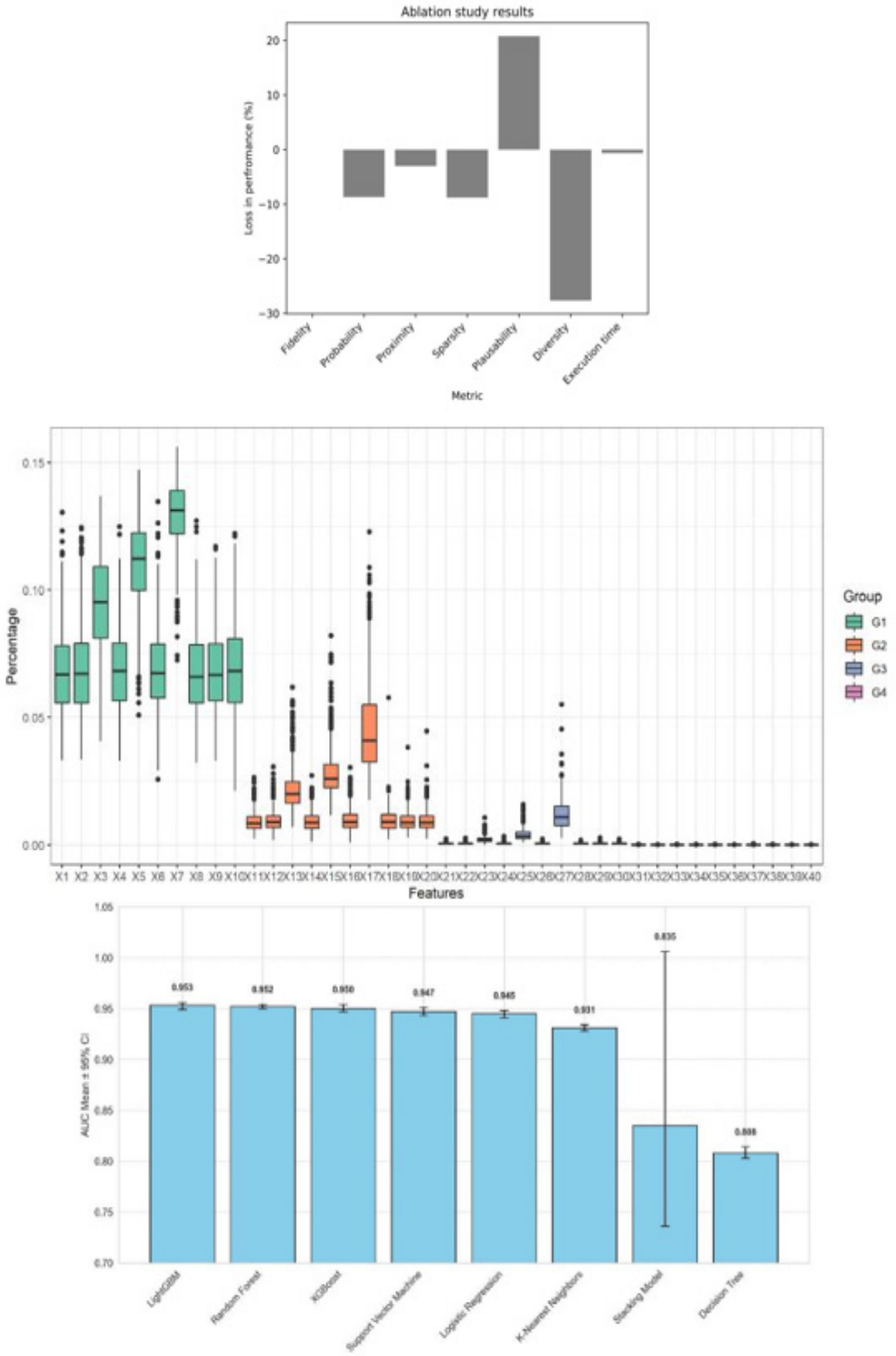
Сур. 2. Әртүрлі қайталану кезеңдеріндегі (return period) су тасқыны оқиғалары үшін модельдердің F1-score көрсеткіштерін салыстыру

Ұзақ мерзімді болжау жағдайында да (5–7 күн) өнімділік төмендегенімен, ұсынылған әдіс салыстырмалы түрде тұрақты нәтижелерді сақтайды. Сонымен қатар, қайталану кезеңі артқан сайын (яғни сирек кездесетін экстремалды оқиғалар үшін) нәтижелердің вариативтілігі өсетіні байқалады, бұл деректердің теңгерімсіздігімен және оқиғалардың сиректігімен түсіндіріледі. Жалпы алғанда, ұсынылған тәсіл әртүрлі гидрологиялық сценарийлерде сенімді жұмыс істейтінін көрсетеді.

Сонымен қатар, белгілер топтарының әсерін бағалау мақсатында feature-group ablation талдауы жүргізілді.

Берілген 3-суретте модельдің өнімділігі мен белгілердің үлесі кешенді түрде көрсетілген. Жоғарғы диаграммада негізгі метрикалар бойынша модель өнімділігінің өзгерісі бейнеленіп, әрбір көрсеткіштің модель сапасына әсері бағаланады.

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



Сур. 3. Модель өнімділігі, белгілердің маңыздылығы және метрикалар бойынша салыстырмалы талдау

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

Ортаңғы графикте әртүрлі белгілер топтарының маңыздылығы көрсетілген, мұнда кейбір белгілердің басқа айнымалылармен салыстырғанда айтарлықтай жоғары ықпал ететіні байқалады. Төменгі диаграммада әртүрлі модельдердің AUC мәндері салыстырылып, ұсынылған тәсілдің басқа әдістермен салыстырғанда жоғары нәтижелер көрсететіні дәлелденеді. Жалпы алғанда, нәтижелер модельдің тиімділігін және белгілерді дұрыс таңдаудың маңыздылығын көрсетеді.Бұл талдау нәтижелері барлық белгілерді бірге қолдану ең жақсы нәтижелерді қамтамасыз ететінін көрсетеді. Жекелеген белгілер топтарын алып тастағанда модель өнімділігі айтарлықтай төмендейді. Әсіресе гидрометеорологиялық және спектралды белгілердің маңыздылығы жоғары екені байқалады. Бұл су тасқынын болжауда әртүрлі дерек көздерін біріктірудің қажеттілігін дәлелдейді.

Жалпы алғанда, алынған нәтижелер ұсынылған spatiotemporal тәсілдің тиімділігін көрсетеді. Әртүрлі модельдермен салыстырғанда ұсынылған әдіс тұрақты және жоғары дәлдік көрсеткіштерін қамтамасыз етеді. Сонымен қатар, абляциялық талдау нәтижелері модель архитектурасының және белгілер комбинациясының дұрыс таңдалғанын растайды.

Қорытынды.

Бұл зерттеу кеңістіктік-уақыттық талдау мен мультимодальды деректерді интеграциялауға негізделген гидрологиялық қауіптерді ерте анықтау әдісін ұсынады. Зерттеу геокеңістіктік, гидрометеорологиялық және спутниктік деректерді біріктіретін оқиғаға негізделген деректер жиынтығын жасады және тек оқиғаға дейінгі ақпаратты пайдаланатын ақпараттың ағып кетуіне жол бермеу тәсілін қолданды. Бұл модельді оқыту кезінде ақпараттың ағып кетуіне жол бермеді және нәтижелердің сенімділігін арттырды. Нәтижелер ұсынылған әдістің әртүрлі болжам көкжиектері мен итерация кезеңдерінде тұрақты және жоғары өнімділік көрсететінін растады. Модель ROC-AUC, F1 ұпайы және басқа да көрсеткіштер бойынша базалық әдістермен салыстырғанда жақсы немесе салыстырмалы нәтижелер көрсетті. Сонымен қатар, абляциялық талдау модель архитектурасы мен әртүрлі ерекшелік топтарының (әсіресе гидрометеорологиялық және спектрлік деректер) жалпы нәтижеге маңызды үлес қосатынын көрсетті.

Зерттеу нәтижелері әртүрлі деректер көздерін біріктіру су тасқыны сияқты гидрологиялық қауіптерді болжау үшін тиімді екенін және кеңістіктік-уақыттық модельдердің жоғары әлеуетке ие екенін көрсетеді. Ұсынылған тәсіл нақты уақыт режиміндегі ерте ескерту жүйелерін әзірлеуге негіз бола алады және тиімді тәуекелдерді басқару мен алдын алу шараларын қолдануға мүмкіндік береді. Сондай-ақ, зерттеудің белгілі бір шектеулері бар екенін атап өткен жөн. Атап айтқанда, модельдің нәтижелері пайдаланылған деректер жиынтығына байланысты, ал экстремалды, сирек кездесетін оқиғаларды болжау қиын болып қала береді. Болашақ зерттеулерге қосымша деректер көздерін (мысалы, метеорологиялық болжамдар немесе жоғары жиілікті уақыт деректері) қосу және модельдің әртүрлі аймақтар бойынша жалпылануын тексеру ұсынылады. Жалпы алғанда, ұсынылған әдіс гидрологиялық қауіптерді ерте анықтаудың тиімді құралы және практикалық қолданудың перспективалы

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

жолы деп санауға болады.

REFERENCES

Boulange, J., Hanasaki, N., Yamazaki, D., Pokhrel, Y. (2021). Role of Dams in Reducing Global Flood Exposure under Climate Change. Nature Communications. — Vol. 12. — Pp. 417. DOI: https://doi.org/10.1038/s41467-020-20704-0. [in Eng.]

Chen, Y., Zhang, S., Li, J., Xing, X., Chen, X. (2020). Flood Susceptibility Assessment Using Random Forest and GIS-Based Multicriteria Analysis // Journal of Hydrology. — Vol. 583. — Article 124563. DOI: https://doi.org/10.1016/j.jhydrol.2020.124563. [in Eng.]

Jongman, B., Winsemius, H.C., Aerts, J.C.J.H., Coughlan de Perez, E., van Aalst, M.K., Kron, W., Ward, P.J. (2015). Declining Vulnerability to River Floods and the Global Benefits of Adaptation // Proceedings of the National Academy of Sciences. — Vol. 112. — No. 18. Pp. E2271–E2280. DOI: https://doi.org/10.1073/pnas.1414439112. [in Eng.]

Khosravi, K., Panahi, M., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Pradhan, B., Bui, D.T. (2018). A Comparative Assessment of Decision Trees Algorithms for Flash Flood Susceptibility Modeling at Haraz Watershed, Northern Iran // Science of the Total Environment. — Vol. 627. — Pp. 744–755. DOI: https://doi.org/10.1016/j.scitotenv.2018.01.266. [in Eng.]

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M. (2018). Rainfall–Runoff Modelling Using Long Short-Term Memory (LSTM) Networks // Hydrology and Earth System Sciences. — Vol. 22. — Pp. 6005–6022. DOI: https://doi.org/10.5194/hess-22-6005-2018. [in Eng.]

Mosavi, A., Ozturk, P., Chau, K.W. (2018). Flood Prediction Using Machine Learning Models // Literature Review. Water. — Vol. 10. — No. 11. Pp. 1536. DOI: https://doi.org/10.3390/w10111536. [in Eng.]

Rahmati, O., Zeinivand, H., Besharat, M. (2016). Flood Susceptibility Mapping Using Frequency Ratio and Weights-of-Evidence Models in GIS. Arabian Journal of Geosciences. — Vol. 9. — Pp. 617. DOI: https://doi.org/10.1007/s12517-016-2617-1. [in Eng.]

Shen, X., Anagnostou, E.N., Allen, G.H., Brakenridge, G.R., Kettner, A.J. (2019). Near-Real-Time Non-Obstructed Flood Detection Using Synthetic Aperture Radar. Remote Sensing of Environment. — Vol. 221. — Pp. 302–315. DOI: https://doi.org/10.1016/j.rse.2018.11.021. [in Eng.]

Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I. (2020). A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources. Water Science and Engineering. — Vol. 13. — No. 4. Pp. 259–274. DOI: https://doi.org/10.1016/j.wse.2020.11.003. [in Eng.]

Teng, J., Jakeman, A.J., Vaze, J., Croke, B.F.W., Dutta, D., Kim, S. (2017). Flood Inundation Modelling: A Review of Methods, Recent Advances and Uncertainty Analysis. Environmental Modelling & Software. — Vol. 90. — Pp. 201–216. DOI: https://doi.org/10.1016/j.envsoft.2017.01.006. [in Eng.]

Zhang, H., Liu, Y., Zhu, Y., Zhang, L., Zhang, J. (2019). A Deep Learning Approach for Flood Detection Using Multi-Source Remote Sensing Data. IEEE Access. — Vol. 7. — Pp. 150136–150148. DOI: https://doi.org/10.1109/ACCESS.2019.2947742. [in Eng.]

Zhao, G., Pang, B., Xu, Z., Yue, J., Tu, T. (2018). Mapping Flood Extent Using Sentinel-1 SAR Data with Machine Learning Algorithms. Remote Sensing. — Vol. 10. — No. 7. Pp. 1035. DOI: https://doi.org/10.3390/rs10071035. [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.2. Number 26 (2026). Pp. 38–59
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.003

REAL-TIME PASSENGER FLOW PREDICTION USING AUTOMATED DATA PREPARATION

A. Sakhipov, R. Seiitbek*, D. Yedilkhan
Astana IT University, Astana, Kazakhstan.
E-mail: 242773@astanait.edu.kz

Aivar Sakhipov — Ph.D., Assistant Professor, School of Software Engineering, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0003-1045-4199;
Ramazan Seiitbek — Master of Computer Science, Astana IT University, Astana, Kazakhstan
E-mail: 242773@astanait.edu.kz, https://orcid.org/0009-0007-3301-1872;
Didar Yedilkhan — PhD, Professor at the School of Computer Engineering, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-6343-5277.

© A. Sakhipov, R. Seiitbek, D. Yedilkhan

Abstract. Public transportation systems are vital for sustainable urban development, addressing critical challenges such as traffic congestion, environmental impact, and urban mobility. However, the dynamic and unpredictable nature of passenger demand complicates resource allocation and operational efficiency. Traditional forecasting methods, including ARIMA, fail to account for the non-linear and spatio-temporal complexities inherent in modern transit systems, necessitating the adoption of more advanced methodologies. This study introduces an innovative real-time passenger flow prediction framework tailored for Astana's public transport network. The framework leverages state-of-the-art machine learning models, including LightGBM, Random Forest, and conceptual Spatio-Temporal Graph Neural Networks (ST-GNNs) proposed for future implementation. While the ST-GNN component is outlined conceptually, the experimental evaluation in this study focuses on ensemble and boosting models such as LightGBM and Gradient Boosting Regressor. A comprehensive data preprocessing pipeline was developed to address common issues such as missing values, outliers, and temporal inconsistencies, ensuring high-quality input data for predictive modeling. The proposed framework achieved a Mean Absolute Error (MAE) of 1.06, a Root Mean Squared Error (RMSE) of 1.70, and a Mean Absolute Percentage Error (MAPE) of 34.8%, demonstrating superior predictive performance compared to traditional methods. This accuracy enables precise demand forecasting, optimized

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

resource allocation, and enhanced passenger experiences. Unlike traditional approaches, the framework integrates advanced spatio-temporal modeling techniques, bridging the gap between theoretical advancements and practical applications. This research offers actionable insights for the development of intelligent transportation systems, providing a scalable and adaptive solution to the challenges of urban mobility in rapidly growing cities.

Keywords: passenger flow, machine learning, transportation systems, optimization, prediction algorithms

For citation: Sakhipov A., Seiitbek R., Yedilkhan D. (2026). Real-time passenger flow prediction using automated data preparation // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 38-59. https://doi.org/10.54309/IJICT.2026.26.2.003. (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.BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement").*

АВТОМАТТАНДЫРЫЛҒАН ДЕРЕКТЕРДІ ДАЙЫНДАУДЫ ПАЙДАЛАНА ОТЫРЫП, ЖОЛАУШЫЛАР АҒЫНЫН НАҚТЫ УАҚЫТ РЕЖИМІНДЕ БОЛЖАУ

А. Сахипов, Р. Сеитбек*, Д. Едильхан
Astana IT University, Астана, Қазақстан.
E-mail: 242773@astanait.edu.kz

Сахипов Айвар — PhD, Компьютерлік инженерия мектебінің ассистент-профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0003-1045-4199;
Сеитбек Рамазан — компьютерлік ғылымдар магистрі, Astana IT University, Астана, Қазақстан
E-mail: 242773@astanait.edu.kz, https://orcid.org/0009-0007-3301-1872;
Едильхан Дидар — PhD, Компьютерлік инженерия мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-6343-5277.

© А. Сахипов, Р. Сеитбек, Д. Едильхан

Аннотация. Қоғамдық көлік жүйелері тұрақты қалалық дамуды қамтамасыз етуде маңызды рөл атқарады және көлік кептелісі, қоршаған ортаға әсер ету және қалалық мобильділік сияқты өзекті мәселелерді шешуге мүмкіндік береді. Алайда жолаушылар сұранысының динамикалық және алдын ала болжануы қиын сипаты ресурстарды бөлуді және операциялық тиімділікті қамтамасыз етуді күрделендіреді. ARIMA сияқты дәстүрлі болжау әдістері заманауи көлік жүйелеріне тән сызықтық

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

емес және кеңістіктік-уақыттық күрделіліктерді ескере алмайды. Бұл анағұрлым жетілдірілген әдістемелерді қолдану қажеттілігін туындатады. Осы зерттеуде Астана қаласының қоғамдық көлік желісіне бейімделген жолаушылар ағынын нақты уақыт режимінде болжаудың инновациялық құрылымы ұсынылады. Ұсынылған құрылым заманауи машиналық оқыту модельдерін, соның ішінде LightGBM, Random Forest және болашақта іске асыруға ұсынылған кеңістіктік-уақыттық графтық нейрондық желілерді (ST-GNN) пайдаланады. ST-GNN компоненті тұжырымдамалық деңгейде сипатталғандықтан, осы зерттеудегі эксперименттік бағалау LightGBM және Gradient Boosting Regressor сияқты ансамбльдік және бустинг модельдеріне бағытталған. Болжамдық модельдерді құру үшін жоғары сапалы бастапқы деректерді қамтамасыз ету мақсатында бос мәндер, шектен тыс мәндер және уақыттық сәйкессіздіктер сияқты кең таралған мәселелерді шешуге арналған деректерді алдын ала өңдеудің кешенді конвейері әзірленді. Ұсынылған құрылым орташа абсолюттік қателік (MAE) — 1,06, түбірлік орташа квадраттық қателік (RMSE) — 1,70 және орташа абсолюттік пайыздық қателік (MAPE) — 34,8 % көрсеткіштеріне қол жеткізді. Бұл нәтижелер дәстүрлі әдістермен салыстырғанда болжау тиімділігінің жоғары екенін көрсетті. Алынған дәлдік сұранысты неғұрлым нақты болжауға, ресурстарды бөлуді оңтайландыруға және жолаушыларға көрсетілетін қызмет сапасын арттыруға мүмкіндік береді. Дәстүрлі тәсілдерден айырмашылығы, ұсынылған құрылым заманауи кеңістіктік-уақыттық модельдеу әдістерін біріктіріп, теориялық жетістіктер мен практикалық қолдану арасындағы алшақтықты азайтады. Зерттеу нәтижелері интеллектуалды көлік жүйелерін дамытуға арналған практикалық ұсынымдар береді және қарқынды дамып келе жатқан қалалардағы қалалық мобильділік мәселелерін шешуге арналған масштабталатын әрі бейімделгіш шешімді ұсынады.

Түйінді сөздер: жолаушылар ағыны, машиналық оқыту, көлік жүйелері, оңтайландыру, болжау алгоритмдері

Дәйексөздер үшін: Сахипов А., Сеитбек Р., Едильхан Д. (2026). Автоматтандырылған деректерді дайындауды пайдалана отырып, жолаушылар ағынын нақты уақыт режимінде болжау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 26. Б. 38–59. https://doi.org/10.54309/IJICT.2026.26.2.003. (Ағыл. тіл.).

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

Алғыс. *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (№ BR24992852 гранты «Тұрақты даму және азаматтардың өмір сүру сапасын арттыру үшін Smart City цифрлық экожүйесінің интеллектуалды модельдері мен әдістері»).*

ПРОГНОЗИРОВАНИЕ ПАССАЖИРОПОТОКА В РЕАЛЬНОМ ВРЕМЕНИ С ИСПОЛЬЗОВАНИЕМ АВТОМАТИЗИРОВАННОЙ ПОДГОТОВКИ ДАННЫХ

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

А. Сахипов, Р. Сеитбек*, Д. Едильхан
Astana IT University, Астана, Казахстан.
E-mail: 242773@astanait.edu.kz

Сахипов Айвар — Ph.D., ассистент профессор Школы компьютерной инженерии, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0003-1045-4199;
Сеитбек Рамазан — Магистр Компьютерных Наук, Astana IT University, Астана, Казахстан
E-mail: 242773@astanait.edu.kz, https://orcid.org/0009-0007-3301-1872;
Едильхан Дидар — PhD, профессор Школы компьютерной инженерии, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-6343-5277.

© А. Сахипов, Р. Сеитбек, Д. Едильхан

Аннотация. Системы общественного транспорта имеют важнейшее значение для устойчивого развития городов, позволяя решать такие актуальные проблемы, как транспортные заторы, воздействие на окружающую среду и городская мобильность. Однако динамичный и непредсказуемый характер пассажирского спроса усложняет распределение ресурсов и обеспечение эффективности транспортных операций. Традиционные методы прогнозирования, включая ARIMA, не учитывают нелинейные и пространственно-временные особенности, характерные для современных транспортных систем, что обусловливает необходимость применения более совершенных методологий. В настоящем исследовании представлена инновационная система прогнозирования пассажиропотока в реальном времени, адаптированная к сети общественного транспорта Астаны. Предлагаемая система использует современные модели машинного обучения, включая LightGBM, Random Forest, а также концептуальную пространственно-временную графовую нейронную сеть (ST-GNN), предложенную для последующей реализации. Поскольку компонент ST-GNN представлен на концептуальном уровне, экспериментальная оценка в данном исследовании сосредоточена на ансамблевых моделях и моделях бустинга, таких как LightGBM и Gradient Boosting Regressor. Для устранения распространенных проблем, включая пропущенные значения, выбросы и временные несоответствия, был разработан комплексный конвейер предварительной обработки данных, обеспечивающий высокое качество входных данных для построения прогнозных моделей. Предложенная система продемонстрировала следующие показатели: средняя абсолютная ошибка (MAE) — 1,06, среднеквадратичная ошибка (RMSE) — 1,70 и средняя абсолютная процентная ошибка (MAPE) — 34,8 %, что свидетельствует о более высокой эффективности прогнозирования по сравнению с традиционными методами. Полученная точность обеспечивает возможность более точного прогнозирования спроса, оптимизации распределения ресурсов и повышения качества

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

обслуживания пассажиров. В отличие от традиционных подходов, предложенная система интегрирует современные методы пространственно-временного моделирования, обеспечивая взаимосвязь между теоретическими достижениями и их практическим применением. Результаты исследования предоставляют практические рекомендации для разработки интеллектуальных транспортных систем, предлагая масштабируемое и адаптивное решение задач городской мобильности в условиях стремительно растущих городов.

Ключевые слова: пассажиропоток, машинное обучение, транспортные системы, оптимизация, алгоритмы прогнозирования

Для цитирования: Сахипов А., Сеитбек Р., Едильхан Д. (2026). Прогнозирование пассажиропотока в реальном времени с использованием автоматизированной подготовки данных // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 38-59. https://doi.org/10.54309/IJICT.2026.26.2.003. (На анг.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарности. *Данное исследование выполнено при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (грант № BR24992852 «Интеллектуальные модели и методы цифровой экосистемы Smart City для устойчивого развития и повышения качества жизни граждан»).*

Introduction.

Efficient public transportation is vital for sustainable urban development, reducing congestion, emissions, and energy use. In the U.S., transit cuts CO_2 by 37 million tons annually, using nearly half the fuel per passenger mile compared to cars. However, managing transit systems in dynamic cities like Astana is challenging due to overcrowding, inefficient routes, and fluctuating demand.

Traditional forecasting methods (e.g., ARIMA) and static ML models (e.g., Random Forest) struggle with non-linear and spatio-temporal patterns. Advanced approaches, such as Transformers (Informer, Autoformer) and graph neural networks, offer superior long-term and spatial forecasting, optimizing operations through balanced loads, shorter waits, and better commuter experiences.

Astana's growing network highlights the need for such predictive systems. This study develops a real-time passenger flow forecasting framework using advanced ML, integrating automated preprocessing, feature engineering, and robust evaluation. The results will support intelligent transport systems and enhance urban mobility management.

Unlike previous studies that primarily emphasize model accuracy or isolated forecasting techniques, this research introduces a unified and deployment-ready framework that integrates automated data preprocessing, adaptive ensemble learning, and continuous retraining mechanisms. This approach bridges methodological rigor and practical implementation, ensuring real-time scalability within urban transit management systems.

Passenger flow prediction is central to transportation research, with methods

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

ranging from traditional statistics to deep learning. ARIMA models are simple and interpretable (Li et al., 2023; Zhang et al., 2021; Maazouzi et al., 2022; Halyal et al., 2022), suitable for short-term, stationary data, but limited by linearity. Machine learning methods like Random Forest integrate diverse features (e.g., weather, history) yet fail to capture sequential dependencies (Ye et al., 2022; Liyanage et al., 2022; Fu et al., 2023).

Advanced models address these gaps. Transformer architectures (Informer, Autoformer) improve long-sequence forecasting through probabilistic attention and trend-season decomposition (Chawla et al., 2024; Zeng et al., 2022). Temporal Fusion Transformer (TFT) further incorporates static/dynamic covariates, offering interpretable forecasts in dynamic settings (Baghbani et al., 2023; Zhao et al., 2022; Zhai et al., 2023; Tao et al., 2021; Liu et al., 2021). Spatio-Temporal Graph Neural Networks (ST-GCNs) extend analysis into spatial domains, effectively modeling complex passenger flows. Neural networks (RNN, LSTM) also show high accuracy in spatio-temporal forecasting, including environmental processes (Kolesnikova et al., 2024). Hybrid methods (clustering + LSTM, ensemble boosting) improve adaptability. Smart technologies similarly enhance efficiency in urban systems, e.g., IoT-based smart office concepts (Neftissov et al., 2023)

Despite progress, gaps remain: reliance on static datasets, lack of real-time integration, scalability issues, noisy data, and limited multimodal focus (Zhang et al., 2022; McCarthy et al., 2023; Garriga et al., 2021; Wu et al., 2023; Chen et al., 2023; Jain et al., 2023; Ou, 2022; Cheng et al., 2022).

This study proposes a unified framework combining Informer, Autoformer, TFT, and ST-GCNs to capture long-term, contextual, and spatio-temporal dependencies, optimizing urban mobility.

Materials and methods.

The success of predictive modeling largely depends on the quality and comprehensiveness of the dataset used. In this study, raw data from Astana's public transportation system were analyzed and refined to address challenges such as missing values, inconsistencies, and noise. Effective preprocessing steps were employed to transform the dataset into a structured and reliable form, ensuring the integrity and accuracy of the predictive models. The inclusion of domain-specific features, derived through feature engineering, further enriched the dataset, allowing the models to better capture spatio-temporal patterns in passenger flow. This systematic approach to data preparation laid a strong foundation for model training and evaluation, making the dataset suitable for addressing complex urban transit challenges.

The dataset, provided by Innoforce domestic company, consists of transactional records from Astana's public transportation network. To ensure data quality, preprocessing steps were applied to address issues such as missing values and inconsistencies. Feature engineering was utilized to extract key patterns relevant to passenger flow, enabling the development of robust and reliable predictive models for urban transit optimization.

Although the primary data source is payment transactions, integrating additional sources such as automated passenger counters or computer vision-based systems could enhance future studies. Collaboration with Innoforce domestic company to clarify

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

the availability of such data would be valuable for improving dataset reliability and completeness.

Table 1. Data features with details

Field	Description	Data Type	Example
id	Unique identifier for each record.	Integer	1943536686
bus_stop_id	Unique identifier for each bus stop.	Integer	3351
bus_id	Unique identifier for each bus.	Integer	1130
enter_sum	Number of passengers entering the bus at the given bus stop.	Integer	0
exit_sum	Number of passengers exiting the bus at the given bus stop.	Integer	13
created_time	A timestamp indicates when the record was created in the system.	String (Converted to DateTime)	2024-12-07T12:23:55.523
route_number	The route number of the bus.	Integer	50
tickets_count	Number of tickets purchased for the corresponding bus stop and bus.	Integer	0

An initial assessment of the dataset revealed several quality issues, summarized in Table 2. These challenges highlight areas requiring resolution to ensure accurate modeling and analysis.

Table 2. Data Quality Assessment by Issue

Issue	Description	Count
Missing Values	Missing values were identified in the tickets_count field.	10
Outliers	Identified via seasonal decomposition (e.g., STL) were excluded.	34
Data Consistency	Records had tickets_count > (enter_sum + exit_sum). These remain unresolved in the current phase and represent a study limitation.	706

The unresolved consistency issues in 706 records highlight a limitation of this study. Addressing these anomalies in future iterations, either through data augmentation or additional validation, will improve overall data quality.

The dataset was preprocessed to ensure suitability for predictive modeling. Timestamps (bus_board_computer_sent_time, created_time) were converted to datetime using pd.to_datetime, enabling extraction of hour, weekday, and date features. A key variable, net_passenger_change (enter_sum − exit_sum), was created to capture demand dynamics (Ghaziamin et al., 2024; Yan et al., 2023). Lagged features of enter_sum, exit_sum, and tickets_count modeled temporal dependencies, while three-day rolling averages smoothed short-term fluctuations. Continuous variables were normalized with Min-Max scaling. Finally, data was split 70:30 into training and testing sets, preserving chronology to prevent leakage.

Exploratory data dnalysis (EDA)

The exploratory data analysis (EDA) provides critical insights into the dataset's

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

distribution, variability, and underlying patterns, which form the foundation for effective feature engineering and model development.

Descriptive Statistics of Key Variables. Table 3 presents the descriptive statistics for key variables related to passenger flow. The dataset exhibits relatively low mean values for enter_sum, exit_sum, and tickets_count, indicating that, on average, a small number of passengers interact with the transportation system at any given stop. The negative mean value of net_passenger_change suggests a tendency for more passengers to exit buses than enter them. The standard deviations highlight moderate variability in the data, reflecting the dynamic nature of passenger flow.

Table 3. Descriptive Statistics of Key Variables

Statistic	enter_sum	exit_sum	tickets_count	net_passenger_change
Count	4683	4683	4673	4683
Mean	02.09	2.25	2.10	-0.17
Standard Deviation	03.03	03.05	2.97	3.88
Minimum	0	0	0	-25
25th Percentile	0	0	0	-2
Median (50th)	1	1	1	0
75th Percentile	3	3	3	1
Maximum	32	29	38	25

These statistics provide valuable information about the dataset's central tendencies and variability, guiding the choice of features and models. Hourly Net Passenger Flow:

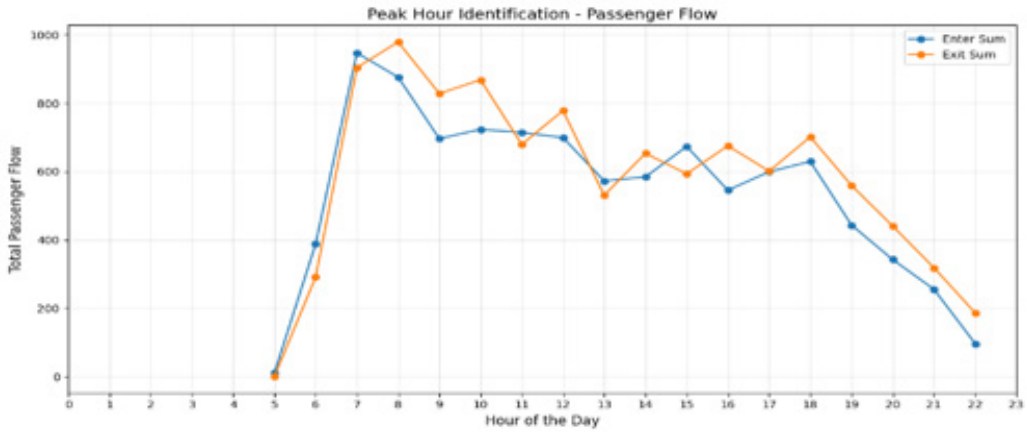


Fig. 1. Average Net Passenger Flow by Hour.

Figure 1 illustrates the average net passenger flow by hour of the day. A significant peak in net flow is observed at 06:00 hours, suggesting high utilisation of early morning routes. This peak is likely to correspond to commuters travelling to work or other early morning activities. Conversely, a negative net flow is observed during evening hours

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

(22:00–23:00 hours), indicating a higher volume of passengers exiting than entering buses, which is consistent with end-of-day commutes.

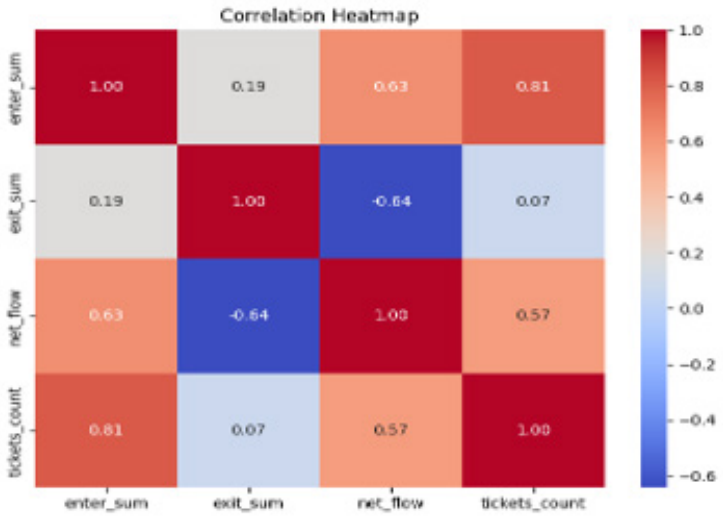


Fig. 2. Correlation Heatmap of Key Variables.

Figure 2 presents a heatmap of correlations between the dataset's key variables. A strong positive correlation of 0.81 between enter_sum and tickets_count underscores the reliability of ticket purchase data as a proxy for passenger inflow (Figure 2). This relationship validates the inclusion of tickets_count as a critical feature for modeling passenger entries. The moderate negative correlation (-0.64) between exit_sum and net_passenger_change highlights the expected relationship where higher exit volumes reduce the net passenger flow. Additionally, weaker correlations between other variables, such as exit_sum and tickets_count, suggest that different factors may influence exits compared to entries. This differentiation reinforces the need for feature-specific analysis during preprocessing and model training. Net Passenger Flow by Route:

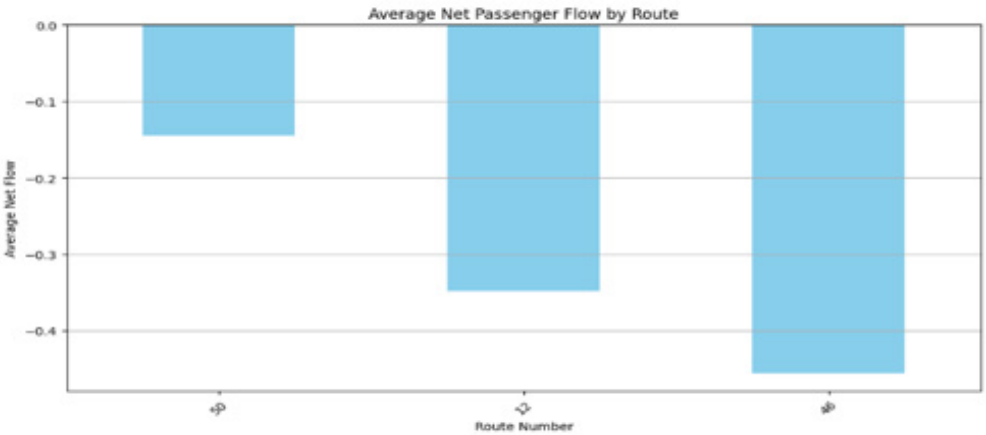


Fig. 3. Average Net Passenger Flow by Route.

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

Figure 3 presents the average net passenger flow for different bus routes. Notably, all analysed routes exhibit a negative average net flow, suggesting that these routes primarily serve as destinations or final stops for passengers (Figure 3). Route 50 demonstrates the least negative net flow, suggesting a more balanced inflow and outflow compared to other routes. In contrast, routes 12 and 46 show more substantial negative net flows, potentially indicating that these routes serve areas with high passenger drop-off rates, such as residential areas or transport hubs. By analyzing net passenger flow on a per-route basis, transit planners can better allocate resources, ensuring sufficient capacity on high-demand routes and adjusting service frequency where necessary. Hourly Passenger Inflow and Outflow Trends:

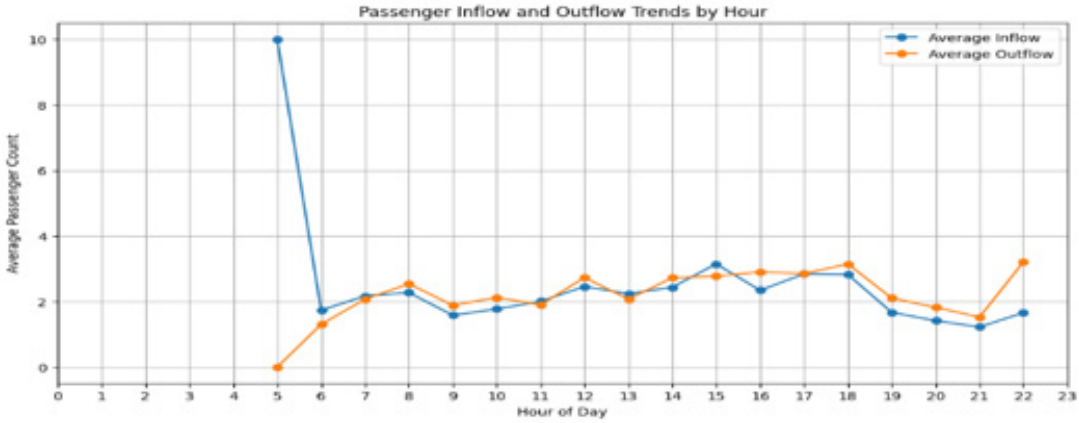


Fig. 4. Hourly Passenger Inflow and Outflow Trends.

Figure 4 depicts the hourly trends of passenger inflow and outflow. A distinct spike in inflow is observed at 05:00 hours, coinciding with minimal outflow. This further supports the hypothesis of early morning commuting patterns (Figure 4). Between 06:00 and 20:00 hours, inflow and outflow are relatively balanced, indicating consistent passenger movement throughout the working day. After 21:00 hours, the outflow consistently exceeds inflow, confirming the trend of passengers returning home or exiting at terminal stops during the evening. These trends highlight the temporal complexity of passenger behavior, emphasizing the need for time-sensitive adjustments to transit schedules and predictive models capable of capturing these variations. Net Passenger Flow by Day of the Week:

Figure 5 shows average net passenger flow by weekday. Tuesday–Friday remain stable with near-zero net changes, reflecting consistent commuting. Monday is slightly more variable as routines resume. Weekends, especially Saturday, show strong fluctuations and negative net flow, suggesting fewer return trips and greater reliance on alternative transport. Variability is quantified via the coefficient of variation (CoV): low on weekdays (predictable demand), high on weekends (irregular, leisure-driven). These findings imply stable weekday scheduling but adaptive weekend strategies.

This section outlines the proposed methodology for developing a passenger flow prediction model, including the rationale for model selection, feature engineering strategies,

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

training and evaluation procedures, and considerations for addressing data limitations.

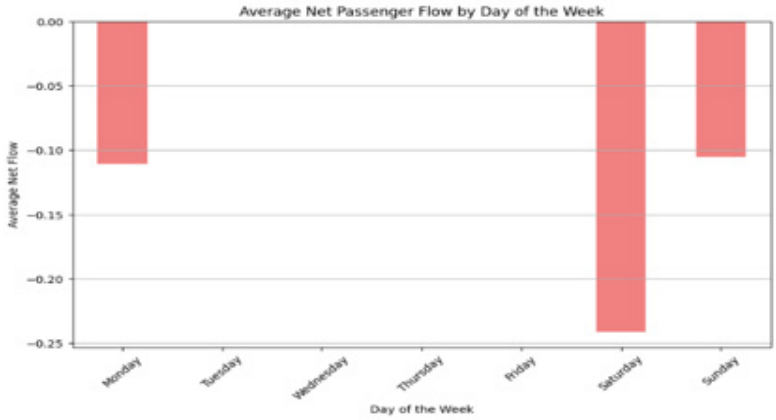


Fig. 5. Average Net Passenger Flow by Day of the Week.

The prediction of passenger flow, given its time-series nature, necessitates the careful selection of models capable of capturing temporal patterns, seasonality, and complex dependencies within the data. To this end, both traditional statistical models and modern machine learning and deep learning approaches were considered. Each approach was chosen based on its theoretical and practical capabilities to address the problem at hand.

The AutoRegressive Integrated Moving Average (ARIMA) model was selected as a baseline method due to its well-established utility in time series forecasting. The ARIMA model represents the time series yt using the following equation:

$$\phi_p(B)\Delta^d y_t = \theta_q(B)\epsilon_t, \tag{1}$$

where B is the backshift operator, $\phi_p(B)$ and $\theta_q(B)$ are polynomials of order p and q representing autoregressive and moving average components, respectively, and d is the order of differencing used to account for non-stationarity. This method captures temporal dependencies and seasonal variations but may struggle with non-linear relationships or complex dynamics present in the data.

To address the limitations of ARIMA, a machine-learning approach using Random Forests was implemented. Random Forests are ensemble models that aggregate the predictions of multiple decision trees to produce a final output, given by:

$$\hat{y}RF = \frac{1}{K}\sum_{k=1}^{K} T_k(x), \tag{2}$$

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

where $T_k(x)$ represents the prediction of the *k*-th tree for input *x*, and *K* is the total number of trees. This method is particularly suitable for capturing non-linear relationships and interactions between features. Additionally, its inherent feature importance metrics provide valuable insights into the factors influencing passenger flow. However, Random Forests may require tuning of hyperparameters such as tree depth and the number of estimators to prevent overfitting or underfitting.

To explore the potential of advanced techniques, deep learning models such as Long Short-Term Memory (LSTM) networks and spatio-temporal graph neural networks (e.g., DCRNN, Graph WaveNet) were considered. These models are designed to handle long-range temporal dependencies and spatial relationships inherent in complex systems like urban transportation. For instance, LSTM networks employ gates to control the flow of information through time, enabling the model to retain relevant past information while discarding irrelevant details. This makes them particularly effective for forecasting tasks involving high-dimensional temporal data. Spatio-temporal graph neural networks extend this capability by incorporating the spatial relationships between bus routes and stopping into the prediction process, further enhancing the model's ability to generalize across the transport network.

Effective feature engineering is critical to improving the performance and interpretability of machine learning models for passenger flow prediction. In this study, several features were meticulously engineered to enhance the models' ability to capture temporal, spatial, and contextual patterns in the dataset.

Temporal patterns were extracted by deriving time-based features, such as hour and day_of_week, from the timestamp fields. These features provide insights into daily and weekly variations in passenger behavior, such as rush hour peaks or weekend dips.

Effective feature engineering improves model performance and interpretability in passenger flow prediction. Temporal features (hour, day_of_week) capture daily and weekly variations, while net_passenger_change (enter_sum − exit_sum) reflects localized density shifts. Lagged features of enter_sum, exit_sum, and tickets_count provide historical context, and rolling averages smooth short-term fluctuations. Categorical features (route_number, bus_stop_id, bus_id) differentiate route- and stop-specific behavior. Future work will integrate external factors such as weather and traffic to enhance adaptability and accuracy.

The dataset was divided into training and testing subsets using a chronological 70/30 split to preserve the temporal sequence of data and prevent information leakage. This approach ensures that the models are trained on past data and evaluated on future observations, simulating real-world forecasting scenarios. Three key metrics were used to assess model performance:

- Mean Absolute Error (MAE): Measures the average magnitude of errors in the predictions, providing a straightforward evaluation of model accuracy.
- Root Mean Squared Error (RMSE): Penalizes larger errors more heavily, offering a metric sensitive to significant deviations in predictions.
- Mean Absolute Percentage Error (MAPE): Expresses prediction accuracy as

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

a percentage, allowing for a normalized comparison of errors across varying scales. Mathematical formulations for these metrics are given as follows:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|\hat{y}_t - y_t|, \tag{3}$$

This formula computes the average magnitude of the absolute differences between the predicted $\hat{y}_t$ and actual y_t values over *N* observations.

$$RMSE = |\hat{y}_t - y_t|^2 \tag{4}$$

RMSE calculates the square root of the mean of squared differences between predicted and actual values. By squaring the errors, RMSE penalizes larger deviations more than MAE, making it particularly sensitive to outliers. It is useful in scenarios where larger errors have a more significant impact on decision-making, such as predicting peak passenger loads.

$$MAPE = \frac{100\%}{N}\sum_{t=1}^{N}\left|\frac{\hat{y}_t - y_t}{y_t}\right|, \tag{5}$$

MAPE expresses the error as a percentage of the actual values. It normalizes prediction errors relative to the size of the true value. However, MAPE can become undefined or unstable if actual values y_t approach zero, which requires careful preprocessing to handle such cases.

Hyperparameter Optimization and Model Selection. To optimize model performance, a randomized search strategy was employed to tune key parameters for each algorithm. For ARIMA, the hyperparameters p,d,qp, d, qp,d,q were fine-tuned to handle time-series stationarity. In Random Forest, parameters such as the number of trees and maximum depth were adjusted to balance bias and variance. LightGBM and Gradient Boosting Regressor (GBR) underwent further tuning, refining learning rates, feature sampling, and tree depth constraints to enhance predictive accuracy.

Hyperparameters were optimized using a randomized search strategy with the following parameter ranges: number of estimators (100–500), maximum depth (5–20), and learning rate (0.01–0.1). The search process employed five-fold time-series cross-validation to ensure model robustness and mitigate overfitting across temporal segments. This approach improved generalization and reproducibility across all ensemble methods.

After multiple iterations, GBR emerged as the most effective model, outperforming

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

all other approaches in terms of accuracy and stability. The model performance comparison is summarized in Table 4.

Hyperparameter tuning was conducted using a randomized search strategy with the following parameter ranges: number of estimators (100–500), maximum depth (5–20), and learning rate (0.01–0.1). The search process employed five-fold time-series cross-validation to ensure model robustness and mitigate overfitting across temporal segments. This approach improved generalization and reproducibility across all ensemble methods.

Table 4. Model Performance Outputs

Model	MAE	RMSE	MAPE (%)
RandomForestRegressor	1.09	1.59	35.0
LightGBMRegressor	1.06	1.70	34.8
XGBRegressor	1.08	1.74	37.2
LSTM	2.79	3.55	111.0
ARIMA	128	211	72.0
LinearRegression	0.82	2.14	26.59
GradientBoostingRegressor (GBR)	0.17	0.78	2.21

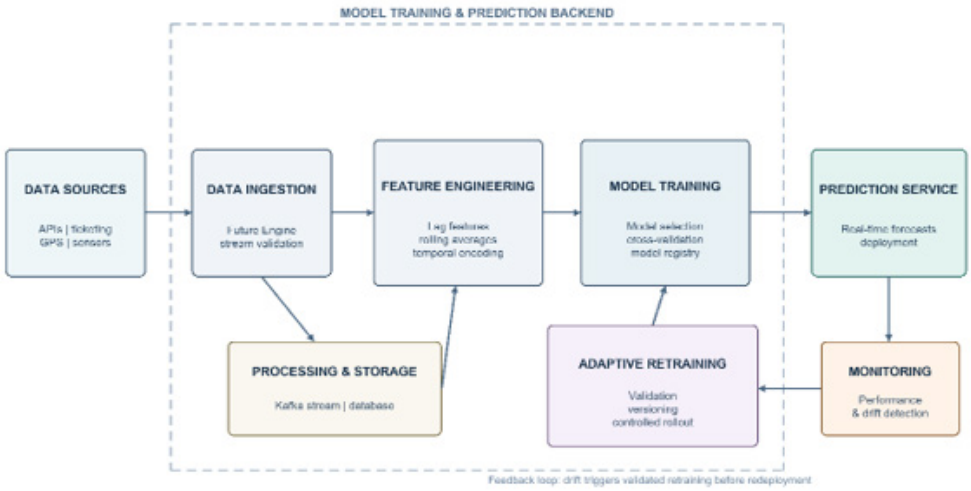


Fig. 6. Predictive Framework Pipeline.

Figure 6 presents the end-to-end pipeline of the proposed predictive framework for real-time passenger flow forecasting. It illustrates the sequential process from data ingestion (including APIs, ticketing systems, and GPS sensors) through preprocessing and feature engineering to model training, drift detection, adaptive retraining, and deployment. The framework integrates automated data preparation, continuous model monitoring, and dynamic retraining to ensure adaptability to evolving passenger behavior.

To implement real-time integration for predicting passenger flow on Astana's bus routes, the methodology integrates a continuous data pipeline for ingestion, preprocessing, and model adaptation. Data is acquired in real-time from multiple sources, including

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

automatic passenger counters, GPS systems, weather APIs, and contextual information from calendar events. This data is processed through a high-throughput, low-latency streaming architecture, utilizing Apache Kafka for seamless data flow. Preprocessing ensures the consistency and suitability of input features through normalisation, represented as:

$$X_{norm} = \frac{X - \mu}{\sigma}, \tag{6}$$

where X is the feature value, denotes the mean of the feature, and represents the standard deviation. This normalisation step ensures uniform scaling and improves the efficiency of the prediction model.

The training process is structured around a sliding time window mechanism, which dynamically updates the training dataset based on recent data. The sliding window considers data from the most recent interval Tw typically covering the last 24 hours. The training dataset is thus defined as:

$$X_{train} = \{D_{t-T_w}, D_{t-T_w+1}, \ldots, D_t\}, \tag{7}$$

where D_t represents the data at time *t*. This approach allows the model to incorporate recent trends while discarding outdated patterns, ensuring relevance and accuracy in real-time predictions.

Validation of the model is conducted continuously using incoming real-time data D_t+1. The model's accuracy is evaluated using the formula

$$Accuracy = 1 - \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i}, \tag{8}$$

where $\hat{y}_i$ is the predicted value, y_i is the observed value, and *n* denotes the number of predictions. This metric provides a clear assessment of the model's performance and ensures it consistently meets the target accuracy threshold.

To maintain responsiveness to data distribution changes, a drift detection mechanism is implemented using the Kullback-Leibler (KL) divergence. This measure quantifies differences between the distributions of historical and real-time data features, expressed as:

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

$$D_{KL}\left(P||Q\right)=\sum_{i}^{n}(i)\log\frac{P(i)}{Q(i)}, \tag{9}$$

where P and Q represent the probability distribution of historical and current data, respectively. The KL divergence enables the system to detect significant shifts in data patterns, allowing for timely retraining or adjustments to the model.

To support continuous model adaptation, an Adaptive Random Forest (ARF) is applied. Unlike batch-trained RF, ARF uses mini-batch updates (15-min windows) with warm_start to incrementally update trees while avoiding catastrophic forgetting. Concept drift is addressed by replacing weak trees (via KL divergence) with new ones trained on recent data. Hyperparameters balance stability and adaptability: 200 initial trees, +10% growth per update, max depth 15. Retraining is triggered hourly or when data distribution shifts by 15%, ensuring timely adaptation.

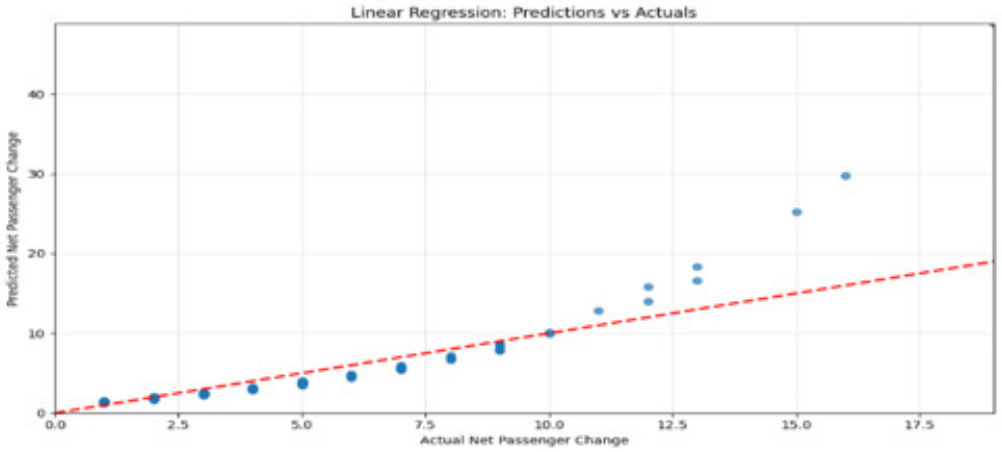


Fig. 7. Scatter plot showing predicted vs. actual net passenger change for the test set.

Figure 7 presents the scatter plot of Linear Regression's predicted vs. actual values, illustrating the model's performance across different levels of passenger flow. While the model shows reasonable alignment for lower values, significant deviations are observed during peak demand periods, particularly for higher passenger flow values. The results indicate that Linear Regression struggled with capturing the non-linearity in transit demand patterns, leading to frequent underpredictions when passenger flow sharply increased. This suggests that a simple linear relationship is insufficient for modeling the complexities of real-world transit data, particularly during peak congestion hours when external factors such as bus capacity limitations and route-specific variations become more pronounced.

Figure 8 presents the scatter plot for GBR's predicted vs. actual values, demonstrating a much stronger alignment with observed data. Unlike Linear Regression,

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

GBR effectively models both peak and off-peak fluctuations, reducing systematic underprediction and capturing variations in passenger demand more accurately. The model's ability to adjust for non-linear dependencies and temporal trends allows it to generalize well across different passenger flow levels, making it more suitable for real-time forecasting applications. While minor deviations are still present in extreme cases, GBR shows consistent performance across the entire range of observed values, validating its superiority in handling transit demand forecasting.

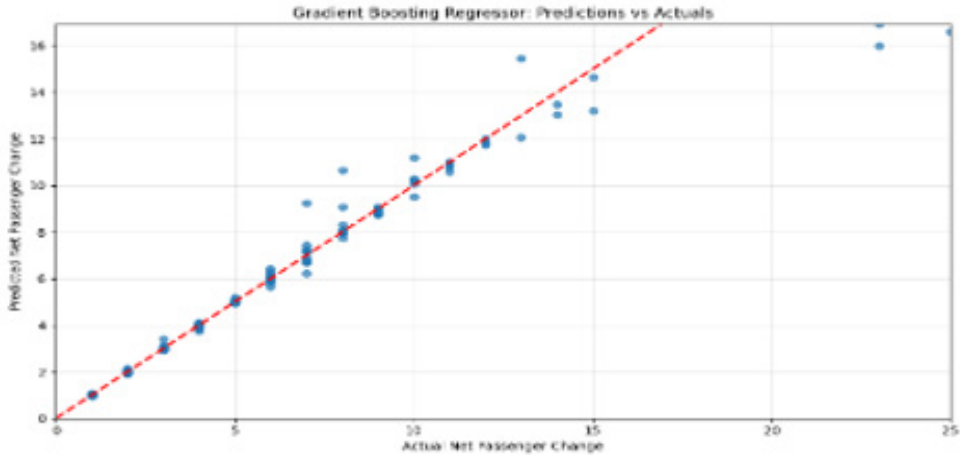


Fig. 8. Scatter plot showing predicted vs. actual net passenger change for the test set.

The Gradient Boosting Regressor (GBR) was deployed for its high accuracy and adaptability, integrated into a dual-layer pipeline: REST API for historical data and Kafka streaming for real-time updates. Automated retraining handles concept drift, keeping the model robust across seasons, holidays, and disruptions. GBR achieved strong performance (MAE = 0.17, RMSE = 0.78, MAPE = 2.21 %) and proved ideal for large-scale transit forecasting. Future improvements include adding contextual factors (weather, traffic, events) and hybridizing with deep learning models. For deployment, GBR is the primary model, with LightGBM and RandomForest as alternatives; continuous retraining ensures long-term stability.

In subsequent iterations, the system will incorporate contextual variables such as weather conditions, public event calendars, and traffic congestion levels through API-based data fusion. These features will serve as exogenous predictors, enhancing the model's responsiveness to irregular demand fluctuations caused by external urban dynamics. Integrating these data sources will further align the predictive framework with real-world smart city infrastructures.

Results and discussion.

Table 4 summarizes the lowest observed MAE, RMSE, and MAPE values for each model across the final set of experiments. These metrics provide a quantitative comparison of absolute and proportional errors, enabling the identification of the most effective model for real-time passenger flow prediction (Table 4).

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

Table 4. Model Performance Outputs

Model	MAE	RMSE	MAPE (%)
RandomForestRegressor	1.09	1.59	35 %
LightGBMRegressor	1.06	1.70	34.8 %
XGBRegressor	1.08	1.74	37.2 %
LSTM	2.79	3.55	111 %
ARIMA	128	211	72 %

Ensemble models outperformed ARIMA and LSTM. LightGBMRegressor showed the best results (MAE = 1.06, MAPE = 34.8 %), followed by RandomForestRegressor (MAE = 1.09, MAPE = 35–44 %) and XGBRegressor (MAE = 1.08, MAPE = 37.2 %). In contrast, LSTM (MAE = 2.79, MAPE = 111 %) and ARIMA (MAE = 128, RMSE = 211, MAPE = 72 %) performed poorly due to data complexity and non-stationarity. Thus, only ensemble models were included in Figure 1 to provide a meaningful comparison.

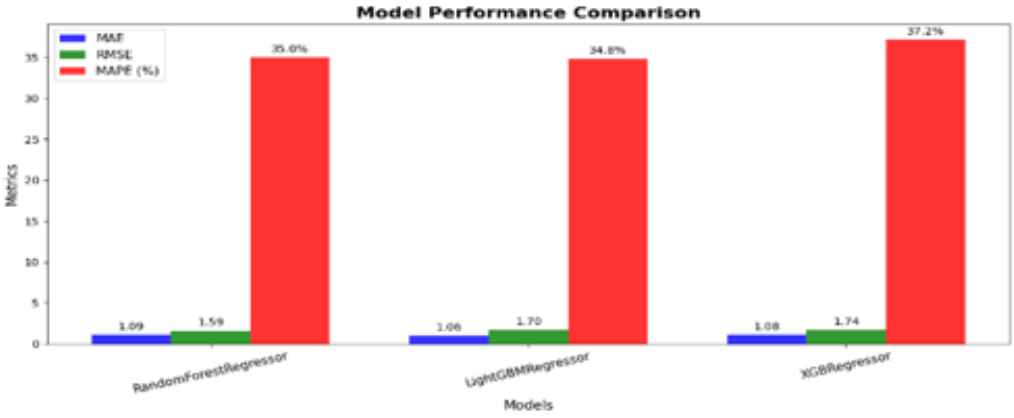


Fig. 9. Model Performance Bar Chart.

This analysis highlights the potential of ensemble-based methods, as it shown in Figure 8 particularly LightGBMRegressor, for robust and accurate real-time passenger flow prediction (Figure 9). By focusing on the models with meaningful results, we avoid overemphasizing suboptimal approaches like ARIMA and LSTM, ensuring a clear and practical assessment for real-world deployment.

The results highlight LightGBMRegressor as the most effective model, achieving the lowest MAE (1.06) and MAPE (34.8%), making it suitable for real-time passenger flow prediction. Its consistent performance demonstrates robustness across various scenarios, particularly for operational contexts requiring accurate and stable forecasting. RandomForestRegressor also performed well, with a slightly higher MAE (1.09) and MAPE (35 %). However, it exhibited variability in proportional errors during off-peak periods, suggesting that additional feature engineering, such as context-specific factors, could enhance its accuracy. XGBRegressor, while slightly behind in performance, showed

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

potential for further improvement through hyperparameter optimization and tailored feature engineering, particularly for imbalanced data scenarios.

The LSTM model and ARIMA were unsuitable for this application. LSTM struggled with the dataset's size and complexity, producing an MAE of 2.79 and MAPE exceeding 100%, indicating the need for larger datasets or advanced architectures to handle complex temporal dynamics. ARIMA's performance (MAE = 128, MAPE = 72 %) reflected its inability to manage non-stationary and non-linear patterns, making it impractical for real-time transit forecasting.

Among all models tested, Gradient Boosting Regressor (GBR) consistently achieved the lowest error rates, with MAE (0.17), RMSE (0.78), and MAPE (2.21 %), highlighting its ability to capture complex, non-linear dependencies in passenger flow dynamics. Unlike ARIMA and LSTM, which struggled with non-stationary trends and high variance, GBR's boosting-based ensemble approach effectively minimized deviations, providing stable and precise forecasting results.

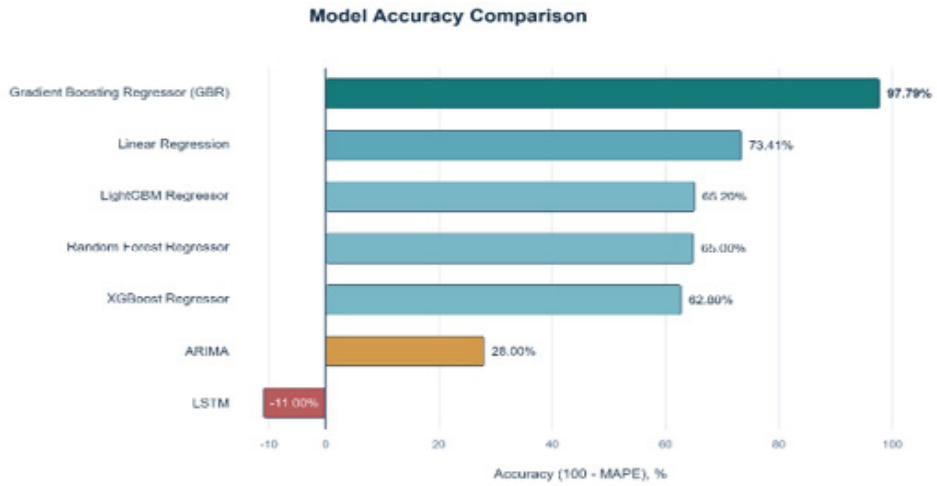


Fig. 10. Model Accuracy Comparison

Figure 10 compares model accuracy (100 − MAPE). Gradient Boosting Regressor (GBR) achieved the highest accuracy (97.79 %), far surpassing other models. Linear Regression performed moderately (73.41 %), while RandomForestRegressor and LightGBMRegressor showed lower accuracy (65–65.2 %) with higher variability. LSTM and ARIMA performed worst, confirming their unsuitability for complex, multi-dimensional transit data. Results highlight GBR as the most effective model, combining accuracy, efficiency, and adaptability.

To evaluate the real-time adaptability of the proposed system, the impact of retraining after drift detection was quantitatively assessed. Prior to adaptation, the model achieved a Mean Absolute Error (MAE) of 0.24 and Root Mean Squared Error (RMSE) of 1.02. Following KL divergence–based drift detection and incremental retraining, these metrics

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

improved to MAE = 0.17 and RMSE = 0.78. This improvement confirms the framework's ability to dynamically adjust to evolving passenger flow patterns, maintaining predictive stability under real-time conditions.

Future improvements include integrating contextual features (weather, events, congestion), enabling real-time adaptation via incremental learning, and exploring hybrid ensemble–deep learning models. Overall, ensemble methods outperform traditional and deep learning models, with GBR recommended for deployment in real-world transit forecasting.

Conclusion.

This study presents a framework for real-time passenger flow prediction using advanced models such as LightGBM, Random Forest, and ST-GCNs. Results show LightGBM as the most effective, enabling accurate forecasting to optimize resources, improve scheduling, and enhance passenger experience. The framework bridges theory and practice, supporting sustainable urban mobility through predictive analytics. Beyond technical performance, the proposed system contributes to data-driven policy development for sustainable transport planning. By providing timely and accurate passenger flow insights, it enables policymakers and transit authorities to optimize routes, allocate resources efficiently, and reduce environmental impacts through informed operational decisions. This aligns the framework with the broader objectives of smart city governance and sustainable urban development.

Limitations include a short dataset and lack of external factors (weather, traffic). A limitation of the current framework is the exclusion of external contextual factors such as weather anomalies, road closures, and major public events, which can significantly influence passenger flow dynamics. These factors will be integrated in future work using multimodal data fusion to enhance prediction stability and adaptability during irregular conditions. Future work should expand data, integrate contextual variables, and test advanced architectures (e.g., transformers, hybrids) in live environments to improve scalability and adaptability. This establishes a foundation for intelligent, efficient, and sustainable transit systems.

REFERENCES

Baghbani A., Bouguila N., Patterson Z. (2023). Short-term passenger flow prediction using a bus network graph convolutional long short-term memory neural network model // Transportation Research Record. — Vol. 2677. — No. 2. Pp. 1331–1340. DOI:10.1177/03611981221112673. [in Eng.].

Bureau of National Statistics, Transport in the Republic of Kazakhstan (January–March 2024), Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. 2024. URL: https://stat.gov.kz/en/industries/business-statistics/stat-transport/publications/157652. [in Eng.].

Cheng C. H., Tsai M. C., Cheng Y. C. (2022). An intelligent time-series model for forecasting bus passengers based on smartcard data, Applied Sciences. — Vol. 12. — No. 9. Pp. 4763. DOI:10.3390/app12094763. [in Eng.].

Chawla P., Hasurkar R., Bogadi C. R., Korlapati N. S., Rajendran R., Ravichandran S., Gao J. Z. (2024). Real-time traffic congestion prediction using big data and machine learning techniques // World Journal of Engineering. — Vol. 21. — No. 1. Pp. 140–155. DOI:10.1108/WJE-07-2021-0428. [in Eng.].

Chen Y., Li T., Sun Y., Wu J., Guo X., Liu D. (2023). Dynamic data-driven computation method for the number of waiting passengers and waiting time in the urban rail transit network, IET Intell. Transp. Syst. — Vol. 17. — No. 1. Pp. 165–179. DOI:10.1049/itr2.12245. [in Eng.].

El Maazouzi Q., Retbi A., Bennani S. (2022). Automatisation hyperparameters tuning process for times series forecasting: application to passenger's flow prediction on a railway network, Int. Arch. Photogramm. Remote Sens.

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

Spatial Inf. Sci. — Vol. 48. — Pp. 53–60. DOI:10.5194/isprs-archives-XLVIII-4-W3-2022-53-2022. [in Eng.].

Fu X., Wu M., Ponnarasu S., Zhang L. A hybrid deep learning approach for real-time estimation of passenger traffic flow in urban railway systems, Buildings. — Vol. 13. — No. 6. Pp. 1514. DOI:10.3390/buildings13061514. [in Eng.].

Garriga M., Aarns K., Tsigkanos C., Tamburri D. A., Heuvel W. V. D. (2021). DataOps for cyber-physical systems governance: the airport passenger flow case // ACM Trans. Internet Technol. — Vol. 21. — No. 2. Pp. 1–25. DOI:10.1145/3432247. [in Eng.].

Ghaziamin P., Bajaj K., Bouguila N., Patterson Z. (2024). A privacy-preserving edge computing solution for real-time passenger counting at bus stops using overhead fisheye camera, in: Proc. 2024 IEEE 18th Int. Conf. on Semantic Computing (ICSC), February. Pp. 25–32. DOI:10.1109/ICSC59802.2024.00011. [in Eng.].

Halyal S., Mulangi R. H., Harsha M. M. (2022). Forecasting public transit passenger demand: with neural networks using APC data, Case Studies on Transport Policy. — Vol. 10. — No. 2. Pp. 965–975. DOI: 10.1016/j.cstp.2022.03.011. [in Eng.].

Jain R., Dhingra S., Joshi K., Rana A., Goyal N. (2023). Enhance traffic flow prediction with real-time vehicle data integration // Journal of Artificial Intelligence. — Vol. 6. — No. 2. DOI:10.32629/jai.v6i2.574. [in Eng.].

Kolesnikova K., Naizabayeva L., Myrzabayeva A., Lisnevskyi R. (2024). Use the neural networks in prediction of environmental processes, in: Proc. 2024 IEEE 4th Int. Conf. on Smart Information Systems and Technologies (SIST). — Astana. Pp. 625–630. DOI: 10.1109/SIST61555.2024.10629330. [in Eng.].

Kramar V., Alchakov V. (2023). Time-series forecasting of seasonal data using machine learning methods, Algorithms. — Vol. 16. — No. 5. Pp. 248. DOI:10.3390/a16050248. [in Eng.].

Liyanage S., Abduljabbar R., Dia H., Tsai P. W. (2022). AI-based neural network models for bus passenger demand forecasting using smart card data // Journal of Urban Management. — Vol. 11. — No. 3. Pp. 365–380. DOI: 10.1016/j.jum.2022.05.002. [in Eng.].

Liu S., Li X., He C. (2021). Study on dynamic influence of passenger flow on intelligent bus travel service model, Transport. — Vol. 36. — No. 1. Pp. 25–37. DOI:10.3846/transport.2021.14343. [in Eng.].

Li B., Guo T., Li R., Wang Y., Gandomi A. H., Chen F. (2023). Self-adaptive predictive passenger flow modeling for large-scale railway systems // IEEE Internet of Things Journal. — Vol. 10. — No. 16. Pp. 14182–14196. DOI:10.1109/JIOT.2023.3270427. [in Eng.].

McCarthy C., Ghaderi H., Jayaraman P. P., Dia H. (2023). A deep learning approach to real-time video analytics for people and passenger counting // in: Handbook on Artificial Intelligence and Transport, Edward Elgar Publishing. Pp. 348–379. DOI:10.4337/9781803929545.00022. [in Eng.].

Marczyk M., Kempski A., Socha M., Cogiel M., Foszner P., Staniszewski M. (2024). Passenger location estimation in public transport: evaluating methods and camera placement impact, IEEE Trans. Intell. Transp. Syst. DOI:10.1109/TITS.2024.3434746. [in Eng.].

Neftissov A., Sarinova A., Kazambaev I., Kirichenko L., Lisnevskyi R., Rzayeva L. (2023). Development of the Smart Office Concept, in: Proc. 2023 IEEE Int. Conf. on Smart Information Systems and Technologies (SIST). — Astana. Pp. 382–386. DOI:10.1109/SIST58284.2023.10223512. [in Eng.].

Ou Y. (2022). AI for real-time bus travel time prediction in traffic congestion management, in: Humanity Driven AI: Productivity // Well-being, Sustainability and Partnership, Springer. Pp. 63–84. DOI:10.1007/978-3-030-72188-6_4. [in Eng.].

Tao Z., Tang J., Hou K. (2021). Online estimation model for passenger flow state in urban rail transit using multi-source data, Comput.-Aided Civil Infrastruct. Eng. — Vol. 36. — No. 6. Pp. 762–780. DOI:10.1111/mice.12671. [in Eng.].

Wu J., Du B., Gong Z., Wu Q., Shen J., Zhou L., Cai C. (2023). A GTFS data acquisition and processing framework and its application to train delay prediction, Int. J. Transp. Sci. Technol. — Vol. 12. — No. 1. Pp. 201–216. DOI: 10.1016/j.ijtst.2022.01.005. [in Eng.].

Ye J., Xu Z., Gou X. (2022). An adaptive Grey–Markov model based on parameters self-optimization with application to passenger flow volume prediction, Expert Systems with Applications. — Vol. 202. — Article 117302. DOI: 10.1016/j.eswa.2022.117302. [in Eng.].

Yan M., Zhou K. (2023). Bus passenger flow prediction using BO-optimized XGBoost, in: Proc. 2023 2nd Int. Conf. on Big Data, Information and Computer Network (BDICN), January. Pp. 134–141. DOI:10.1109/BDICN58493.2023.00035. [in Eng.].

Zeng L., Li Z., Yang J., Xu X. (2022). CEEMDAN-IPSO-LSTM: a novel model for short-term passenger flow prediction in urban rail transit systems, Int. J. Environ. Res. Public Health. — Vol. 19. — No. 24. Pp. 16433. DOI:10.3390/ijerph192416433. [in Eng.].

Zhang X., Lauber L., Liu H., Shi J., Xie M., Pan Y. (2022). Travel time prediction of urban public transportation

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

based on detection of single routes, PLoS One. — Vol. 17. — No. 1. e0262535. DOI: 10.1371/journal.pone.0262535. [in Eng.].

Zhai X., Shen Y. (2023). Short-term bus passenger flow prediction based on graph diffusion convolutional recurrent neural network // Applied Sciences. — Vol. 13. — No. 8. Pp. 4910. DOI:10.3390/app13084910. [in Eng.].

Zhang X., Yan M., Xie B., Yang H., Ma H. (2021). An automatic real-time bus schedule redesign method based on bus arrival time prediction // Advanced Engineering Informatics. — Vol. 48. Article 101295. DOI: 10.1016/j.aei.2021.101295. [in Eng.].

Zhao Y., Ma Z. (2022). Naïve Bayes-based transition model for short-term metro passenger flow prediction under planned events // Transportation Research Record. — Vol. 2676. — No. 9. Pp. 309–324. DOI:10.1177/03611981221086645. [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.2. Number 26 (2026). Pp. 60–82
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.004

A REVIEW OF PREDICTIVE MODELS APPROACHES TO STOCK MARKET FORECASTING IN KAZAKHSTAN

K. Tokayev*, A. Moldagulova
Satbayev University, Almaty, Kazakhstan.
E-mail: tokayev.k@stud.satbayev.university

Kuanysh Tokayev — PhD student, Department of Software Engineering, Satbayev University, Almaty, Kazakhstan
E-mail: Tokayev.K@stud.satbayev.university, https://orcid.org/0009-0006-5110-811X;
Aiman Moldagulova — Candidate of Mathematical and Physical Sciences, Professor, Department of Information Systems, Satbayev University, Almaty, Kazakhstan
https://orcid.org/0000-0002-1596-561X.

© K. Tokayev, A. Moldagulova

Abstract. These days, the financial world is driven more and more by data, so predictive analytics has become increasingly important for navigating modern markets. This is especially true for places like Kazakhstan, where the stock exchange is expanding rapidly and the country's becoming more integrated into global finance. In this review, recent developments in machine learning models for stock market forecasting are examined, with particular attention to their potential applications in Kazakhstan. The most common models - such as widespread use of LSTM networks - the types of data utilized, and the main evaluation metrics reported in the literature are summarized. Most studies tend to focus on complex models and large amounts of high-frequency data. However, in markets like Kazakhstan, hybrid, multi-scale models that also incorporate local sentiment and fundamental company data may offer the best approach. Ultimately, the key challenges involve ensuring high data quality and model transparency, as well as adapting global technological advances to fit local market realities. Through these efforts, stronger financial systems can be established to support more informed decision-making.

Keywords: stock market forecasting, predictive modeling, machine learning, emerging markets

For citation: Tokayev K., Moldagulova A. (2026). A review of predictive models approaches to stock market forecasting in Kazakhstan // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 60–82. https://doi.org/10.54309/IJICT.2026.26.2.004. (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

ҚАЗАҚСТАНДАҒЫ ҚОР НАРЫҒЫН БОЛЖАУДЫҢ БОЛЖАМДЫҚ МОДЕЛЬДЕРІНЕ ШОЛУ

К. Токаев*, А. Молдагулова
Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Казақстан.
E-mail: tokayev.k@stud.satbayev.university

Токаев Куаныш — PhD докторанты, Бағдарламалық инженерия кафедрасы, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
E-mail: Tokayev.K@stud.satbayev.university, https://orcid.org/0009-0006-5110-811X;
Молдагулова Айман — физика-математика ғылымдарының кандидаты, профессор, Ақпараттық жүйелер кафедрасы, Қ. И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-1596-561X.

© К. Токаев, А. Молдагулова

Аннотация. Қазіргі таңда қаржы әлемі деректерге барған сайын тәуелді болып келеді, сондықтан болжамды аналитика заманауи нарықтарды талдауда және бағдарлауда маңызды рөл атқарады. Бұл әсіресе қор нарығы қарқынды дамып, жаһандық қаржы жүйесіне интеграциялану үдерісі күшейіп келе жатқан Қазақстан үшін өзекті. Бұл шолуда қор нарығын болжауға арналған машиналық оқыту модельдерінің соңғы әзірлемелері қарастырылып, оларды Қазақстанда қолданудың әлеуетті мүмкіндіктеріне ерекше назар аударылады. Әдебиеттерде кеңінен қолданылатын модельдер, атап айтқанда LSTM желілері, пайдаланылатын деректер түрлері және негізгі бағалау көрсеткіштері жинақталып талданады. Зерттеулердің басым бөлігі күрделі модельдер мен жоғары жиілікті деректердің үлкен көлемін пайдалануға бағытталған. Алайда Қазақстан сияқты нарықтар үшін жергілікті нарықтық көңіл-күйді және компаниялардың іргелі көрсеткіштерін ескеретін гибридті, көпмасштабты модельдер тиімді тәсіл болуы мүмкін. Жалпы алғанда, негізгі мәселелер деректердің жоғары сапасы мен модельдердің ашықтығын қамтамасыз ету, сондай-ақ жаһандық технологиялық жетістіктерді жергілікті нарық жағдайларына бейімдеумен байланысты. Осы бағыттағы жұмыстарды дамыту арқылы неғұрлым тұрақты қаржы жүйесін қалыптастырып, негізделген шешімдер қабылдауға қолдау көрсетуге болады.

Түйінді сөздер: қор нарығын болжау, болжамды модельдеу, машиналық оқыту, дамушы нарықтар

Дәйексөздер үшін: Токаев К., Молдагулова А. (2026). Қазақстандағы қор нарығын болжаудың болжамдық модельдеріне шолу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 60–82. https://doi.

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

org/10.54309/IJICT.2026.26.2.004. (Ағыл. тіл.).

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

ОБЗОР ПРОГНОСТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ФОНДОВОГО РЫНКА В КАЗАХСТАНЕ

К. Токаев*, А. Молдагулова

Казахский национальный технический университет им. К. Сатпаева, Алматы, Казахстан.

E-mail: tokayev.k@stud.satbayev.university

Токаев Қуаныш — докторант PhD, кафедра программной инженерии, Казахский национальный технический университет им. К. Сатпаева, Алматы, Казахстан
E-mail: Tokayev.K@stud.satbayev.university, https://orcid.org/0009-0006-5110-811X;
Молдагулова Айман — кандидат физико-математических наук, профессор, кафедра информационных систем, Казахский национальный технический университет им. К.Сатпаева, Алматы, Казахстан
https://orcid.org/0000-0002-1596-561X.

© К. Токаев, А. Молдагулова

Аннотация. В настоящее время финансовый мир всё больше ориентируется на данные, поэтому прогнозная аналитика приобретает всё большее значение для анализа и понимания современных финансовых рынков. Это особенно актуально для таких стран, как Казахстан, где фондовый рынок стремительно развивается, а страна становится всё более интегрированной в глобальную финансовую систему. В данном обзоре рассматриваются последние достижения в области моделей машинного обучения для прогнозирования фондового рынка с особым вниманием к их потенциальному применению в Казахстане. Обобщаются наиболее распространённые модели, в частности широкое применение сетей LSTM, типы используемых данных, а также основные показатели оценки, представленные в научной литературе. Большинство исследований сосредоточено на сложных моделях и больших объёмах высокочастотных данных. Однако для таких рынков, как Казахстан, наиболее перспективным подходом могут быть гибридные многоуровневые модели, учитывающие также локальные настроения участников рынка и фундаментальные показатели компаний. В конечном итоге ключевыми задачами остаются обеспечение высокого качества данных и прозрачности моделей, а также адаптация глобальных технологических достижений к особенностям локальных рыночных условий. Реализация этих подходов позволит сформировать более устойчивую финансовую систему и обеспечить принятие более обоснованных решений.

Ключевые слова: прогнозирование фондового рынка, прогнозное моделирование, машинное обучение, развивающиеся рынки

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

Для цитирования: Токаев К., Молдагулова А. (2026). Обзор прогностических моделей для прогнозирования фондового рынка в казахстане // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 60–82. https://doi.org/10.54309/IJICT.2026.26.2.004. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The exponential growth of data generation has fundamentally transformed financial markets. By 2025, global data volumes are projected to reach 181 zettabytes, while the big data market is expected to grow from $220.2 billion in 2023 to $401.2 billion by 2028 (Vassakis et al., 2018). This growth reflects a fundamental shift in how organizations collect, analyze, and utilize information for decision-making. In the financial sector, the ability to interpret and act on real-time data has evolved from a competitive advantage to an essential capability for market participants.

For developing economies like Kazakhstan, which is experiencing rapid financial sector growth, predictive analytics offers a transformative opportunity. The Kazakhstan Stock Exchange (KASE) is seeing increased activity as the country's economy integrates more closely with global markets. In such a dynamic environment, the ability to accurately predict financial trends could help policymakers, investors, and institutions anticipate risks, optimize investment strategies, and support macroeconomic stability. However, forecasting market behavior is highly complex, influenced by a multitude of interrelated factors, including macroeconomic indicators, political events, social media sentiment, and investor psychology (Min, 2020).

For decades, investors have relied on three core approaches to financial market analysis: fundamental analysis, technical analysis, and sentiment analysis (Nti et al., 2019). Fundamental analysis examines a security's intrinsic value through economic and financial indicators, including earnings, interest rates, and geopolitical events. This approach assumes that prices fluctuate around intrinsic value due to market noise but eventually revert to equilibrium (Arya, 2020). However, fundamental analysis has limitations: it is subject to analyst judgment, relies on historical reports, and may not capture rapid market changes (Min, 2020).

Technical analysis, in contrast, focuses on quantitative data, using historical price movements and trading volumes to predict future price behavior. Analysts employ statistical techniques and charting tools to detect market trends, momentum, and cycles (Nti et al., 2019). While more objective, technical analysis has its limitations in capturing external events or sudden sentiment shifts that can drastically affect market dynamics. That's where sentiment analysis comes into play, an increasingly important tool in the financial analyst's toolkit. It involves extracting psychological and emotional cues from unstructured textual data such as news headlines, social media posts, and expert commentary (Singh et al., 2022). Powered by natural language processing (NLP), sentiment analysis can reveal investor confidence or panic before such sentiments manifest in the market. Nevertheless, it remains supplementary rather than foundational, primarily due to the complexity of

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

modeling linguistic nuances and the lack of standardized integration with quantitative models.

Each of these methodologies provides unique insights, but none is sufficient on its own. Therefore, there is a growing consensus that conjoint analytical frameworks, capable of integrating fundamental, technical, and sentiment analysis, are necessary for reliable predictions in volatile financial environments. A critical enabler of such integration is the construction of a data lake, a centralized repository that accommodates structured (e.g., historical prices) and unstructured data (e.g., financial news articles). Realizing such a data ecosystem introduces its own challenges, particularly in addressing the well-established 5V characteristics of Big Data: volume, velocity, variety, veracity, and value (Naeem et al., 2021).

These challenges intensify with real-time analysis requirements. Processing high-frequency stock data while simultaneously extracting sentiment from news sources demands substantial computational resources and scalable infrastructure. Legacy batch-oriented pipelines often lack the necessary fault tolerance and flexibility. For emerging markets like Kazakhstan with limited local data, leveraging models and insights from mature markets through transfer learning and adaptation becomes particularly valuable.

Given this context, the present review examines research that applies machine learning–based predictive models to stock market forecasting. The emphasis is on algorithms such as regression models, decision trees, ensemble methods, and deep learning architectures, including recurrent and long short-term memory networks, which have demonstrated potential to capture complex, nonlinear relationships in financial data. Studies that integrate multiple paradigms into hybrid frameworks are also considered, as such multidimensional approaches reflect the increasing need for more robust and adaptable forecasting strategies in volatile markets. An example of such a hybrid framework is illustrated in Figure 1, where structured historical stock data from the Kazakhstan Stock Exchange (KASE) and unstructured financial news articles can be combined through the real-time preprocessing pipeline before being fed into predictive models. This type of integration demonstrates how technical data, fundamental drivers, and sentiment information can be jointly leveraged to enhance predictive accuracy—an approach highly relevant for emerging markets such as Kazakhstan.

This review does not attempt to provide an exhaustive account of conventional forecasting methods that lack machine learning components, nor does it include purely theoretical discussions without empirical validation. Instead, the focus is on research that demonstrates practical applications and measurable predictive performance, with particular attention to insights that may inform the development of forecasting models for emerging financial markets such as Kazakhstan.

Materials and methods.

Predictive modeling has emerged as a central component in financial market analysis, offering researchers and practitioners the capacity to anticipate future trends and behaviors by identifying patterns within large volumes of historical data. As financial markets have become increasingly data-driven and interconnected, traditional forecasting

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

methods have given way to more sophisticated, data-centric approaches. The adoption of machine learning (ML) and deep learning (DL) models, particularly in the context of time-series analysis, has transformed the financial forecasting landscape, especially in emerging economies such as Kazakhstan. These models offer the potential to model complex nonlinear relationships, adapt to changing market dynamics, and incorporate a wide range of features including textual sentiment, macroeconomic indicators, and historical pricing data.

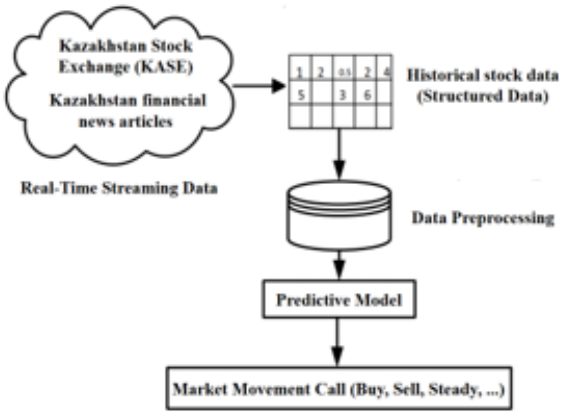


Fig. 1. Hybrid predictive framework for stock market forecasting in Kazakhstan.

The growing interest in ML-based financial forecasting stems from the limitations of classical econometric methods like ARIMA or linear regression. Although these methods have long been the standard, they often rely on strong assumptions of stationarity, linearity, and normally distributed errors, which rarely hold in financial time series. In contrast, predictive models based on ML can flexibly capture nonlinear relationships and temporal dependencies without the need for extensive pre-specification. A systematic review by Nti, Adekoya, and Weyori provides a comprehensive summary of how ML techniques have evolved in financial forecasting, highlighting a shift from fundamental and technical analysis toward more intelligent systems capable of self-learning and adaptation (Nti et al., 2019).

Among the most widely adopted models for time-series prediction is the Long Short-Term Memory (LSTM) network, a variant of the recurrent neural network (RNN) architecture designed specifically to mitigate the vanishing gradient problem and capture long-term dependencies. LSTM networks maintain internal memory cells that allow them to store relevant information over extended sequences, making them especially useful for capturing trends and seasonality in financial data. Fazeli and Houghten demonstrated the effectiveness of LSTM for stock price forecasting, achieving low mean squared error (MSE) scores on financial datasets across multiple markets between 2014 and 2019 (Fazeli et al., 2019). Their work showed how LSTM models outperformed standard feedforward networks and even traditional RNNs in handling sequential financial inputs.

Despite their success, LSTM networks are not without drawbacks. The model's

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

complexity can lead to long training times and a higher risk of overfitting, especially in markets with limited historical data or high levels of noise. To overcome this, researchers have investigated hybrid modeling approaches that combine the strengths of multiple algorithms. Li and Soo mentioned a hybrid model integrating ARIMA with LSTM, using the former to capture linear patterns and the latter to model nonlinear residuals (Lee et al., 2018). Their results indicated improved forecasting accuracy over either model alone, especially in volatile environments like the Taiwanese stock market. However, such hybrid models also introduce challenges in terms of hyperparameter tuning and system interpretability.

Another promising avenue in predictive modeling involves the integration of convolutional neural networks (CNNs) with LSTM architectures. Alonso-Monsalve et al. applied this hybrid CNN-LSTM model to cryptocurrency markets, achieving high forecasting accuracy on volatile assets such as Bitcoin and Ethereum (Alonso-Monsalve et al., 2020). By using CNNs to extract spatial features from structured input representations and LSTM to model sequential dependencies, this architecture proved highly effective in markets characterized by high-frequency fluctuations. Nevertheless, these models often demand large training datasets and substantial computational resources, posing limitations for researchers or institutions operating in data-scarce or resource-constrained environments.

While DL techniques have shown strong predictive capabilities, their black-box nature has raised concerns regarding interpretability and trustworthiness in financial contexts. Financial analysts and decision-makers often require explanations for model predictions, especially in regulatory environments where transparency is critical. This has led to renewed interest in ensemble learning methods, such as Random Forest (RF) and Gradient Boosting (GB), which offer greater interpretability while maintaining high performance. These models construct predictions by aggregating the outcomes of multiple decision trees, thereby reducing variance and increasing robustness to overfitting. Although RF and GB may not capture temporal dependencies as naturally as LSTM, they can be extremely useful when supplemented with lagged features or applied to cross-sectional data involving firm-specific indicators, macroeconomic signals, or sentiment metrics.

Sentiment analysis has become an increasingly important feature in financial forecasting, particularly in the age of social media and real-time news feeds. Numerous studies have confirmed that investor sentiment extracted from news articles, Twitter posts, and financial blogs can have predictive value over short time horizons. Valle-Cruz et al. analyzed Twitter sentiment to guide investment decisions during the COVID-19 pandemic, revealing a strong correlation between public mood and market movement (Valle-Cruz et al., 2021). Similarly, Lee and Soo employed recurrent convolutional neural networks (RCNNs) to forecast stock trends using news data, demonstrating that combining textual sentiment with quantitative indicators significantly boosts predictive accuracy (Lee et al., 2018). These findings underscore the value of unstructured data in enriching predictive models and suggest that future financial forecasting systems should

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

be equipped to process both numerical and textual data.

Kazakhstan's financial markets offer a unique testing ground for the development and application of such predictive techniques. As a frontier market, it presents both challenges and opportunities due to its relatively lower liquidity, limited transparency, and evolving regulatory landscape. Sembiyeva et al. explored the use of regression and AI-based models to forecast the green bond market in Kazakhstan, achieving strong statistical performance and drawing attention to the role of predictive analytics in sustainable finance (Sembiyeva et al., 2024). Their study illustrates how emerging markets can benefit from adopting modern data science techniques, especially when local data limitations are addressed through innovative modeling strategies.

The infrastructure supporting predictive modeling is critical for time-sensitive financial applications. Modern distributed architectures enable real-time data ingestion and processing, enhancing system responsiveness to sudden market shifts (Peddireddy, 2023). Vassakis et al. emphasize that predictive modeling's utility hinges on efficient data pipelines and computing frameworks capable of handling financial big data's volume, velocity, and variety (Vassakis et al., 2018).

As predictive modeling techniques continue to evolve, the intersection of big data, machine learning, and domain-specific financial knowledge will define the next frontier of market forecasting. The inclusion of explainable AI, better feature engineering techniques, and real-time adaptation capabilities will further enhance the relevance of these tools. These developments are particularly significant for policymakers, institutional investors, and researchers operating in emerging markets like Kazakhstan, where reliable and interpretable forecasting models can inform both economic planning and investment strategy.

The effectiveness of predictive modeling in financial markets also hinges on the selection and transformation of relevant features. Feature engineering remains a critical component of model development, often determining whether a predictive system will succeed or fail. In financial contexts, common features include price-based indicators such as moving averages, relative strength index (RSI), Bollinger Bands, and various momentum measures. These features attempt to capture technical patterns in asset behavior that may precede price changes. However, the increasing availability of alternative data, such as satellite imagery, ESG metrics, or geospatial data, has expanded the scope of feature selection significantly. As observed by Gandomi and Haider, big data's contribution to predictive modeling is not only about volume but also about the variety and richness of information that can now be leveraged for forecasting purposes (Gandomi et al., 2015).

The rise of big data technologies has transformed predictive modeling workflows in finance. Modern distributed computing platforms enable faster processing of both streaming and batch data, while NoSQL databases and cloud-based storage facilitate ingestion of heterogeneous data types at scale. Vassakis et al. emphasize the importance of scalable infrastructures for advanced analytics in financial applications, especially when integrating high-velocity data streams like stock tick data, social media sentiment, and news feeds (Vassakis et al., 2018).

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

Moreover, hyperparameter optimization has become an essential part of training predictive models. As financial markets exhibit high volatility and nonstationarity, even small changes in model configuration can lead to significant differences in performance. Automated machine learning (AutoML) frameworks such as Google's AutoML or H2O.ai offer streamlined solutions by testing numerous model types and parameter combinations to find the best-performing pipeline (Majidi et al., 2022). Nonetheless, in financial applications, it is critical to ensure that optimization techniques do not lead to overfitting or information leakage. Rolling window validation and walk-forward analysis are therefore commonly used to maintain the temporal integrity of the data, as discussed by Nti et al. in their evaluation of machine learning techniques for financial forecasting (Nti et al., 2019).

Another emerging trend is the incorporation of attention mechanisms into sequence modeling, particularly in transformer-based models. Originally developed for natural language processing, transformers have been increasingly applied to time-series forecasting due to their ability to model long-range dependencies without relying on sequential recurrence. Models like BERT and its time-series derivatives have shown promise in capturing complex temporal relationships in financial datasets. The attention mechanism allows the model to focus on relevant time points or data features dynamically, improving interpretability and potentially enhancing predictive power. While still a nascent area in financial modeling, the use of transformers is likely to grow, especially as more financial institutions invest in large-scale AI research.

While accuracy remains a core metric for model evaluation, other considerations such as robustness, adaptability, and explainability are becoming increasingly important. In practical settings, a highly accurate model that fails during regime changes or market shocks may be of little value. This has led to growing interest in adaptive modeling techniques that can adjust their parameters or structure in response to changing market dynamics. Reinforcement learning (RL) offers one such approach, where models learn to optimize a reward function based on feedback from the environment. For example, Carta et al. proposed a hybrid model for intelligent stock trading based on deep Q-learning, exemplifying how RL can be applied to decision-making in financial contexts, with the model learning from market responses to improve trading strategies over time (Carta et al., 2021).

The trade-off between model complexity and interpretability remains a central tension in financial predictive modeling. While deep learning architectures often yield superior performance, they can be opaque and difficult to interpret. This poses a problem not only for end users seeking to understand model outputs but also for regulatory bodies demanding transparency. Explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are therefore gaining traction. These tools provide post-hoc explanations of model predictions, allowing analysts to identify which features were most influential in a given decision. Sembiyeva et al., in their analysis of Kazakhstan's green bond market, emphasized the importance of model transparency, particularly in policy-oriented financial research where stakeholder trust is paramount (Sembiyeva et al., 2024).

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

The reliability of predictive modeling also depends on data quality and preprocessing techniques. Financial datasets often contain missing values, outliers, or inconsistencies that can severely impair model performance if not properly handled. Data normalization, smoothing, imputation, and anomaly detection are all critical preprocessing steps. Batini and Scannapieco propose a multi-dimensional data quality framework that includes accuracy, completeness, consistency, and timeliness, which can be tailored to financial datasets for rigorous quality assurance (Batini et al., 2016). For emerging financial systems such as Kazakhstan's, where digitization is still expanding, establishing high-quality data pipelines from the outset is crucial for long-term success in predictive analytics. Moreover, when working with textual data such as financial news or social media, natural language processing (NLP) techniques must be employed to clean, tokenize, and vectorize text inputs. Word embedding models like GloVe or FastText convert text into dense numerical representations, allowing predictive models to ingest and analyze linguistic information. Pencheva et al. demonstrated the utility of NLP for sentiment-driven forecasting, using news-based feature engineering to improve model accuracy in cryptocurrency markets (Pencheva et al., 2020).

Ensemble learning remains one of the most effective ways to improve predictive performance while managing model variance. By combining multiple models, each with its own strengths and weaknesses, ensemble methods can often outperform any single predictor. Techniques like bagging, boosting, and stacking allow modelers to aggregate results in a structured way. In financial modeling, stacking is particularly useful when combining models that specialize in different data modalities, such as combining an LSTM for price data with a random forest for fundamental indicators. Valle-Cruz et al. highlighted the benefits of model stacking in their COVID-19 investment prediction study, where they fused sentiment analysis models with quantitative regressors to achieve more stable forecasts (Valle-Cruz et al., 2021).

Risk management is another critical area where predictive modeling plays a pivotal role. Predicting future asset returns is just one part of financial decision-making; equally important is assessing the likelihood and impact of adverse outcomes. Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR), and expected shortfall are widely used metrics, and predictive models are increasingly being used to forecast these risk measures under various market scenarios. Bayesian neural networks (BNNs) are gaining popularity for this purpose because they offer probabilistic outputs, allowing for better quantification of uncertainty. Saba and Babajide's review of artificial intelligence applications in banking highlighted how AI models can provide early warnings of liquidity or credit risks, enabling proactive risk mitigation (Madu-Saba et al., 2024).

In the context of Kazakhstan and other frontier markets, predictive modeling offers significant potential to overcome traditional barriers to financial development. These markets often lack high-frequency trading infrastructure, large institutional investor bases, or extensive historical data. Nonetheless, by using transfer learning or pretraining models on data from more developed markets and fine-tuning them on local data, it is possible to build reliable predictive systems even in data-scarce environments. Furthermore, predictive

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

models can help attract foreign investment by improving market transparency and reducing information asymmetry, aligning with broader economic development goals.

The rapid evolution of hardware acceleration technologies, such as GPUs and TPUs, has further empowered the use of deep learning models for financial forecasting. These technologies allow for faster training of large-scale models, facilitating experimentation with deeper architectures and larger datasets. As processing power becomes increasingly accessible via cloud services, even resource-intensive models like transformers and deep LSTMs are now within reach for academic and industrial researchers. According to Costa et al., the combination of cloud computing and AI enables the democratization of advanced analytics, allowing institutions from emerging economies to access tools that were previously restricted to major financial centers (Costa et al., 2024). This technological democratization is particularly relevant for countries like Kazakhstan, where financial systems are undergoing digital transformation and can leapfrog legacy infrastructure.

In addition to market prices and volumes, sentiment data derived from social media and news articles has become a vital input for predictive modeling. The rationale lies in the behavioral finance theory, which posits that investor sentiment significantly influences asset prices. Tools such as VADER, TextBlob, and BERT-based sentiment classifiers are used to quantify textual data into sentiment scores. These scores can then be incorporated into predictive models as explanatory variables. For instance, a surge in positive sentiment around a stock may signal upcoming price increases, particularly in markets where retail investors dominate trading volumes. Nasir et al. provided evidence that sentiment-based variables improved forecasting accuracy in the cryptocurrency market, highlighting the importance of non-traditional data sources in modeling financial behavior (Nasir et al., 2019).

Another promising direction involves graph-based machine learning techniques, particularly graph neural networks (GNNs), which model relationships between entities such as firms, assets, or market indices. Financial systems are inherently networked, with intricate connections between market participants, sectors, and geographies. GNNs allow for the encoding of these relationships into the model architecture itself, enabling the prediction of cascading effects in systemic risk analysis or supply chain disruptions. As noted by Ghosh et al., such techniques offer a more nuanced understanding of financial contagion and interdependencies, which are especially relevant in globally integrated markets (Ghosh et al., 2014). While still emerging, these models offer a pathway to more holistic and interconnected forecasting frameworks.

Model evaluation in predictive analytics also extends beyond traditional error metrics such as MAE, RMSE, or R^2. In financial contexts, metrics must reflect the economic utility of the prediction. For example, a model might achieve a high R^2 but still perform poorly when translated into trading decisions due to high transaction costs or poor timing. Therefore, profit-based evaluation metrics such as Sharpe ratio, maximum drawdown, and cumulative returns are frequently used to assess model effectiveness in a real-world context. Dash and Dash demonstrated this approach in their financial

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

forecasting models by integrating predictive accuracy with portfolio performance simulations, thus aligning evaluation more closely with investor goals (Dash et al., 2016).

Moreover, economic cycles and structural breaks present significant challenges to predictive modeling. Financial time series are nonstationary, meaning their statistical properties can change over time due to macroeconomic events, regulatory shifts, or technological innovations. These regime shifts can invalidate model assumptions or learned patterns, leading to forecast degradation. To mitigate this, regime-switching models, such as Markov-switching autoregressive models or hidden Markov models (HMMs), have been employed to capture distinct market states and adjust predictions accordingly. Alternatively, ensemble models can incorporate regime-specific predictors and dynamically allocate weights depending on the prevailing conditions. As Guresen et al. highlighted in their work comparing ANN and SVM for stock price prediction, the adaptability of a model to changing conditions is as crucial as its baseline accuracy (Guresen et al., 2011).

Transfer learning has emerged as another tool for enhancing model generalization, especially in data-scarce environments. By pretraining models on large, related datasets and fine-tuning them on the target data, transfer learning enables better performance when labeled examples are limited. This approach is particularly useful in financial applications where specific markets may not have sufficient historical data for training from scratch. For instance, a model trained on global equity markets can be fine-tuned for local market conditions in Kazakhstan, thereby accelerating development and improving accuracy. Cao et al. have shown the benefits of transfer learning in improving forecasting for illiquid and emerging markets by leveraging broader data resources (Cao et al., 2023).

In risk-sensitive domains like finance, interpretability and trust are paramount, especially when predictive models are used to guide regulatory or high-stakes decisions. Black-box models, despite their predictive power, often face resistance from stakeholders who need clear justifications for recommendations. Techniques such as counterfactual explanations, decision trees extracted from neural networks, and rule-based surrogate models offer partial solutions by providing simplified versions of complex models. In response, institutions are increasingly embedding explainability audits into their model development lifecycle, ensuring accountability and compliance with emerging AI governance frameworks.

The convergence of predictive modeling and economic forecasting is further exemplified by macro-financial modeling efforts that combine traditional econometric models with machine learning. Hybrid models may incorporate ARIMA or vector autoregression (VAR) as components of broader machine learning pipelines, benefiting from the interpretability of classical models and the flexibility of AI. Azqueta-Gavaldón et al. proposed models where economic indicators such as GDP, interest rates, and inflation are fused with real-time market data to predict macro-level financial trends, demonstrating that machine learning can complement, rather than replace, economic intuition (Azqueta-Gavaldón, 2017). This synergy is essential for central banks and policy institutions in Kazakhstan and similar economies, where forward-looking indicators are crucial for

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

planning and stability.

As financial systems become more digitized, predictive modeling techniques are also playing a growing role in fraud detection and anti-money laundering (AML) efforts. By analyzing transaction patterns, user behavior, and contextual signals, models can identify anomalous activity in real time. Techniques such as clustering, isolation forests, and neural network-based autoencoders are commonly employed to detect deviations from expected norms. Hernández et al. applied machine learning algorithms to detect suspicious banking activities and achieved higher accuracy than traditional rule-based systems, illustrating the potential for AI-driven compliance mechanisms (Hernandez et al., 2024). In Kazakhstan, where digital banking and e-commerce are expanding rapidly, integrating such predictive systems into national financial infrastructure can significantly enhance trust and operational security.

The integration of real-time data streams is reshaping financial forecasting. With high-frequency trading proliferation, online learning techniques and streaming analytics enable dynamic adjustments to market changes. According to Fischer and Krauss, ensemble learning models that continually adapt to new market conditions perform substantially better than static models in highly volatile environments (Fischer et al., 2018). This capability is vital for markets in transition, such as those in Central Asia, where macroeconomic indicators can shift quickly due to policy changes or external shocks.

Moreover, the role of feature engineering remains critical in financial predictive modeling despite the rise of end-to-end learning systems. Crafting meaningful features from raw data, such as moving averages, volatility indicators, sentiment scores, or event flags, significantly enhances model performance. Even in deep learning contexts, hybrid approaches combining raw data with engineered features often yield superior results. Feature importance ranking techniques, such as SHAP and LIME, improve performance while offering transparency regarding prediction drivers. Zheng et al. demonstrate how integrating domain knowledge into feature engineering provides robustness and mitigates overfitting in financial neural networks (Zheng et al., 2021).

Furthermore, predictive models are increasingly being evaluated for their fairness and potential for bias, particularly in credit scoring and lending decisions. Machine learning models trained on historical data may inadvertently learn patterns that reflect existing inequalities or systemic biases. For instance, if past lending practices were discriminatory, a model may learn to replicate those patterns unless corrective measures are taken. Techniques such as adversarial debiasing, reweighing, and fairness-constrained optimization are now being applied to mitigate bias and ensure equitable predictions. In high-stakes contexts, such as access to capital or insurance pricing, fairness-aware models are not only a regulatory requirement but also a strategic imperative for fostering inclusive financial systems. Khandani et al. discuss the importance of designing credit risk models that balance predictive accuracy with fairness constraints (Khandani et al., 2010).

Another important frontier in predictive modeling is the fusion of structured

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

and unstructured data. Financial data is not confined to rows and columns; it also exists in the form of earnings call transcripts, CEO interviews, satellite images of retail parking lots, and customer reviews. Natural Language Processing (NLP) models, especially transformer-based architectures like BERT and RoBERTa, are now being used to extract actionable insights from textual data. These models can classify sentiment, detect topics, or even predict market movements based on linguistic patterns. Chen et al. demonstrated that incorporating insights from news headlines into financial models significantly improved short-term prediction accuracy for stock indices (Chen et al., 2024). This fusion of data types enhances situational awareness and enables a richer understanding of market behavior, especially during periods of uncertainty or crisis.

Model retraining and lifecycle management are key components of sustainable predictive modeling practices. As financial markets evolve, statistical relationships between variables may degrade, necessitating periodic updates. Model drift, where prediction accuracy deteriorates due to changing input distributions, requires ongoing monitoring and validation. Continuous integration and deployment (CI/CD) pipelines, often referred to as MLOps, facilitate automated retraining, version control, and model deployment. According to Zbikowski, MLOps frameworks ensure both scalability and consistency in model performance across time and use cases (Zbikowski, 2015).

The use of ensemble learning methods also continues to expand in financial modeling. By combining the strengths of multiple models, ensemble methods such as bagging, boosting, and stacking help reduce variance, bias, and improve generalizability. Random Forests, Gradient Boosting Machines (like XGBoost and LightGBM), and Voting Classifiers have demonstrated strong performance in predicting asset prices, credit defaults, and customer churn. According to Patel et al., ensemble methods consistently outperformed single-model approaches across various financial forecasting tasks, especially when calibrated carefully with appropriate hyperparameters and feature sets (Patel et al., 2015). These techniques are also more resilient to outliers and overfitting, making them suitable for datasets with noisy or heterogeneous financial variables.

An important yet often underexplored aspect of predictive modeling in finance is the psychological and behavioral interpretation of model outcomes. Predictions influence human decisions, which in turn affect markets. This feedback loop means that models not only reflect but also shape market realities. For instance, if a predictive model becomes widely adopted, its recommendations may drive collective behaviors that reinforce or invalidate its own forecasts. This is particularly evident in algorithmic trading, where model-based strategies compete and interact in real time. Bollen et al. showed that collective sentiment derived from social media could be used to forecast DJIA movements, highlighting how information dissemination and behavioral response co-evolve in financial ecosystems (Bollen et al., 2011).

Finally, the regulatory landscape around predictive analytics in finance is tightening. As models become more influential in decision-making, regulators are demanding greater transparency, auditability, and accountability. Frameworks such as the EU's AI Act and the U.S. Securities and Exchange Commission (SEC)'s guidelines on AI in trading emphasize

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

the need for explainable, robust, and fair algorithms. For financial institutions operating across jurisdictions, this creates a complex compliance environment that necessitates not only technical but also legal and ethical considerations in model design. As noted by Li et al., predictive modeling in finance must evolve within a governance framework that balances innovation with responsibility (Wu et al., 2020).

As predictive modeling continues to evolve, the role of cross-disciplinary approaches is becoming increasingly important. The convergence of finance, data science, behavioral economics, and computer science fosters the development of innovative models that are both technically sophisticated and practically relevant. For example, the integration of network theory and graph analytics provides fresh perspectives on market interconnectedness and systemic risk. By modeling financial entities, such as banks, funds, or even individual assets, as nodes within a network, researchers can quantify contagion effects, identify critical nodes, and simulate crisis propagation. Barabási's work on complex networks has influenced financial risk management by highlighting the non-linear dependencies inherent in market systems (Barabási, 2016). Applying these insights in Kazakhstan's emerging markets offers opportunities to preempt systemic shocks through predictive indicators derived from network properties.

Moreover, cloud computing and distributed systems have transformed the scalability and accessibility of predictive modeling. High-performance computing platforms enable training of large-scale models with vast datasets, including alternative data sources like satellite imagery or transaction-level records, supporting more granular and timely forecasts. Parallel processing frameworks facilitate efficient big data handling, accelerating model training and inference. Khandani et al. emphasize that scalable infrastructure is prerequisite for deploying advanced credit risk models requiring frequent retraining and testing in production environments (Khandani et al., 2010). For developing economies, leveraging cloud platforms can offset local infrastructure constraints and accelerate predictive analytics adoption.

Uncertainty quantification is a critical yet often underappreciated component of financial market prediction. Predictive models involve probabilistic estimates with varying degrees of uncertainty, and communicating this uncertainty is essential for risk management and decision-making. Bayesian methods and probabilistic programming frameworks enable explicit modeling of uncertainty in parameters and predictions. Gaussian Process Regression offers both point predictions and confidence intervals, valuable for cautious investment strategies. Monte Carlo simulations combined with predictive models enable scenario analysis and stress testing under various market conditions. According to Hernández et al., incorporating uncertainty quantification improves forecast robustness and interpretability, enabling users to better balance risk and reward (Hernandez et al., 2024).

Ethical considerations in predictive modeling, particularly in finance, are increasingly gaining attention. The deployment of predictive algorithms can have profound social and economic impacts, influencing credit access, investment outcomes, and even employment opportunities. Transparency, accountability, and the prevention of harm are

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

key ethical principles that must guide model development and application. Data privacy is a major concern, especially with sensitive financial information. Techniques such as differential privacy and federated learning are emerging as ways to protect individual data while still enabling effective model training. Federated learning allows models to be trained across decentralized data sources without sharing raw data, preserving confidentiality. Kairouz et al. provide an extensive review of privacy-preserving machine learning techniques relevant for sensitive financial applications (Kairouz et al., 2021). Ethical frameworks must be embedded into predictive modeling workflows to ensure that technological progress benefits society equitably.

The emergence of explainable AI (XAI) frameworks continues to address the challenge of balancing model complexity with interpretability. In finance, the ability to explain why a model made a particular prediction is crucial for regulatory compliance and stakeholder trust. Techniques such as SHAP values, counterfactual explanations, and surrogate models offer insights into the inner workings of complex algorithms. These explanations help detect biases, validate assumptions, and uncover causal relationships, thereby increasing the credibility of models. Dos Santos and Da Silva underscore the role of XAI in fostering collaboration between data scientists, domain experts, and business users, leading to more informed and responsible decision-making (Santos et al., 2024). The shift toward explainability also supports the integration of human expertise, enabling hybrid intelligence systems where AI augments rather than replaces human judgment.

Another noteworthy trend is the application of reinforcement learning (RL) to financial decision-making. Unlike traditional supervised learning, RL focuses on learning optimal policies through interaction with the environment, making it particularly suited for sequential decision problems such as portfolio optimization and algorithmic trading. Deep reinforcement learning, which combines neural networks with RL, has shown promise in learning complex trading strategies that adapt to evolving market dynamics. Deng et al. demonstrated that deep RL agents could outperform benchmark trading algorithms by optimizing reward functions aligned with financial goals (Deng et al., 2017). However, challenges remain in stabilizing training, handling sparse rewards, and ensuring interpretability, which are active areas of research.

The role of alternative data in predictive modeling cannot be overstated. Beyond conventional financial and economic indicators, data sources such as social media sentiment, web search trends, credit card transaction data, and even environmental sensor data offer novel predictive signals. Integrating these diverse datasets requires sophisticated data fusion and multi-modal learning techniques. Multi-task learning frameworks can leverage related tasks to improve overall model performance by sharing representations. Chen et al. showed that combining social media data with traditional market variables led to significant improvements in short-term stock movement prediction (Chen et al., 2024). For emerging financial markets, harnessing alternative data can compensate for limited traditional data availability and provide competitive advantages.

Finally, the future of predictive modeling in finance is poised to benefit from advances in quantum computing. Although still nascent, quantum machine learning

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

algorithms promise exponential speedups for certain optimization and sampling problems relevant to finance. Quantum annealing and variational quantum circuits are being explored for portfolio optimization, risk analysis, and option pricing. While practical applications remain experimental, early work suggests that quantum-enhanced models could tackle problems that are currently computationally intractable (Goudarzi et al., 2023). The intersection of quantum computing and financial predictive modeling represents a frontier that could revolutionize the field over the coming decades.

Research methodology.

This review adopts a systematic literature synthesis approach. Studies from 2000–2023 were collected through major databases (Scopus, Web of Science, IEEE Xplore, and Google Scholar) using keywords such as stock market forecasting, machine learning, deep learning, LSTM, hybrid models, sentiment analysis, and Kazakhstan stock market. Papers were included if they presented predictive models applied to financial markets with reported metrics. Emphasis was placed on hybrid architectures and approaches relevant to emerging markets.

Results and discussion.

As illustrated in Figure 2, Long Short-Term Memory (LSTM) models and their variants (including CMH-LSTM) are, by a significant margin, the most widely employed techniques, representing the approach in nearly half of the reviewed studies. This underscores their dominant capability and popularity in capturing complex, non-linear sequential dependencies in financial time series data. Beyond LSTMs, regression models and decision trees form a secondary tier of commonly used techniques, while methods like autoencoders, reinforcement learning, and other hybrid models are currently explored in a smaller subset of research.

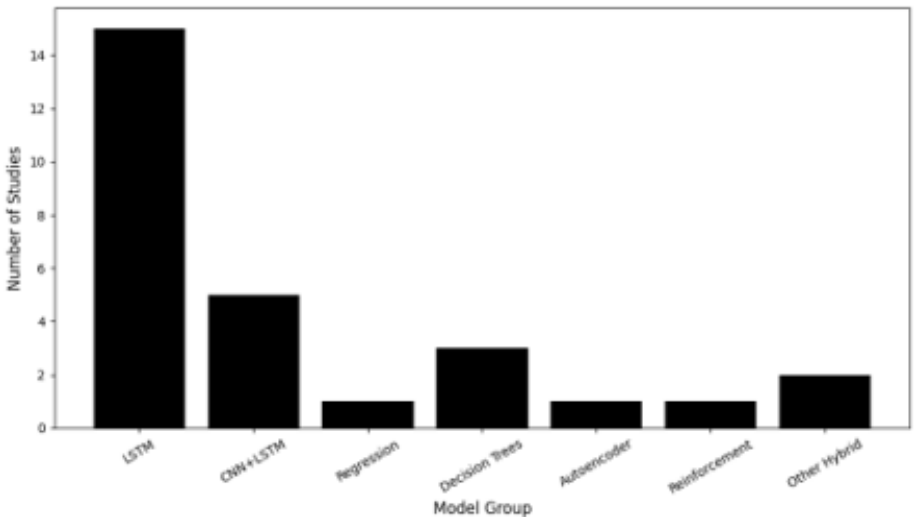


Fig. 2. Distribution of model technologies in financial market prediction studies.

To provide a comprehensive overview of the state-of-the-art predictive models in financial forecasting, the following Table 1 summarizes key studies, their methodologies,

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

and outcomes.

Table 1. Review of predictive modeling approaches in financial market forecasting

Methods	Frequency	Period	Market	Predictor	Metric Used
Fazeli & Houghten	Daily	2014-2019	American	LSTM	MSE
Sembiyeva et al	Yearly	2018-2022	Kazakhstan	Regression	R-squared, Statistical validations
Alonso-Monsalve	Monthly	2018-2019	Cryptocurrency	CNN+LSTM	Accuracy, Statistical validations
Lee & Soo	Daily	2009-2018	Taiwanese	Arima+LSTM RNN	MAE, RMSE
Tsai & Wang	Quarterly	2002 - 2007	Taiwanese	DT + ANN	Accuracy
Matsubara et al	Daily	2001 - 2013	American, Japanese	Deep Autoencoder	Accuracy, Statistical validations
Singh et al	Weekly	1982-1990	Japanese	LSTM	Accuracy, Statistical validations
Agrawal et al.	Daily	2016-2018	Indian	Optimal LSTM	Accuracy, MSE
Al-Thelaya et al.	Daily	2010-2018	Bahrain	AutoSLSTM: LSTM autoencoder	MAE, RMSE, R^2
Orimoloye et al.	Weekly	2010-2018	S&P 500, Nikkei 225	DeepFFNN, Shallow Models	Accuracy, MAPE
Alonso-Monsalve et al.	1-Minute (Tick)	2018-2019	Crypto currencies	CLSTM: CNN + LSTM	Accuracy, Statistical validations
Borovkova & Tsiamas	5-Minute	2014	American	LSTM	AUC, ROC
Mohapatra et al.	5-Minute	2018-2019	Cryptocurrency	LSTM+ Twitter Sentiment	RMSE, F1 Score
Chen et al.	Daily	2004-2018	Chinese	LSTM with Attention	MAE, MSE
Demir et al.	Hourly	2014-2018	Belgian	2NN, CNN, ResNet	RMSE, MAE, PCC, Diebold-Mariano
JuHyok et al.	Daily	1987-2018	Chinese, American	LSTM	Accuracy, Precision, Recall, F1 Score
Kelotra & Pandey	Daily	2017-2018	Indian	Deep-ConvLSTM	MSE, RMSE
Roy	Multi-Timeframe	2015-2021	US & European	Thick Data + DRL	Sharpe Ratio
Labiad et al.	10/30/60-Minute	2016-2017	Moroccan	LSTM	Accuracy, Precision, Recall, F1 Score
Goudarzi & Bazzana	Millisecond	2020-2022	Global (HFT)	Random Forest+Order Book Data	Precision, Latency
Lai et al.	Daily	2007-2017	Taiwanese	RCN-CNN+ LSTM	RMSE
Li & Liao	Daily	2003-2008	Chinese	LSTM	Accuracy, Precision, Recall, F1 Score
Ismailova et al.	Daily	2018-2023	Kazakhstan (KASE)	Hybrid CNN-BiLSTM	Accuracy, MAE
Li et al.	15-Minute	2008-2015	Brazilian	LSTM	Accuracy, Precision, Recall, F1 Score
Oncharoen & Vateekul	Daily	2006-2017	American	CNN + LSTM	Accuracy
Oncharoen & Vateekul	Daily	2006-2016	American	CNN + LSTM	F1 Score

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

The distribution of data frequencies, presented in Figure 3, indicates a strong preference for higher-frequency data. Daily data is the most prevalent time frame, forming the basis for a clear majority of the studies. High-frequency (intraday) data is also extensively used, highlighting a significant research focus on capturing short-term market dynamics and microstructural patterns. In contrast, longer-term horizons such as weekly, monthly, quarterly, and yearly data are utilized less frequently, though the adoption of multi-timeframe analysis suggests a growing interest in integrating trends from different scales.

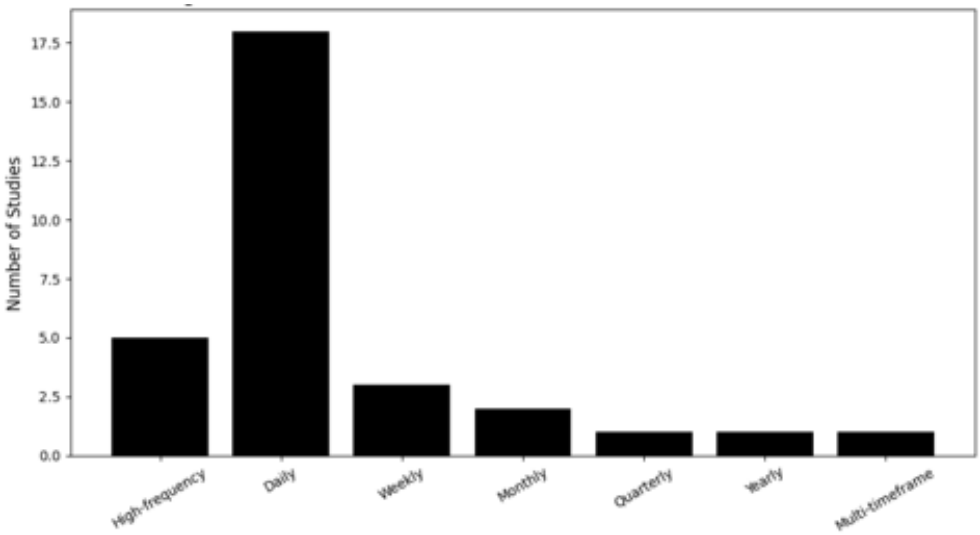


Fig. 3. Distribution of time frames used in market prediction research.

Figure 4 details the usage of performance metrics across the literature. The evaluation landscape is dominated by two primary groups: accuracy-based metrics (e.g., Accuracy, F1-Score) and error-based metrics (e.g., MAE, MSE, RMSE). Their concurrent high prevalence indicates that the field places a strong, dual focus on both the classification correctness of predictive models (e.g., predicting price direction) and the precision of their regression outputs (e.g., predicting exact price values). Statistical validation techniques and financial metrics (e.g., Sharpe Ratio, returns) are also employed, though to a lesser extent, suggesting that a portion of the research validates model performance through economic significance rather than purely statistical measures.

The findings of this review reveal clear trends in predictive modeling and present important considerations for their application in the Kazakhstan Stock Exchange (KASE). The overwhelming dominance of LSTM-based models is justified by their strength in analyzing sequential data, making them a prime candidate for initial forecasting efforts in Kazakhstan (Fischer et al., 2018). However, their complexity requires caution. For a developing market like KASE, with potentially less data volume, these models risk overfitting (Livieris et al., 2020). A more robust approach may lie in the hybrid frameworks that combine LSTMs with more interpretable models like decision trees.

The strong focus on high-frequency and daily data in global research aligns with

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

modern, high-speed trading. Yet, for an emerging market like Kazakhstan, which is highly sensitive to macroeconomic shifts and commodity prices, this focus may be misplaced (Smales, 2014). The most valuable insights for investors and policymakers might instead come from models that effectively integrate lower-frequency, fundamental data. This suggests a promising opportunity for multi-scale models that analyze daily patterns within the context of broader economic trends.

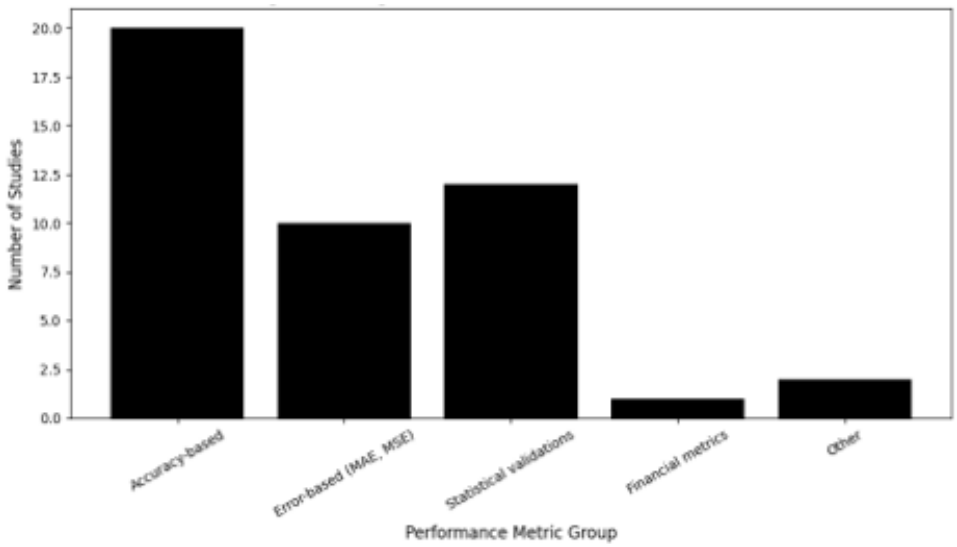


Fig. 4. Usage of performance metrics across studies.

Furthermore, the near-universal reliance on accuracy and error-based metrics only tells half the story. For a model to be truly valuable for KASE, its performance must be measured by financial utility, such as risk-adjusted returns - not just statistical precision. Finally, leveraging sentiment analysis on local news sources could be a particularly powerful tool for Kazakhstan, as investor sentiment may be a stronger predictor in less efficient markets.

Conclusion.

Predictive modeling in finance has evolved significantly from traditional statistical methods to embrace machine learning, deep learning, and hybrid approaches. The field continues to advance through improved data accessibility, increased computational capabilities, and growing emphasis on ethical and regulatory considerations. For emerging financial markets such as Kazakhstan, adopting these innovations offers substantial opportunities for enhanced forecasting accuracy, improved risk management, and potentially greater financial inclusion.

However, significant challenges remain. Ensuring data quality, maintaining model transparency and fairness, and managing the reflexive impact of models on market behavior are ongoing concerns. Future research must address these issues through interdisciplinary approaches and strengthened collaboration among academics, industry practitioners, and regulators. Through such efforts, predictive modeling can continue transforming the

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

financial landscape, contributing to more informed, equitable, and resilient economic systems.

REFERENCES

Azqueta-Gavaldón A. (2017). Developing news-based economic policy uncertainty index with unsupervised machine learning, Economics Letters. — Vol. 158. — Pp. 47–50. Sep. 2017, DOI: 10.1016/j.econlet.2017.06.034. [In Eng.].

Arya K. (2020). Fundamental Analysis Underlies Investment Growth. [In Eng.].

Al-Thelaya K.A., El-Alfy E.-S.-M., and Mohammed S. (2019). Forecasting of Bahrain stock market with deep learning: Methodology and case study, in Proc. 8th Int. Conf. Modeling Simulation Appl. Optim. (ICMSAO), Pp. 1–5. [In Eng.].

Agrawal M., Khan A.U., and Shukla P.K. (2019). Stock price prediction using technical indicators: A predictive model using optimal deep learning, Learning. — Vol. 6. — No. 2. Pp. 7. [In Eng.].

Alonso-Monsalve S., Suárez-Cetrulo A. L., Cervantes A., and Quintana D. (2020). Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators, Expert Systems with Applications. — Vol. 149. — Pp. 113250. DOI: https://doi.org/10.1016/j.eswa.2020.113250. [In Eng.].

Borovkova S. and Tsiamas I. (2019). An ensemble of LSTM neural networks for high-frequency stock market classification, J. Forecasting. — Vol. 38. — No. 6. Pp. 600–619. [In Eng.].

Batini C. and Scannapieco M. (2016). Data and Information Quality. Cham, Switzerland: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-24106-7. [In Eng.].

Barabási A.L., Network Science (2016). Cambridge University Press. [In Eng.].

Bollen J., Mao H., and Zeng X. (2011). Twitter mood predicts the stock market // Journal of Computational Science. — Vol. 2. — No. 1. Pp. 1-8. [In Eng.].

Carta S., Ferreira A., Podda A. S., Recupero D. R., and Sanna A. (2021). Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting // Expert Systems with Applications. — Vol. 164. — Pp. 113820. DOI: https://doi.org/10.1016/j.eswa.2020.113820. [In Eng.].

Chen P., Boukouvalas Z., and Corizzo R. (2024). A Deep Fusion Model for Stock Market Prediction with News Headlines and Time Series Data // Neural Computing and Applications. — Vol. 36. — No. 34. Pp. 21229–21271. Dec. 2024, DOI: 10.1007/s00521-024-10303-1. [In Eng.].

Chen Y., Wu J., and Bu H. (2018). Stock market embedding and prediction: A deep learning method, in Proc. 15th Int. Conf. Service Syst. Service Manage. (ICSSSM). Pp. 1–6. [In Eng.].

Costa C. J., Aparicio M., Aparicio S., and Aparicio J. T. (2024). The Democratization of Artificial Intelligence: Theoretical Framework // Applied Sciences. — Vol. 14. — No. 18. Pp. 8236. DOI: 10.3390/app14188236. [In Eng.].

Cao H., Gu H., Guo X., and Rosenbaum M. (2023). Transfer learning for portfolio optimization, arXiv preprint arXiv:2307.13546. [In Eng.].

Deng Y., Bao F., Kong Y., Ren Z., and Dai Q. (2017). Deep direct reinforcement learning for financial signal representation and trading // IEEE Transactions on Neural Networks and Learning Systems. — Vol. 28. — No. 3. Pp. 653-664. [In Eng.].

Dash R. and Dash P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques // The Journal of Finance and Data Science. — Vol. 2. — No. 1. Pp. 42-57. [In Eng.].

Demir S., Mincev K., Kok K., and Paterakis N.G. (2019). Introducing technical indicators to electricity price forecasting: A feature engineering study for linear, ensemble, and deep machine learning models, Appl. Sci. — Vol. 10. — No. 1. Pp. 255. [In Eng.].

dos Santos A. Q., da Silva J. R., Lins R. F., and Prudencio R. B. C. (2024). Explainable Machine Learning Techniques for Criticality Prediction of Software Change Requests, in Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), Pp. 520–528, SBC. [In Eng.].

Fazeli A. and Houghten S. (2019). Deep learning for the prediction of stock market trends, in Proc. IEEE Int. Conf. Big Data (Big Data), Pp. 5513–5521. [In Eng.].

Fischer T. and Krauss C. (2018). "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research. — Vol. 270. — No. 2. Pp. 654–669. [In Eng.].

Gandomi A. and Haider M. (2015). Beyond the hype: Big data concepts, methods, and analytics // International Journal of Information Management. — Vol. 35. — No. 2. Pp. 137-144. [In Eng.].

Ghosh A.R., Qureshi M.S., Kim J.I., and Zalduendo J. (2014). "Surges // Journal of International Economics. — Vol. 92. — No. 2. Pp. 266–285, Mar. 2014, DOI: 10.1016/j.jinteco.2013.11.007. [In Eng.].

Guresen E., Kayakutlu G., and Daim T. U. (2011). Using artificial neural network models in stock market

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

index prediction, Expert Systems with Applications. — Vol. 38. — No. 8. Pp. 10389-10397. [In Eng.].

Goudarzi M. and Bazzana F. (2023). Identification of high-frequency trading: A machine learning approach, Res. Int. Bus. Finance. — Vol. 66. — No. 102078. DOI: 10.1016/j.ribaf.2023.102078. [In Eng.].

Hernandez Aros L., Bustamante Molano L.X., Gutierrez-Portela F., Moreno Hernandez J.J., and Rodríguez Barrero M.S. (2024). Financial fraud detection through the application of machine learning techniques: a literature review // Humanities and Social Sciences Communications. — Vol. 11. — No. 1. Pp. 1–22. DOI: 10.1057/s41599-023-02004-8. [In Eng.].

Ismailova A. et al., (2024). Forecasting stock market prices using deep learning methods, Int. J. Elect. Comput. Eng. — Vol. 14. — No. 5. Pp. 5601-5611. DOI: 10.11591/ijece.v14i5.pp5601-5611. [In Eng.].

JuHyok U., Lu P., Kim C., Ryu U., and Pak K. (2020). A new LSTM based reversal point prediction method using upward/downward reversal point feature sets // Chaos, Solitons Fractals. — Vol. 132. Article 109559. [In Eng.].

Khandani A.E., Kim A.J., and Lo A.W. (2010). "Consumer credit-risk models via machine-learning algorithms," Journal of Banking & Finance. — Vol. 34. — No. 11. Pp. 2767–2787, Nov. 2010, DOI: 10.1016/j.jbankfin.2010.06.001. [In Eng.].

Kelotra A. and Pandey P. (2020). Stock market prediction using optimized deep-ConvLSTM model, Big Data. — Vol. 8. — No. 1. Pp. 5–24. [In Eng.].

Kairouz P. et al. (2021). "Advances and open problems in federated learning // Foundations and Trends® in Machine Learning. — Vol. 14. — No. 1–2. Pp. 1–210. DOI: 10.1561/2200000083. [In Eng.].

Labiad B., Berrado A., and Benabbou L. (2018). Short term prediction framework for Moroccan stock market using artificial neural networks, in Proc. 12th Int. Conf. Intell. Syst., Theories Appl. (SITA), Pp. 1–6. [In Eng.].

Lee C.-Y. and Soo V.-W. (2017). Predict Stock Price with Financial News Based on Recurrent Convolutional Neural Networks // 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI). DOI: https://doi.org/10.1109/taai.2017.27. [In Eng.].

Lai C. Y., Chen R.-C., and Caraka R. E. (2019). Prediction stock price based on different index factors using LSTM, in Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC), Pp. 1–6. [In Eng.].

Li W. and Liao J. (2017). A comparative study on trend forecasting approach for stock price time series, in Proc. 11th IEEE Int. Conf. Anti-Counterfeiting, Secur., Identificat. (ASID)*. Pp. 74–78. [In Eng.].

Li X., Wu P., and Wang W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Inf. Process. Manage. — Vol. 57. — No. 5. Article 102212. [In Eng.].

Livieris I. E., Pintelas E., and Pintelas P. (2020). A CNN–LSTM model for gold price time-series forecasting. Neural Computing and Applications. — Vol. 32. — No. 23. Pp. 17351–17360. [In Eng.].

Madu-Saba M.B., Babajide G. A., Musa R. O., and Nwozor U. F. (2024). Fraud Control Practices Adopted by Small and Medium Scale Enterprises for Business Sustainability in Enugu State, Nigeria // ESUT Journal of Social Sciences. — Vol. 9. — No. 2. [In Eng.].

Min J. (2020). Financial Market Trend Forecasting and Performance Analysis Using LSTM, arXiv:2004.01502 [cs, q-fin]. Available: https://arxiv.org/abs/2004.01502I. [In Eng.].

Matsubara T., Akita R., and Uehara K. (2018). Stock Price Prediction by Deep Neural Generative Model of News Articles. IEICE TRANSACTIONS on Information and Systems. — Vol. E101-D. — No. 4. Pp. 901–908. DOI: https://search.ieice.org/bin/pdf_link.php?fname=e101d_4_901&category=D&lang=E&year=2018&abst=.[In Eng.].

Majidi F., Openja M., Khomh F., and Li H. (2022). An empirical study on the usage of automated machine learning tools, in Proc. IEEE Int. Conf. Software Maintenance and Evolution (ICSME), Pp. 59–70. DOI: https://doi.org/10.1109/ICSME55016.2022.00017. [In Eng.].

Mohapatra S. et al. (2019). KryptoOracle: A real-time cryptocurrency price prediction platform using Twitter sentiments. in Proc. IEEE Int. Conf. Big Data, Dec. Pp. 2983-2990. DOI: 10.1109/BigData47090.2019.9005598. [In Eng.].

Naeem M. et al. (2021). Trends and Future Perspective Challenges in Big Data // Advances in Intelligent Data Analysis and Applications. — Vol. 253. Pp. 309–325. Available: https://link.springer.com/chapter/10.1007/978-981-16-5036-9_30. [In Eng.].

Nasir M.A., Huynh T.L.D., Nguyen S.P., and Duong D. (2019). Forecasting cryptocurrency returns and volume using search engines, Financial Innovation. — Vol. 5. — No. 25. Pp. 1–13. DOI: 10.1186/s40854-018-0119-. [In Eng.].

Nti I. K., Adekoya A. F., and Weyori B. A. (2019). A systematic review of fundamental and technical analysis of stock market predictions // Artificial Intelligence Review. — Vol. 53. —No. 1. Pp. 3007–3057. DOI: https://doi.org/10.1007/s10462-019-09754-z. [In Eng.].

Orimoloye L.O. et al. (2020). Comparing the effectiveness of deep feedforward neural networks and shallow architectures for predicting stock price indices // Expert Syst. Appl. — Vol. 139. — Art. no. 112828. DOI:

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

10.1016/j.eswa.2019.112828. [In Eng.].

Oncharoen P. and Vateekul P. (2018). Deep learning for stock market prediction using event embedding and technical indicators, in Proc. 5th Int. Conf. Adv. Inform., Concept Theory Appl. (ICAICTA), Aug. Pp. 19–24. [In Eng.].

Oncharoen P. and Vateekul P. (2018). Deep learning using risk-reward function for stock market prediction, in Proc. 2nd Int. Conf. Comput. Sci. Artif. Intell. (CSAI). Pp. 556–561. [In Eng.].

Pencheva I., Esteve M., and Mikhaylov S. J. (2020). Big Data and AI—A transformational shift for government: So, what next for research? // Public Policy and Administration. — Vol. 35. — No. 1. Pp. 24–44. DOI: https://doi.org/10.1177/0952076718780537. [In Eng.].

Patel J., Shah S., Thakkar P., and Kotecha K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques // Expert Systems with Applications. — Vol. 42. — No. 1. Pp. 259-268. [In Eng.].

Peddireddy K. (2023). Kafka-based architecture in building data lakes for real-time data streams // International Journal of Computer Applications. — Vol. 185. — No. 9. Pp. 1–3. [In Eng.].

Roy G. (2022). Multi-Timeframe Algorithmic Trading Bots Using Thick Data Heuristics with Deep Reinforcement Learning, Ph.D. dissertation, Univ. London. [In Eng.].

Sembiyeva L., Zhagyparova A., Zhumadillayeva A., Zholamanova M., Bekbolsynova A., and Zhanabergenova M. (2024). Applying Advanced Artificial Intelligence to Predict the Green Bond Market in Kazakhstan: Fostering Sustainable Financial Instruments and Environmental Objectives'. [In Eng.].

Smales L.A. (2014). News sentiment and the investor fear gauge // Finance Research Letters. — Vol. 11. — No. 2. Pp. 122–130. [In Eng.].

Singh A., Bhardwaj G., Srivastava A.P., Bindra A., Chaudhary P., and Ritika. (2022). Application of Neural Network to Technical Analysis of Stock Market Prediction, IEEE Xplore. [In Eng.].

Tsai C. F. and Wang S. P. (2009). Stock price forecasting by hybrid machine learning techniques, in Proceedings of the international multiconference of engineers and computer scientists. — Vol. 1. — Pp. 60. [In Eng.].

Valle-Cruz D., Fernandez-Cortez V., López-Chau A., and Sandoval-Almazán R. (2021). Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods, Cognitive Computation. — Vol. 14. — DOI: https://doi.org/10.1007/s12559-021-09819-8. [In Eng.].

Vassakis K., Petrakis E., and Kopanakis I. (2018). Big Data Analytics: Applications, Prospects and Challenges, Mobile Big Data. — Vol. 10. — Pp. 3–20. DOI: https://doi.org/10.1007/978-3-319-67925-9_1. [In Eng.].

Zbikowski K. (2015). Using Volume Weighted Support Vector Machines with walk forward testing and feature selection for the purpose of creating stock trading strategy // Expert Systems with Applications. — Vol. 42. — No. 4. Pp. 1797-1805. [In Eng.].

Zheng Y., Yang Y., and Chen B. (2021). Incorporating prior financial domain knowledge into neural networks for implied volatility surface prediction // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Pp. 3968–3975. DOI: 10.1145/3447548.3467423. [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.2. Number 26 (2026). Pp. 83–96
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.005

AI TECHNOLOGIES IN LANGUAGE EDUCATION: PRACTICAL ASPECTS AND CHALLENGES OF STUDENT USAGE

F.N. Abdraimova, A.A. Kereibayeva*, D.S. Dyussenova, D.A. Aliyeva, T.Zh. Toktarova
International Information Technology University, Almaty, Kazakhstan.
E-mail: a.kereyibaeva@iitu.edu.kz

Fariza N. Abdraimova — master, senior lecturer, «Department of Languages», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-6931-23154;
Aigerim A. Kereibayeva — PhD, senior lecturer, «Department of Languages», International Information Technology University, Almaty, Kazakhstan
E-mail: a.kereyibaeva@iitu.edu.kz, https://orcid.org/0000-0002-0330-9033;
Dinara S. Dyussenova — master, assistant-professor, «Department of Languages», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0003-4514-9856;
Dinara A. Aliyeva — Candidate of Pedagogical Sciences, Professor, «Department of Languages», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0001-8058-7152;
Tolkyn Zh. Toktarova — Candidate of Philological Sciences, Associate Professor, «Department of Languages», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0001-6674-3251.

© F.N. Abdraimova, A.A. Kereibayeva, D.S. Dyussenova, D.A. Aliyeva, T.Zh. Toktarova

Abstract. This article examines the practical aspects of utilizing artificial intelligence (AI) technologies in language education for university students, identifying the associated cognitive and methodological challenges. The relevance of the study stems from the contradiction between the rapid integration of AI and the insufficient research regarding the actual strategies students employ when using these tools. The aim of the research is to identify the specific characteristics of AI application in language learning and

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

to demonstrate that, without proper methodological support, its use diminishes the depth of analytical activity. The empirical foundation includes the testing of an educational chatbot, «Русский язык_ментор» (Russian Language_Mentor) through focus groups involving 30 students, utilizing methods of observation, analysis of student performance outcomes, and qualitative content analysis. Results indicate that while AI is effective for reproductive tasks, its responses to analytical assignments tend to be generalized. Furthermore, a significant asymmetry was identified between the correctness of the AI-generated answers and the students' actual depth of understanding. The study concludes that methodologically grounded AI integration is essential to foster the development of students' critical thinking.

Keywords: Artificial Intelligence in Education, Language Pedagogy, AI Chatbots, Critical Inquiry, Digital Instructional Strategies

For citation: F.N. Abdraimova, A.A. Kereibayeva, D.S. Dyussenova, D.A. Aliyeva, T.Zh. Toktarova (2026). AI technologies in language education: practical aspects and challenges of student usage // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 83–96. https://doi.org/10.54309/IJICT.2026.26.2.005. (In Eng.).

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

ТІЛ БІЛІМІНДЕ ЖАСАНДЫ ИНТЕЛЛЕКТ ТЕХНОЛОГИЯЛАРЫ: СТУДЕНТТЕР ҚОЛДАНУЫНЫҢ ПРАКТИКАЛЫҚ АСПЕКТІЛЕРІ МЕН МӘСЕЛЕЛЕРІ

Ф.Н. Абдраимова, А.А. Керейбаева*, Д.С. Дюсенова, Д.А. Алиева, Т.Ж. Токтарова

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

Абдраимова Фариза Нуржановна — магистр, тілдер кафедрасының аға оқытушысы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-6931-2315;
Керейбаева Айгерим Абильгазиевна — PhD, тілдер кафедрасының аға оқытушысы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: a.kereyibaeva@iitu.edu.kz, https://orcid.org/0000-0002-0330-9033;
Дюссенова Динара Сагимжановна — магистр, ассистент-профессор, тілдер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-4514-9856;
Алиева Динара Асылхановна — педагогика ғылымдарының кандидаты, профессор, тілдер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-8058-7152;
Токтарова Толқын Жанибековна — филология ғылымдарының кандидаты, доцент, тілдер кафедрасы, Халықаралық ақпараттық технологиялар университеті,

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

Алматы, Қазақстан
https://orcid.org/0000-0001-6674-3251.

© Ф.Н. Абдраимова, А.А. Керейбаева, Д.С. Дюссенова, Д.А. Алиева, Т.Ж. Токтарова

Аннотация. Мақала студенттердің тілдік білімінде жасанды интеллект (ЖИ) технологияларын қолданудың практикалық аспектілерін қарастырып, онымен байланысты когнитивті және әдістемелік мәселелерді айқындайды. Зерттеудің өзектілігі ЖИ-дің белсенді енгізілуі мен білім алушылардың оны пайдаланудағы шынайы стратегияларының жеткілікті зерттелмеуі арасындағы қайшылықтан туындап отыр. Зерттеудің мақсаты — тіл үйретуде ЖИ-ді қолданудың ерекшеліктерін анықтау және әдістемелік сүйемелдеусіз оны пайдалану талдау қызметінің тереңдігін төмендететінін көрсету. Эмпирикалық база 30 студенттің қатысуымен фокус-топ форматында «Русский язык_ментор» (Орыс тілі_ментор) оқу чат-ботын апробациядан өткізуді қамтиды. Зерттеуде бақылау, іс-әрекет өнімдерін талдау және сапалы контент-талдау әдістері қолданылды. Нәтижелер көрсеткендей, ЖИ репродуктивті тапсырмаларды орындауда тиімді, алайда аналитикалық тапсырмалар кезінде жауаптар жалпылама сипатқа ие болады. Жауаптың дұрыстығы мен оны түсіну тереңдігі арасындағы асимметрия анықталды. Студенттердің сыни ойлауын дамыту үшін ЖИ-ді әдістемелік тұрғыдан негізделген интеграциялау қажеттілігі туралы қорытынды жасалады.

Түйінді сөздер: жасанды интеллект, тілдік білім беру, чат-бот, сыни ойлау, цифрлық педагогика

Дәйексөздер үшін: Ф.Н. Абраимова, А.А. Керейбаева, Д.С. Дюссенова, Д.А. Алиева, Т.Ж. Токтарова (2026). Тіл білімінде жасанды интеллект технологиялары: студенттер қолдануының практикалық аспектілері мен мәселелері // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 83–96. https://doi.org/10.54309/IJICT.2026.26.2.005. (Ағыл. тіл.).

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

ТЕХНОЛОГИИ ИИ В ЯЗЫКОВОМ ОБРАЗОВАНИИ: ПРАКТИЧЕСКИЕ АСПЕКТЫ И ПРОБЛЕМЫ ПРИМЕНЕНИЯ СТУДЕНТАМИ

Ф.Н. Абдраимова, А.А. Керейбаева*, Д.С. Дюсенова, Д.А. Алиева, Т.Ж. Токтарова
Международный университет информационных технологий, Алматы, Казахстан.
E-mail: a.kereyibaeva@iitu.edu.kz

Абдраимова Фариза Нуржановна — магистр, старший преподаватель кафедры языков, Международный университет информационных технологий, Алматы, Казахстан https://orcid.org/0000-0002-6931-2315;
Керейбаева Айгерим Абильгазиевна — PhD, старший преподаватель кафедры

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

языков, Международный университет информационных технологий, Алматы, Казахстан
E-mail: a.kereyibaeva@iitu.edu.kz, https://orcid.org/0000-0002-0330-9033;
Дюссенова Динара Сагимжановна — магистр, ассистент-профессор кафедры языков, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0003-4514-9856;
Алиева Динара Асылхановна — кандидат педагогических наук, профессор кафедры языков, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0001-8058-7152;
Токтарова Толқын Жанибековна — кандидат филологических наук, доцент кафедры языков, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0001-6674-3251.

© Ф.Н. Абдраимова, А.А. Керейбаева, Д.С. Дюсенова, Д.А. Алиева, Т.Ж. Токтарова

Аннотация. В статье рассматриваются практические аспекты использования технологий искусственного интеллекта в языковом образовании студентов и выявляются связанные с этим когнитивные и методические проблемы. Актуальность исследования обусловлена противоречием между активным внедрением ИИ и недостаточной изученностью реальных стратегий его использования обучающимися. Цель исследования-выявить особенности применения ИИ в языковом обучении и показать, что без методического сопровождения его использование снижает глубину аналитической деятельности. Эмпирическая база включает апробацию учебного чат-бота «Русский язык_ментор» в формате фокус-групп с участием 30 студентов. Использованы методы наблюдения, анализа продуктов деятельности и качественного контент-анализа. Результаты показывают, что ИИ эффективен при выполнении репродуктивных заданий, однако при аналитических задачах ответы носят обобщенный характер. Выявлена асимметрия между корректностью ответа и глубиной его понимания. Делается вывод о необходимости методически обоснованной интеграции ИИ для развития критического мышления студентов.

Ключевые слова: искусственный интеллект, языковое образование, чат-бот, критическое мышление, цифровая педагогика

Для цитирования: Ф.Н. Абраимова, А.А. Керейбаева, Д.С. Дюссенова, Д.А. Алиева, Т.Ж. Тохтарова (2026). Технологии ИИ в языковом образовании: практические аспекты и проблемы применения студентами // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 83–96. https://doi.org/10.54309/IJICT.2026.26.2.005. (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 integration of Artificial Intelligence (AI) technologies into higher education is precipitating a transformation not only of instructional tools but of the very architecture of learning activities. This shift is particularly salient in language education, where the core processes of textual interpretation, argumentation, and communicative competence remain paramount. Among the most pervasive instruments are generative language models and educational chatbots that provide instantaneous feedback and personalized access to pedagogical content (Amaro, 2023; Kohnke, 2023). Within the framework of the present study, the «Русский язык_ментор» chatbot serves as a representative tool designed to facilitate explanation, training, and the methodological scaffolding of student learning.

While extant literature focuses heavily on AI's didactic potential, empirical data regarding authentic student engagement remains fragmented. Specifically, there is an evidentiary gap concerning student usage strategies, prompt formulation, and the degree of critical perception toward AI-generated outputs. This contradiction is most visible in the training of technical students, where AI is often employed to expedite task completion at the expense of analytical depth, fostering superficial learning strategies. Initial piloting suggests that student-AI interaction is non-homogeneous: while some utilize the tool for verification, a significant trend toward uncritical acceptance of AI-generated responses is observed, whereas instances of independent cognitive refinement remain infrequent.

The relevance of this study is underscored by the urgent need to analyze these real-world practices and identify their influence on the development of linguistic competence. Theoretically, this work seeks to redefine the role of AI within the modern linguodidactic paradigm not merely as an auxiliary tool, but as a transformative factor in academic engagement. The object of this study is the process of language acquisition within the context of integrating Artificial Intelligence technologies into the higher education environment. The subject of this research encompasses specific student-AI interaction patterns, including usage strategies, the nature of prompt formulation, and emerging cognitive and methodological challenges.

The primary aim is to identify the practical dimensions of AI usage in language learning and to demonstrate that, absent methodologically organized scaffolding, AI application shifts learning activities toward the streamlining of cognitive operations and a reduction in analytical depth.

To achieve the stated aim and systematically investigate the influence of AI on the structure of student learning, the following research objectives were formulated and addressed:

- To define the functional and didactic capabilities of the chatbot as a pedagogical tool;
- To investigate authentic student-AI interaction patterns within authentic learning environments;
- To identify typical usage strategies, such as verification, clarification, or verbatim borrowing;
- To determine the correlation between task complexity and the quality of AI-

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

generated responses;

- To isolate the fundamental limitations of AI in performing analytical and interpretative tasks.

The scientific novelty lies in the empirical identification of the «asymmetry» between task performance and conceptual mastery, providing a nuanced taxonomy of real-world student-AI interaction strategies. The practical significance involves the development of principles for methodologically grounded AI integration in language education, oriented toward preserving and fostering students' analytical thinking and cognitive autonomy.

This study is structured around the following research questions:

1. What strategies for interacting with AI tools do students develop while performing linguistic tasks?

2. How does the nature of the learning assignment influence the quality and comprehensiveness of AI-generated responses?

3. To what extent do students engage in critical processing of AI outputs versus uncritical reproduction?

The hypothesis posits that active AI utilization creates a disconnect between the correctness of the final output and the depth of the student's understanding. We anticipate that while students demonstrate high success in reproductive tasks, they encounter significant difficulties in assignments requiring analysis, argumentation, and interpretation, a phenomenon attributed both to the constraints of large language models and the current absence of established strategies for critical AI engagement.

Materials and methods.

The primary research instrument utilized in this study was the «Русский язык_ментор» (available at https://chatgpt.com/g/g-698addcc621481919f70a0c695b42cfb-russkii-iazyk-mentor), an educational chatbot developed on the ChatGPT platform as part of the internal university project: *«Integration of Artificial Intelligence (AI) into the Development of Professional Competencies of IITU Students through Language Education.»* This specialized virtual assistant implements instructional, evaluative, and developmental functions aimed at fostering both linguistic proficiency and critical thinking skills. Methodologically, the tool is predicated on the 15-week syllabus of the «Russian Language» course (Level B), where each query is structured by study week to ensure logical consistency. Within each weekly module, the system provides 4–5 sequential tasks that transition the learner from foundational knowledge to complex application, specifically targeting grammatical, lexical, and communicative competencies alongside analytical and argumentative capabilities.

The research was conducted using a focus group format, shifting the analytical emphasis from final learning outcomes to the interaction process itself. This qualitative approach facilitated the direct observation of student behavioral strategies among 30 undergraduate students (1st and 2nd year) from the International IT University (IITU). The sample was characterized by a heterogeneous level of language proficiency, enabling the identification of variations in chatbot usage-specifically how the system's dialogue-based model provided immediate formative feedback and diagnostic reviews of student responses.

The study adhered to strict ethical and data privacy protocols. Formal institutional

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

approval was obtained as part of the university's internal research framework. Prior to the focus group sessions, all participants were briefed on the research objectives and provided informed consent. Participation was entirely voluntary, ensuring that the use of the AI tool did not impact academic grading. To safeguard student privacy, all recorded interactions and pedagogical artifacts were de-identified, with personal identifiers removed during the analysis phase to maintain participant anonymity.

The research process was structured into three distinct operational phases:

- The Preparatory Phase: encompassing the development of specialized pedagogical tasks and the conceptual modeling of chatbot integration;
- The Empirical Phase: involving the piloting of the chatbot via focus group sessions to capture real-time interaction data;
- The Interpretative Phase: consisting of the qualitative synthesis and critical analysis of the gathered empirical evidence.

To address the research objectives, a multifaceted and triangulated methodological framework was employed, including an extensive literature review, direct participant observation, and a product analysis of learning artifacts such as presentations and essays. Qualitative content analysis was utilized to isolate recurring behavioral paradigms, complemented by a comparative analysis of AI outputs relative to task complexity and an interpretative analysis based on curriculum-standard correlations. Ultimately, this comprehensive methodology facilitates a robust analysis of student-AI dynamics within an authentic educational environment, establishing a definitive foundation for identifying the systemic patterns inherent in AI-mediated language acquisition.

Results and discussion.

Fundamental research regarding the application of Artificial Intelligence (AI) in language education emphasizes the transformative potential of chatbots for personalized learning and oral proficiency. Scholars analyze ChatGPT as a sophisticated instrument for generating contextually relevant responses (Baskara et al., 2023), noting its capacity to improve lexical acquisition while addressing cognitive load (Xiao, 2024; Feng, 2024). Recent investigations (2024–2026) in the domain of Russian as a Foreign Language (RFL) further underscore generative AI's role in providing granular feedback and simulating naturalistic dialogue (Yiğit, 2024; Becerik-Gerber, 2025). However, while these studies reveal broad functional advantages-such as the automation of routines scholarly evidence remains limited regarding the specific requirements of technical disciplines and the cognitive strategies non-philological students employ during interaction. This gap is particularly relevant when considering the cognitive-discursive processes of IT specialists in digital society, where linguistic competence must be integrated into high-complexity technical workflows (Kenzhebulatova, 2025).

Student engagement with AI manifests through diverse formats characterized by varying degrees of cognitive involvement (Holmes et al., 2019; Sun et al., 2024). Certain practices align with a «delegating» model, where AI serves as a source of pre-packaged solutions a burgeoning phenomenon of cognitive outsourcing (Mollick, 2024). According to Cognitive Load Theory, such reliance on automated outputs may result in a significant

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

reduction in the depth of information processing, potentially hindering long-term knowledge retention (Sweller, 1988).Conversely, the «verificative» strategy utilizes AI as a mechanism for validation and refinement, corresponding with the Technology Acceptance Model (TAM) where AI is perceived as a utilitarian tool that enhances user confidence (Davis, 1989). A third research trajectory positions AI as a collaborative partner in the co-construction of knowledge, emphasizing the necessity of active information synthesis (Favero et al., 2026). Building upon these foundations, the educational chatbot developed for this study was intentionally designed to avoid finalized solutions, instead engaging students in iterative processes of analysis and clarification to catalyze cognitive activation rather than displacement (Zhang, 2025).

This investigation utilized a focus group methodology to achieve a profound analysis of chatbot utilization. By moving beyond aggregate statistical data, the format captured nuanced student responses and lived experiences, uncovering difficulties often obscured by the standardized constraints of traditional surveys (Jung et al., 2019; Belanovsky, 2001; Chekhovsky, 2017; Renandya, 2024). The participants comprised 30 undergraduate students from the International IT University (IITU). During the empirical phase, it became evident that linguistic proficiency significantly influenced interaction: in 18 cases, the chatbot functioned primarily as a verification tool, whereas in 12 instances, it served as a source for pre-packaged linguistic formulations.

The research targeted technical students for whom linguistic competence is critical but not a primary field of study. Despite a robust conceptual understanding, a significant portion of participants encountered difficulties in synthesizing thoughts or selecting appropriate rhetorical models. Utilizing an activity-based inquiry where students independently formulated queries, observations revealed that 21 out of 30 analyzed artifacts were dominated by clarifying and verificative queries, whereas advanced requests for complex synthesis were recorded in only 9 instances. System responses were analyzed via longitudinal transcripts, with comparative evaluation benchmarking outputs against course criteria such as definitional precision. This approach allowed researchers to bypass «social desirability bias» and capture authentic usage: uncritical acceptance of AI outputs occurred in 19 presentations, while proactive engagement was evident in 11 works. While response velocity and structural coherence were highlighted in 26 cases, limitations regarding analytical superficiality were identified in 22 instances.

Notably, student-AI interaction followed the internal logic of pedagogical tasks, occurring at specific points of «cognitive friction» such as terminological deconstruction. Data analysis reveals that in 24 out of 30 analyzed focus group artifacts, query formulations mirrored curriculum prompts with negligible modifications. This suggests the chatbot was perceived by the technical student cohort as a functionally integrated component of academic activity rather than a peripheral tool. However, behavioral variations persisted: 18 participants utilized the system primarily to clarify specific sub-components before independent refinement, while 12 participants regarded the generated response as a definitive, finalized solution. These synthesized findings, which categorize the nature of student engagement and system efficacy, are detailed in Table 1.

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

Table 1 – Typology of Chatbot Queries

Query Category	Task Exemplar	AI Response Characteristics	Pedagogical Efficacy
Conceptual	«Define a micro-theme in a scientific text. How does it differ from the primary theme?»	Concise definitions, comparative synthesis, and illustrative examples.	Formation of foundational theoretical concepts and terminological clarity.
Analytical	«Identify the communicative objective of the provided text regarding AI in medicine.»	Validation of student hypotheses and generalized feedback on discourse intent.	Partial advancement of analytical skills and interpretive accuracy.
Structural	«Deconstruct the provided text on scientific advancement into its themes and micro-themes.»	Accurate structural breakdown, though often lacking an explanation of the underlying methodology.	Identification of text hierarchy with limited procedural understanding of the method.
Logical	«Identify the logical contradiction within this argument regarding scientific progress.»	Precise identification of logical fallacies or inconsistencies in reasoning.	Cultivation of critical inquiry and the development of logical reasoning patterns.
Applied	«Define and integrate the terms ‘digitalization’ and ‘fundamental science’ into a professional context.»	Comprehensive definitions complemented by contextual examples and supplementary tasks.	Expansion of academic engagement and practical application of specialized vocabulary.

As demonstrated in Table 1, chatbot performance was most consistent at the Conceptual level. When engaged with foundational terminology-such as «micro-theme» or «scientific discourse»-the system provided reliable, standardized outputs. In a representative interaction, the chatbot defined a micro-theme as a «semantic unit disclosing a specific aspect of the primary theme» and successfully operationalized the concept within the context of «climate change.» Focus group participants frequently utilized the system for iterative validation, using follow-up inquiries like: «How does a micro-theme differ from a sub-topic?» These refinements confirmed the system’s internal logic and consistency. From a didactic perspective, this deductive movement serves as a productive scaffold during the introductory phase of new conceptual frameworks, rendering AI outputs «operationally accessible.»

However, granularity diminishes during analytical tasks involving the interpretation of Information Technology (IT) discourse. This is evidenced by a reduced level of textual grounding. When prompted to «justify a conclusion based on linguistic features,» the chatbot validated the student’s identification of an informative function but remained abstract (e.g., «the text elucidates algorithmic principles»). It failed to reference specific diagnostic markers such as terminological density or generalizing syntactic structures. Even when students attempted to «strengthen» prompts by requesting specific sentences, the output remained unchanged, providing isolated fragments without diagnostic analysis. For example, highlighting the term «learning algorithm» without explicating its relationship to the text’s informative orientation provides a finalized result but fails to facilitate evidence-based reasoning-a cornerstone of academic literacy.

A comparable phenomenon appeared in structural decomposition. While the chatbot accurately delineated micro-themes in neural network texts, it reverted to generalized

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

descriptions when prompted for diagnostic criteria. It failed to reference objective markers like lexical keywords or logical transitions. Conversely, tasks involving logical inconsistencies demonstrated higher precision. The system accurately isolated deliberate contradictions by relying on formalistic inconsistency detection rather than nuanced discourse interpretation.

Regarding IT terminology, the system demonstrated significant applied potential by providing accurate definitions for terms like «machine learning.» However, systemic limitations emerged as task complexity increased. Requests for context-specific applications often resulted in «stylistic attenuation,» where formulations became less precise or incorporated colloquial elements. Participants characterized these outputs as «comprehensible but insufficiently academic.» This transition from stable terminological definitions to attenuated analytical outputs underscores a direct correlation between task complexity and response granularity (see Table 2).

Table 2 – Task Complexity vs. Chatbot Response Quality

Complexity Level	Qualitative Characteristics of Response	Primary Functional Limitation
Foundational (Basic)	Precise, highly structured, and contextually accurate.	No significant limitations identified.
Intermediate (Average)	Factually correct but characterized by high-level abstraction.	Lack of direct evidence-based reference to the source text.
Advanced (Enhanced)	Partially comprehensive with intermittent accuracy.	Omission of critical details and absence of analytical logic.
Expert (Complex)	Primarily directional or heuristic; lacks exhaustive depth.	Absence of definitive solutions and failure to demonstrate methodology.

At the foundational level, encompassing queries such as «explain the term» or «provide a definition,» the chatbot demonstrates its most consistent results. However, as complexity shifts toward intermediate tasks-requiring academic justification a transition toward generalized formulations occurs. This functional divergence is synthesized in Figure 1.

Empirical evidence from focus group participants (e.g., Student A, 2nd year; see Figure 1A) confirms that while the chatbot successfully validates student responses through a structured feedback loop (status, explanation, follow-up), it often lacks granular textual evidence. As noted in the student's report, the system «does not explain in detail why the answer is correct» and fails to «highlight specific phrases from the text,» reinforcing the identified «procedural gap» between correct output and methodological mastery.

Furthermore, when transitioned to intermediate tasks requiring morphological precision (see Figure 1B), the gap becomes more pronounced. While the specialized mentor remains accurate and concise, the general-purpose LLM provides a higher degree of grammatical systematicity. This suggests that the chatbot functions primarily as an introductory scaffolding tool rather than an exhaustive linguistic reference.

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

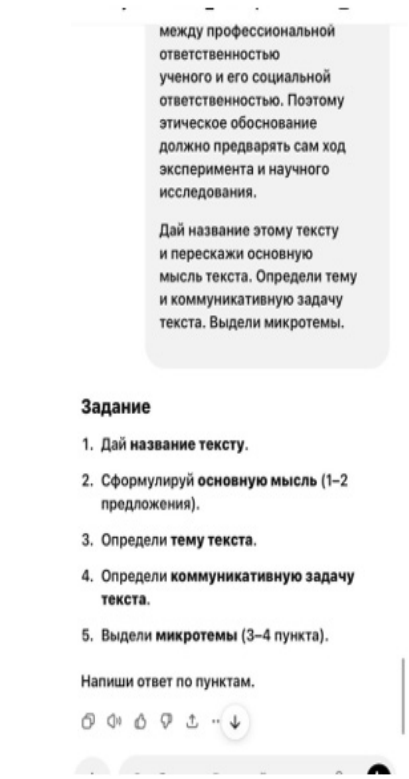


Fig. 1A. Instructional Scaffolding Automated Validation

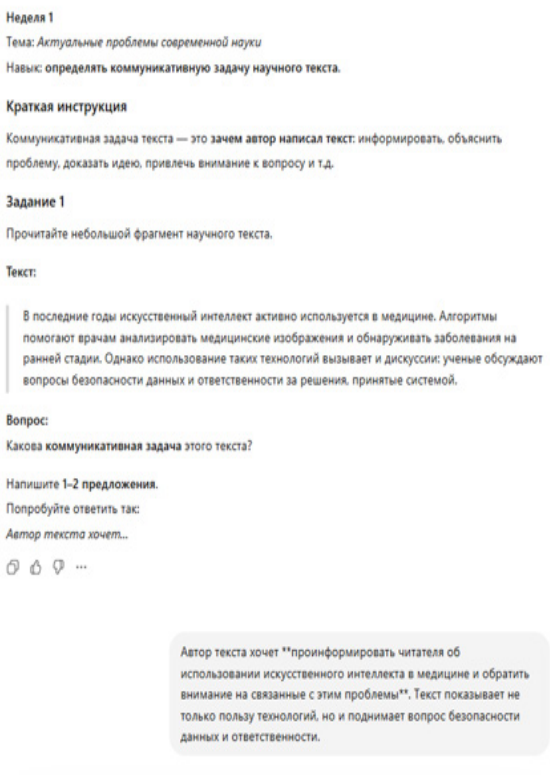


Fig. 1B. Comparative Morphological Analysis and and Procedural gap

At advanced levels, system responses often fragment, prompting a dualistic evaluation. Student B (2nd year) noted that on complex tasks, the chatbot merely repeated rules rather than completing assignments. Participants reinterpreted this as a «productive constraint» that prevents «cognitive outsourcing» and directs students toward independent work. Furthermore, Student C (1st year; see Figure 2) highlighted functional distinctions: while general LLMs offer superior systematicity for «deeper classification,» the specialized «Русский язык_ментор» is favored for its accessible, simplified delivery. Ultimately, the specialized bot serves as an introductory scaffolding mechanism rather than an exhaustive linguistic reference.

The most telling limitations are observed in high-complexity tasks necessitating a synthesis of analysis and interpretation. Students formulated sophisticated queries such as: *«Prove the popular-science nature of the text» or «Compare two paragraphs and explicate their logical connection.»* In these scenarios, the system generally fails to develop a comprehensive solution, instead reverting to a theoretical baseline by reproducing definitions. For example, when asked to «prove the communicative objective using specific phrases,» the chatbot often paraphrases the definition, effectively bypassing the requirement for empirical evidence.

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

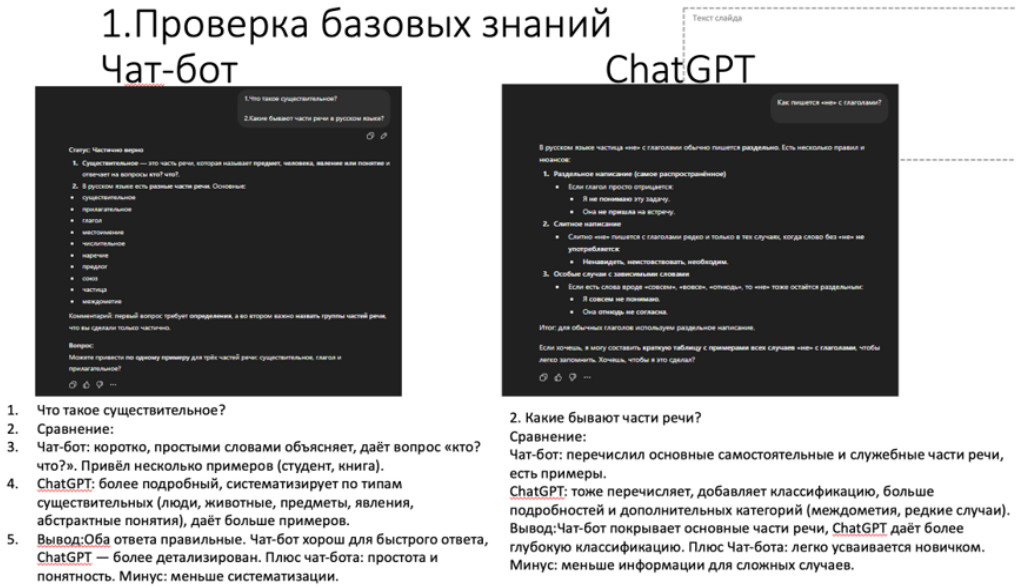


Fig. 2. Student-generated comparative analysis slide (Student C, 1st year)

This aspect elicited divergent evaluations from the participant cohort. At the level of immediate feedback, the behavior was perceived as a functional deficit; students noted that the output felt «unfinished» or «did not complete the assignment.» However, during in-depth oral discussions, a nuanced perspective emerged. Some students reinterpreted this limitation as a productive constraint-a situation necessitating a return to the source text to independently construct argumentation. As one participant noted: *«the bot sets the direction but does not do the work for you, which forces you to think.»*

From a pedagogical standpoint, this phenomenon is significant when viewed through the framework of instructional scaffolding. The chatbot functions as a cognitive anchor: it actualizes theoretical knowledge and delineates analytical trajectories without supplanting the student's intellectual agency. Unlike «ready-made» solutions, this interaction preserves the requirement for the student to complete the mental process.

Nevertheless, efficacy is contingent upon prior proficiency and subsequent pedagogical mediation. Without these, partial responses may be mistaken for finalized outputs, leading to fragmented understanding. This is evident in tasks where the absence of a detailed explanation may go unnoticed by the learner. Ultimately, these findings suggest the strategic utility of a blended learning model. The AI fulfills functions of preliminary preparation and self-verification, while the instructor provides in-depth analysis and quality control.

A consistent finding is the asymmetry between response accuracy and the depth of substantiation. The chatbot demonstrates high precision in defining concepts but exhibits «collapsed» reasoning in interpretation tasks. These results corroborate research by Baskara & Mukarto (2023) and Yiğit (2024) regarding the systemic limitations of Large

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

Language Models (LLMs) in replicating knowledge.

The influence of chatbots is inherently dualistic. They streamline information access for technical students facing high workloads but may inadvertently reinforce superficial learning strategies in the absence of mediation. The focus group results underscore the necessity of transitioning from situational AI use toward methodologically rigorous curriculum integration. This requires a task system that accounts for the tool's limitations and the purposeful cultivation of AI literacy, enabling students to critically evaluate, cross-reference, and refine AI-generated outputs through independent analytical effort.

Conclusion.

Artificial Intelligence (AI) instruments have become an intrinsic component of contemporary academic practice, perceived by students as a natural extension of the educational environment. This study provided a comprehensive examination of AI's role in language education within a technical university, where the cohort's preference for velocity and efficiency significantly shapes chatbot interaction. Such findings necessitate the development of disciplined, evidence-based approaches to cultivating AI literacy.

The scientific novelty of this research lies in the empirical substantiation of the «procedural gap» and its pedagogical consequences. The data demonstrates that utilizing chatbots without structured methodological support may result in superficial understanding, where students reproduce accurate results without mastering the underlying analytical mechanisms.

Future trajectories for this research, conducted under the project *«Integration of Artificial Intelligence (AI) in the Development of Professional Competencies of IITU Students through Language Education,»* are multidimensional. The identified usage patterns allow for a transition from a declarative analysis of attitudes toward a rigorous study of actual interaction strategies within the educational ecosystem. This provides the basis for a comprehensive typology of AI utilization models-ranging from supplementary toolsets to integrated instructional participants.

A promising direction for future inquiry involves a comparative analysis between proficiency levels (e.g., Level A vs. Level B) to identify qualitative distinctions in how AI is leveraged across different stages of linguistic development. Furthermore, there is significant potential to operationalize these findings into practice-oriented models for educators, encompassing structured task scenarios and strategic interventions designed to mitigate the risks of cognitive outsourcing and the erosion of student autonomy.

Additionally, a compelling trajectory involves investigating the longitudinal cognitive transformations resulting from sustained AI interaction, including shifts in academic motivation and metacognitive reflection. Expanding the empirical database to include varied instructional modalities will enhance the generalizability of these patterns.

Finally, a pivotal direction lies in leveraging chatbots to foster an inclusive educational ecosystem. Aligning with UNESCO (2021), AI can lower barriers for learners with diverse requirements, such as dyslexia or physical disabilities, by functioning as a vital support mechanism. Integrating AI within inclusive pedagogy presents an opportunity to design adaptive instructional scenarios that prevent information saturation. Ultimately,

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

this research establishes the empirical foundation for novel pedagogical models centered on the purposeful and effective integration of AI within language education.

REFERENCES

Amaro I., Della Greca A., Francese R., Tortora G., Tucci C. (2023). AI unreliable answers: A case study on ChatGPT // Artificial Intelligence in HCI. Pp. 23–40. DOI: 10.1007/978-3-031-35894-4_2 [In Eng.]

Baskara G.B., Mukarto M. (2023). Exploring the Implications of ChatGPT for Language Learning in Higher Education // ERIC. Pp. 1–15. DOI:10.21093/ijeltal.v7i2.1387 [In Eng.]

Becerik-Gerber B. et al. (2025). AI-powered scaffolding: Rethinking English language teaching practices in the digital age // Global Journal of Foreign Language Teaching. Pp. 45–58. DOI:10.66355/rzwgva56 [In Eng.]

Belanovskiy S.A. (2001). Metod fokus-grupp: ucheb. posob. (Focus Group Method: Study Guide). — M.: Nikkolo-Media. Pp. 1–280. [In Eng.]

Cheng J. (2025). Research status of artificial intelligence in Russian language teaching: a literature review // Region - Educational Research and Reviews. — Vol. 7. — No. 7. DOI: 10.32629/rerr.v7i7.4285 [In Eng.]

Chekhovskiy I.V. (2009). O znachimosti oshibok, voznikayushchikh pri ispol'zovanii metoda fokus-grupp (On the significance of errors arising in the use of the focus group method) // RUDN Journal of Sociology. No. 2. Pp. 83–92. [In Rus.].

Davis F. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology // MIS Quarterly. — Vol. 13. — No. 3. Pp. 319–340. DOI: 10.2307/249008[In Eng.]

Favero L., Pérez-Ortiz J.A., Käser T. et al. (2026). AI in Education Beyond Learning Outcomes. — arXiv. DOI: 10.48550/arXiv.2602.04598 [In Eng.]

Feng L. (2024). Investigating the effects of artificial intelligence-assisted language learning strategies on cognitive load and learning outcomes: A comparative study // Journal of Educational Computing Research. — Vol. 62. — No. 8. Pp. 1741–1774. — DOI: 10.1177/07356331241268349 [In Eng.]

Holmes W., Bialik M. & Fadel C. (2019). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning // ResearchGate, 2019. DOI: 10.58863/20.500.12424/4276068 [In Eng.]

Jung H. & Ro E. (2019). Validating common experiences through focus group interaction // Journal of Pragmatics, 2019. — Vol. 143. — Pp. 169–184. DOI: 10.1016/j.pragma.2019.02.019. [In Eng.]

Kohnke L., Moorhouse B. L., Zou D. (2023). ChatGPT for language teaching and learning // RELC Journal. — Vol. 54. — No. 2. Pp. 537–550. DOI: 10.1177/0033688223116 [In Eng.]

Kenzhebulatova D., Abdraim F., Bublikova O., Togzhanova L., Rassilova A. (2025). Cognitive-discursive processes of IT specialists in the digital society // Procedia Computer Science. — Vol. 272. — Pp. 392-397. ISSN 1877-0509. DOI: 10.1016/j.procs.2025.10.220. [In Eng.]

Mollick E. & Mollick L. (2023). Assigning AI: Seven Approaches for Students. — arXiv. DOI: 10.48550/arXiv.2306.10052 [In Eng.]

OpenAI. (2024). Russian Language_Mentor (Version GPT-4): Custom GPT // OpenAI Platform. URL: https://chatgpt.com/g/g-698addcc621481919f70a0c695b42cfb-russkii-iazyk-mentor (in Eng.)

Renandya W. (2024). Modeling and scaffolding using AI in language teaching. — Willy Renandya Blog. Pp. 1–8. [In Eng.]

Sun Y., Wu Y. (2026). The efficacy of artificial intelligence - powered scaffolding in individual acquisition efficiency of EFL in tertiary educational context. Front Psychol., 2026. — Vol. 16. — DOI: 10.3389/fpsyg.2025.1613285. [In Eng.]

Sweller J. (1988). Cognitive Load During Problem Solving: Effects on Learning. — Cognitive Science. — Vol. 12. — No. 2. Pp. 257–285. DOI: 10.1207/s15516709cog1202_4[In Eng.]

UNESCO (2021). Artificial Intelligence in Education: Guidance for Policy-makers. — UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000376703 [In Eng.]

Xiao L. (2024). The promises and challenges of AI-based chatbots in education // PMC. Pp. 1–12. DOI:10.1016/j.heliyon.2024.e37238 [In Eng.]

Yiğit Ç. (2024). Potentials and applications of ChatGPT in foreign language education. — RIUNET. Pp. 1–20. — DOI:10.4995/eurocall.2024.19957 [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.2. Number 26 (2026). Pp. 97–108
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.006
IRSTI / ГТАХР / МРНТИ / 004.931

HYBRID STACKING FRAMEWORK FOR CROP CLASSIFICATION USING UAV DATA

G. Azieva[1], M. Yessenova[2]*, A. Abzhapparova[3], G. Abdikerimova[2], P. Schmidt[4]
[1]Esil University, Astana, Kazakhstan;
[2]Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan;
[3]M.Kh. Dulaty Taraz University, Taraz, Kazakhstan;
[4]University of Economics, Bratislava, Slovakia.
E-mail: moldir_11.92@mail.ru

Gulmira Azieva — Senior Lecturer, Department of Information Systems and Technologies, ESIL Universety, Astana, Kazakhstan
https://orcid.org/ 0000-0002-7329-6768;
Moldir Yessenova — PhD, Senior Lecturer, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: moldir_11.92@mail.ru, https://orcid.org/0000-0002-2644-0966;
Aigerim Abzhapparova — Senior Lecturer of the Department of Physics and informatics, M.Kh. Dulaty Taraz University, Taraz, Kazakhstan
https://orcid.org/0000-0002-7500-9179;
Gulzira Abdikerimova — PhD, Associate Professor, Department of Information Systems, Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan
https://orcid.org/0000-0002-4953-0737;
Peter Schmidt — doctor of Philosophy (PhD), professor, head of the Department of Applied Informatics, Bratislava, University of Economics, Bratislava, Slovak Republic
https://orcid.org/0000-0001-5928-2821.

© G. Azieva, M. Yessenova, A. Abzhapparova, G. Abdikerimova, P. Schmidt

Abstract. This study presents a two-level hybrid stacking framework for automatic crop classification using multispectral UAV orthomosaics. The proposed architecture combines several gradient boosting methods (LightGBM, XGBoost, and CatBoost), classical tree-based ensemble models (Random Forest and Extra Trees), and an Attention-based multilayer perceptron. During training, out-of-fold probability estimates generated by the base learners are used as inputs to a second-level Extra Trees meta-classifier, which produces the final prediction. The feature representation includes original spectral bands together with eight widely used vegetation indices (NDVI, NDRE, GNDVI,

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

SAVI, MSR, EVI, SIPI, and MSAVI). For each index and band, statistical descriptors such as mean and standard deviation are calculated to capture spatial variability within field segments. The experimental evaluation was carried out on real multispectral UAV data collected in Eastern Kazakhstan. The proposed model achieved high classification accuracy (approximately 95 %) and a macro-averaged F1-score close to 0.95. In addition, full-field spatial segmentation results demonstrated stable performance at the level of about 99%, indicating strong consistency across the study area. Overall, the findings suggest that combining heterogeneous machine learning models within a stacking framework improves robustness and generalization in crop classification tasks.

Keywords: crop classification, multispectral UAV data, orthomosaics, vegetation indices, hybrid stacking, two-level ensemble, machine learning, precision agriculture, spatial segmentation, digital agriculture

For citation: G. Azieva, M. Yessenova, A. Abzhapparova, G. Abdikerimova, P. Schmidt (2026). Hybrid stacking framework for crop classification using UAV data // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 97–108. https://doi.org/10.54309/IJICT.2026.26.2.006. (In Kaz.).

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

Acknowledgment: *This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP25793889 "Development of a system for identifying diseases and pests of agricultural crops based on artificial intelligence").*

UAV ДЕРЕКТЕРІ НЕГІЗІНДЕ АУЫЛ ШАРУАШЫЛЫҒЫ ДАҚЫЛДАРЫН ЖІКТЕУГЕ АРНАЛҒАН ГИБРИДТІ СТЕКИНГ МОДЕЛІ

Г.Т. Азиева[1], М.Б. Есенова[2]*, А.К. Абжаппарова[3], Г.Б. Абдикеримова[2], P. Schmidt[4]

[1]Esil University, Астана, Қазақстан;
[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[3]М.Х. Дулати атындағы Тараз университеті, Тараз, Қазақстан;
[4]Братислава экономикалық университеті, Словакия.
E-mail: moldir_11.92@mail.ru

Азиева Гульмира Тагибергеновна — магистр, «Ақпараттық жүйе және технологиялар» кафедрасының аға оқытушысы, Esil University, Астана, Қазақстан https://orcid.org/ 0000-0002-7329-6768;

Есенова Молдир Балкаировна — PhD, Ақпараттық жүйелер кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
E-mail: moldir_11.92@mail.ru, https://orcid.org/0000-0002-2644-0966;

Абжаппарова Айгерим Кахармановна — «Физика және информатика» кафедрасының аға оқытушысы, М.Х.Дулати атындағы Тараз университеті, Тараз, Қазақстан

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

https://orcid.org/0000-0002-7500-9179;
Абдикеримова Гульзира Бахытбековна — PhD, Ақпараттық жүйелер кафедрасының доценті, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0000-0002-4953-0737;
P. Schmidt — философия докторы (PhD), профессор, «Қолданбалы информатика» кафедрасының меңгерушісі, Братислава экономикалық университеті, Братислава қ., Словакия
https://orcid.org/0000-0001-5928-2821.

© Г.Т. Азиева, М.Б. Есенова, А.К. Абжаппарова, Г.Б. Абдикеримова, P. Schmidt

Аннотация. Бұл зерттеуде UAV-дан алынған мультиспектрлі ортомозаикаларды қолдана отырып, дақылдарды автоматты түрде жіктеуге арналған екі қабатты гибридті қабаттастыру құрылымы ұсынылған. Ұсынылған архитектура бірнеше градиент күшейту әдістерін (LightGBM, XGBoost және CatBoost), классикалық ағаш негізіндегі ансамбльдік модельдерді (Random Forest және Extra Trees) және назар аудару механизміне негізделген көп қабатты перцептронды біріктіреді. Оқыту кезінде базалық жіктеуіштермен алынған ықтималдық бағалары екінші қабатты метакласификаторға (Extra Trees) кіріс ретінде пайдаланылады, бұл соңғы болжамды жасайды. Белгілердің көрінісі бастапқы спектрлік жолақтарды сегіз кеңінен қолданылатын өсімдік индекстерімен (NDVI, NDRE, GNDVI, SAVI, MSR, EVI, SIPI және MSAVI) бірге қамтиды. Әрбір индекс пен жолақ үшін орташа және стандартты ауытқу сияқты статистикалық сипаттамалар далалық сегменттердегі кеңістіктік өзгергіштікті ескеру үшін есептеледі. Шығыс Қазақстанда UAV алған нақты әлемдегі мультиспектрлі деректер бойынша эксперименттік бағалау жүргізілді. Ұсынылған модель жоғары жіктеу дәлдігіне (шамамен 95 %) және 0,95-ке жақын макроорташа F1-метрикалық ұпайына қол жеткізді. Сонымен қатар, бүкіл егістік бойынша кеңістіктік сегменттеу нәтижелері шамамен 99 % тұрақты өнімділікті көрсетті, бұл бүкіл зерттеу аймағында жоғары тұрақтылықты көрсетеді. Жалпы алғанда, нәтижелер гетерогенді машиналық оқыту модельдерін қабаттасқан құрылымда біріктіру дақылдарды жіктеу тапсырмаларында беріктік пен жалпылау мүмкіндігін жақсартатынын көрсетеді.

Түйінді сөздер: ауыл шаруашылығы дақылдарын жіктеу, мультиспектралды UAV деректері, орттомозаика, вегетациялық индекстер, гибридті стекинг,екі деңгейлі ансамбль, машиналық оқыту, дәл егіншілік, кеңістіктік сегментация, цифрлық ауыл шаруашылығы

Дәйексөздер үшін: Г.Т. Азиева, М.Б. Есенова, А.К. Абжаппарова, Г.Б. Абдикеримова, P. Schmidt (2026). UAV деректері негізінде ауыл шаруашылығы дақылдарын жіктеуге арналған гибридті стекинг моделі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. 97–108. https://doi.org/10.54309/IJICT.2026.26.2.006. (Қаз. 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

Алғыс: *Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырады (Грант № AP25793889 «Жасанды интеллект негізінде ауыл шаруашылығы дақылдарының аурулары мен зиянкестерін анықтау жүйесін әзірлеу»).*

ГИБРИДНАЯ МОДЕЛЬ СТЕКИНГА ДЛЯ КЛАССИФИКАЦИИ СЕЛЬСКОХОЗЯЙСТВЕННЫХ КУЛЬТУР ПО ДАННЫМ UAV

Г.Т. Азиева[1], М.Б. Есенова[2]*, А.К. Абжаппарова[3], Г.Б. Абдикеримова[2], P. Schmidt[4]
[1]Esil University, Астана, Республика Казахстан;
[2]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;
[3]Таразский университет имени М.Х.Дулати, Тараз, Казахстан;
[4]Братиславский экономический университет, Словакия.
E-mail: moldir_11.92@mail.ru

Азиева Гульмира Тагибергеновна — магистр, старший преподаватель кафедры «Информационные системы и технологии», Esil University, Астана, Казахстан
https://orcid.org/ 0000-0002-7329-6768;
Есенова Молдир Балкаировна — PhD, старший преподаватель кафедры информационных систем Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан
E-mail: moldir_11.92@mail.ru, https://orcid.org/0000-0002-2644-0966;
Абжаппарова Айгерим Кахармановна — старший преподаватель кафедры «Физика и информатика», Таразский университет имени М.Х.Дулати, Тараз, Казахстан
https://orcid.org/0000-0002-7500-9179;
Абдикеримова Гульзира Бахытбековна — PhD, доцент кафедры информационных систем Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0002-4953-0737;
P. Schmidt — доктор философии (PhD), профессор, заведующий кафедрой «Прикладная информатика», Братиславского экономического университета, г. Братислава, Словацкая Республика
https://orcid.org/0000-0001-5928-2821.

© Г.Т. Азиева, М.Б. Есенова, А.К. Абжаппарова, Г.Б. Абдикеримова, P. Schmidt

Аннотация. В данном исследовании представлена двухуровневая гибридная структура стекирования для автоматической классификации сельскохозяйственных культур с использованием мультиспектральных ортомозаик, полученных с помощью БПЛА. Предложенная архитектура сочетает в себе несколько методов градиентного бустинга (LightGBM, XGBoost и CatBoost), классические ансамблевые модели на основе деревьев (Random Forest и Extra Trees) и многослойный персептрон на основе

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

механизма внимания. Во время обучения оценки вероятности, полученные базовыми классификаторами, используются в качестве входных данных для метаклассификатора второго уровня Extra Trees, который выдает окончательное предсказание. Представление признаков включает исходные спектральные полосы вместе с восемью широко используемыми индексами растительности (NDVI, NDRE, GNDVI, SAVI, MSR, EVI, SIPI и MSAVI). Для каждого индекса и полосы рассчитываются статистические дескрипторы, такие как среднее значение и стандартное отклонение, для учета пространственной изменчивости внутри сегментов поля. Экспериментальная оценка проводилась на реальных мультиспектральных данных, полученных с помощью БПЛА в Восточном Казахстане. Предложенная модель достигла высокой точности классификации (приблизительно 95%) и макроусредненного F1-мерного показателя, близкого к 0,95. Кроме того, результаты пространственной сегментации всего поля продемонстрировали стабильную производительность на уровне около 99%, что указывает на высокую согласованность по всей исследуемой области. В целом, результаты показывают, что объединение гетерогенных моделей машинного обучения в рамках стековой структуры повышает устойчивость и обобщающую способность в задачах классификации сельскохозяйственных культур.

Ключевые слова: классификация сельскохозяйственных культур, мультиспектральные данные БПЛА, ортомозаики, вегетационные индексы, гибридный стекинг, двухуровневый ансамбль, машинное обучение, точное земледелие, пространственная сегментация, цифровое сельское хозяйство

Для цитирования: Г.Т. Азиева, М.Б. Есенова, А.К. Абжаппарова, Г.Б. Абдикеримова, P. Schmidt (2026). Гибридная модель стекинга для классификации сельскохозяйственных культур по данным UAV // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 97–108. https://doi.org/10.54309/IJICT.2026.26.2.006. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность: *Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант No AP25793889 «Разработка системы выявления болезней и вредителей сельскохозяйственных культур на основе искусственного интеллекта»)*

Кіріспе.

Соңғы жылдары ұшқышсыз ұшу аппараттарына (UAV) негізделген мультиспектралды түсірілім дәл егіншілік жүйелерінде кеңінен қолданылып келеді. Бұл технология дақылдардың жай-күйін бағалауға, өсу динамикасын бақылауға және агротехникалық шешімдер қабылдауға мүмкіндік береді (Shu et al., 2022; Pádua et al., 2020). Жоғары кеңістіктік рұқсатқа ие UAV-ортомозаикалары өсімдік жамылғысының құрылымдық және спектралдық ерекшеліктерін айқын көрсетеді. Сондықтан мұндай деректер ауыл шаруашылығы дақылдарын автоматты түрде жіктеу міндетінде маңызды ақпарат көзі болып саналады (Huang et al., 2025; Deng et al., 2024). Сонымен бірге спектралдық өзгергіштік, жарықтандыру жағдайының

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

тұрақсыздығы және бір класс ішіндегі әркелкілік классификация нәтижесіне елеулі әсер етеді (Tan et al., 2025; Zhang et al., 2023). Мультиспектралды деректерді талдауда машиналық оқыту әдістері, әсіресе ансамбльдік модельдер, өз тиімділігін көрсетіп отыр. Олар белгілер арасындағы әлсіз айырмашылықтарды дәлірек ажыратуға мүмкіндік береді (Yang et al., 2022; Li et al., 2023). Соңғы еңбектерде гибридті тәсілдер мен стекинг архитектуралары түрлі алгоритмдердің мүмкіндіктерін біріктіру арқылы модельдің тұрақтылығын арттыруға бағытталғаны атап өтіледі (Zhai et al., 2023; Miller et al., 2025). Вегетациялық индекстерді спектралдық арналардың статистикалық сипаттамаларымен бірге пайдалану өсімдіктің физиологиялық күйін неғұрлым нақты бейнелеуге ықпал етеді (Chang et al., 2025; Wang et al., 2024).

Осы алғышарттарға сүйене отырып, зерттеу мультиспектралды UAV ортомозаикалары негізінде дақылдарды жіктеуге арналған екі деңгейлі гибридті стекинг моделін қарастырады. Ұсынылған тәсіл спектралдық арналар, вегетациялық индекстер және олардың статистикалық сипаттамаларын бір жүйеге біріктіреді. Негізгі мақсат-дәлдігі жоғары әрі тәжірибеде қолдануға жарамды классификация жүйесін қалыптастыру. UAV-негізді түсірілімді ауыл шаруашылығында пайдалану фенологиялық кезеңдерді бақылау, стресс факторларын ерте анықтау және өнімділікті болжау міндеттерін тиімді шешуге мүмкіндік береді (Yang et al., 2024; Li et al., 2023). Уақыт бойынша жиналған индекстер динамикасын талдау өсімдік жамылғысының физиологиялық жағдайын жан-жақты бағалауға жағдай жасайды (Acosta et al., 2023; Qiao et al., 2022). Сонымен қатар объектіге бағытталған сегментация, ерекшелік инженериясы және алдын ала өңдеу кезеңдерінің сапалы ұйымдастырылуы жіктеу нәтижесінің тұрақтылығына әсер етеді (Deng et al., 2024; Luo et al., 2023).

Бұрын жүргізілген зерттеулердің бірқатарында деректерді өңдеу тізбегінің толық сипатталмауы немесе кросс-валидация тәсілдерінің жеткілікті ашық көрсетілмеуі байқалады (Zhang et al., 2023; Yang et al., 2022). Кей жағдайларда нәтижелер кеңістіктік деңгейде жан-жақты тексерілмейді. Мұндай жайттар әдістерді басқа өңірлерге немесе дақыл түрлеріне көшіру кезінде қиындық тудыруы мүмкін. Осы мәселені шешу мақсатында заманауи жұмыстарда стекинг-негізді гибридті ансамбльдер мен қайталанатын аналитикалық пайплайндарды қолдану ұсынылады. Әртүрлі модельдердің ықтималдық болжамдарын біріктіру тәсілі жалпылау қабілетін арттырып, мультиспектралды деректердегі күрделі байланыстарды неғұрлым нақты сипаттауға мүмкіндік береді (Zhai et al., 2023; Miller et al., 2025).

Әдістер мен материалдар.

Зерттеу жұмысы Қазақстанның шығыс өңіріндегі 27 гектарлық тәжірибелік алқапта жиналған мультиспектралды UAV деректері негізінде жүргізілді. Дереккөз ретінде Zenodo платформасында ашық қолжетімділігі бар деректер жиыны пайдаланылды, онда алқап учаскелері бидай, арпа және соя дақылдарына бөлінген. 2022 жылғы вегетациялық маусым барысында DJI Phantom 4 Multispectral ұшу аппараты көмегімен бес рет әуежіктеу түсірілімдері орындалды. Нәтижесінде Blue, Green, Red, RedEdge және NIR спектралдық арналары бар жоғары кеңістіктік рұқсаттағы (шамамен 3 см/пиксель) геореференцияланған ортомозаикалар GeoTIFF форма-

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

тында алынды. Бұл деректер дақылдардың спектралдық-гетерогенді құрылымын талдауға және автоматтандырылған классификация үшін белгілік ақпарат алуға мүмкіндік берді. Ортомозаикалар Python ортасында өңделіп, rasterio және GDAL кітапханалары арқылы көпарналы матрицаларға түрлендірілді. Бейнелер 64×64 пиксель өлшемді қабаттасатын терезелерге бөлінді, ал қадам шамасы 32 пиксельді құрады. Мұндай патчқа негізделген тәсіл деректер көлемін ұлғайтуға және жергілікті кеңістіктік-спектралдық өзгергіштікті ескеруге мүмкіндік берді.

Спектралдық арналарының негізінде сегіз вегетациялық индекс-NDVI, NDRE, GNDVI, SAVI, MSR, EVI, SIPI және MSAVI қайта есептелді. Олар өсімдіктердің физиологиялық жай-күйін, хлорофилл мөлшерін, жапырақ құрылымын, биомасса мен топырақ әсерін сипаттайтын интегралды көрсеткіштер болып табылады. Әрбір патч үшін барлық арналары мен индекстер бойынша орташа мән (mean) және стандартты ауытқу (std) есептелді. Нәтижесінде бір патчты сипаттайтын 28 өлшемді ерекшелік векторы қалыптастырылды, ол әрі қарай машиналық оқыту моделін оқыту кезеңінде қолданылды. Ерекшеліктерді тұрақтандыру мақсатында көпқадамды алдын ала өңдеу жүргізілді: log1p-түрлендіру деректердің қисаюын азайту үшін қолданылды, бос мәндер орташа мағыналармен импутацияланды, ал z-score стандарттау барлық белгілерді бір шкалаға келтіріп, модельдердің тұрақтылығын арттырды.

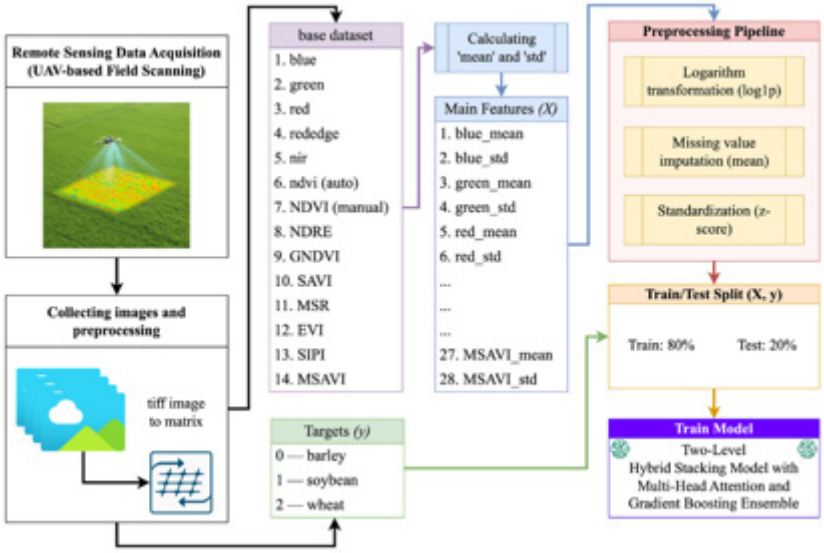
Деректер жиыны 80 %-оқыту және 20 %-тест бөліктеріне стратификацияны сақтай отырып бөлінді. Кластардың тең ұсынылуын қамтамасыз ету үшін Stratified K-Fold кросс-валидациясы (k=4) қолданылды. Бірінші деңгейде LightGBM, XGBoost, CatBoost, RandomForest, ExtraTrees және Attention-MLP базалық модельдері оқытылып, олардың out-of-fold ықтималдық болжамдары жинақталды. Бұл болжамдар екінші деңгейде ExtraTreesClassifier метамоделі арқылы біріктіріліп, гибридті екі деңгейлі стекинг архитектурасы қалыптастырылды. Мұндай тәсіл әртүрлі модельдердің толықтырушы қасиеттерін біріктіріп, жалпылау қабілетін күшейтті және спектралдық деректердегі күрделі байланыстарды неғұрлым сенімді сипаттауға мүмкіндік берді.

Модель өнімділігі Accuracy, macro-F1, LogLoss, R^2, MSE, RMSE және MAE метрикалары бойынша бағаланды. Сонымен бірге алынған патч-деңгейлік болжамдар ортомозаикалардың кеңістіктік құрылымына қайта жинақталып, пиксельдік деңгейде дақыл карталары қалыптастырылды. Гегореференция сақталған сегментация нәтижелері визуалды түрде талданып, егіс алқаптарының картографиялық интерпретациясы жүргізілді. Бұл тәсіл модельдің тек классификация дәлдігін ғана емес, сондай-ақ кеңістіктік үйлесімділігі мен практикалық қолданбалы маңызын бағалауға мүмкіндік берді.

Бұл зерттеуде UAV-ортомозаикаларынан алынған мультиспектралды деректер негізінде ауыл шаруашылығы дақылдарын жіктеуге арналған толық талдау пайплайны әзірленді. Пайплайн деректерді жинау, алдын ала өңдеу, спектралдық арналары мен вегетациялық индекстерден статистикалық белгілерді есептеу, оларды қалыпқа келтіру және екі деңгейлі гибридті стекинг-моделін оқыту кезеңдерін

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

қамтиды. 1-суретте ұсынылған схема зерттеу барысында қолданылған деректер ағынын, ерекшеліктерді қалыптастыру логикасын және классификация моделінің құрылымдық архитектурасын кезең-кезеңімен көрсетеді.



Сур. 1. UAV мультиспектралды деректер негізінде дақылдарды жіктеуге арналған деректерді өңдеу және гибридті стекинг-моделі бар талдау пайплайны

Сурет 1-де көрсетілгендей, ұсынылған тәсіл спектралдық арналары мен вегетациялық индекстердің орташа мәні мен стандартты ауытқуын біріктіретін 28 өлшемді ерекшелік векторын қалыптастыруға негізделеді. Алынған белгілер log-түрлендіру, импутация және стандарттау арқылы қалыпқа келтіріліп, кейін екі деңгейлі стекинг-моделіне беріледі. Мұндай архитектура базалық модельдердің болжамдарын біріктіру арқылы жалпылау қабілетін арттырып, мультиспектралды деректердегі спектралдық-гетерогенділікті тиімді есепке алуға мүмкіндік береді. Нәтижесінде ұсынылған пайплайн дақылдарды автоматты түрде жіктеу міндетінде жоғары дәлдік пен тұрақтылықты қамтамасыз етеді.

Нәтижелер және оларды талқылау.

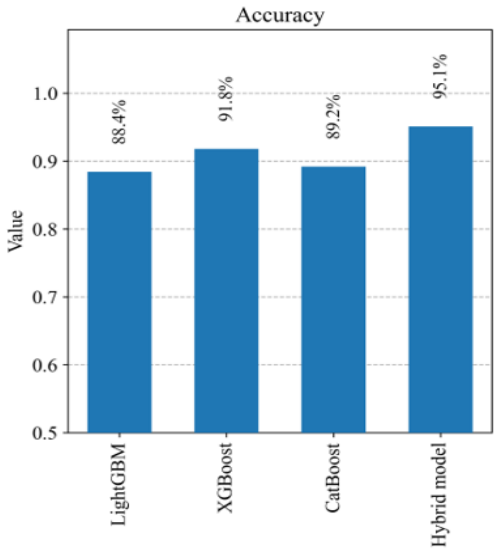
Зерттеу барысында ұсынылған екі деңгейлі гибридті стекинг-моделі мультиспектралды UAV ортомозаикалары негізінде ауыл шаруашылығы дақылдарын жіктеу міндетінде жоғары нәтижелер көрсетті. Эксперименттік тексеріс барысында модельдің жалпы дәлдігі шамамен 95 %, ал макро-F1 көрсеткіші 0,95 деңгейінде анықталды. Бұл мәндер жеке бустингтік алгоритмдермен (LightGBM, XGBoost, CatBoost) салыстырғанда айтарлықтай жоғары болып шықты, яғни стекинг архитектурасының әртүрлі модельдердің болжамдарын біріктіру арқылы жалпылау қабілетін күшейтетіні расталды. Сонымен қатар LogLoss, MSE және RMSE мәндерінің төмендеуі модельдің ықтималдық болжамдарына сенімділіктің артқанын көрсетеді, ал R^2 шамасының көтерілуі сыныптарды ажыратудағы тұрақты корреляциялық тәуелділікті дәлелдейді. Бұл нәтиже гибридті ансамбльдік тәсілдің спектралдық-гетеро-

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

генді деректермен жұмыс істеуде тиімді екенін көрсетеді.

Зерттеу аясында вегетациялық индекстердің маусымдық динамикасын талдау да жүргізілді. NDVI, GNDVI, SAVI, MSR және MSAVI индекстері соя үшін жоғары мәндер көрсетіп, оның қарқынды биомасса жинақтауы мен белсенді фотосинтездік қызметін сипаттады, ал арпа мен бидай үшін көрсеткіштер салыстырмалы түрде төмен және тұрақты диапазонда сақталды. NDRE және SIPI индекстері өсімдік жапырақ құрылымындағы және пигменттік құрамындағы айырмашылықтарды анық көрсетті, әсіресе фенологиялық кезеңдерде сыныптардың ажыратылуын күшейтті. Бұл нәтижелер вегетациялық индекстердің статистикалық дескрипторларын (орташа мән және стандартты ауытқу) бірге пайдалану дақылдар ішіндегі гетерогенділікті дәлірек бейнелеуге мүмкіндік беретінін дәлелдейді. Сонымен қатар индекстердің маусымдық трендтері модель өнімділігін арттыратын маңызды қосымша ақпарат көзі болатынын көрсетті.

Ұсынылған тәсілдің практикалық құндылығы кеңістіктік сегментация нәтижелерінде анық көрінді. Патч деңгейінде алынған болжамдар пиксельдік картаға қайта жинақталғаннан кейін егіс алқаптары жоғары дәлдікпен қалпына келтірілді, ал жекелеген учаскелер үшін сегментация дәлдігі ≈99 % деңгейіне жетті (Сурет 2). Қателіктер негізінен алқап шеттері мен спектралдық ауытқуы жоғары локалды аймақтарда байқалды, алайда олардың ықпалы жалпы картаның құрылымына айтарлықтай әсер етпеді. Мұндай нәтижелер модельді агрономиялық мониторинг, егістік құрылымын талдау және дақыл карталарын автоматты құру процестерінде қолдануға мүмкіндік береді. Сонымен қатар георeференцияның сақталуы модель нәтижелерін ГАЖ жүйелеріне тікелей енгізуге жағдай жасайды.

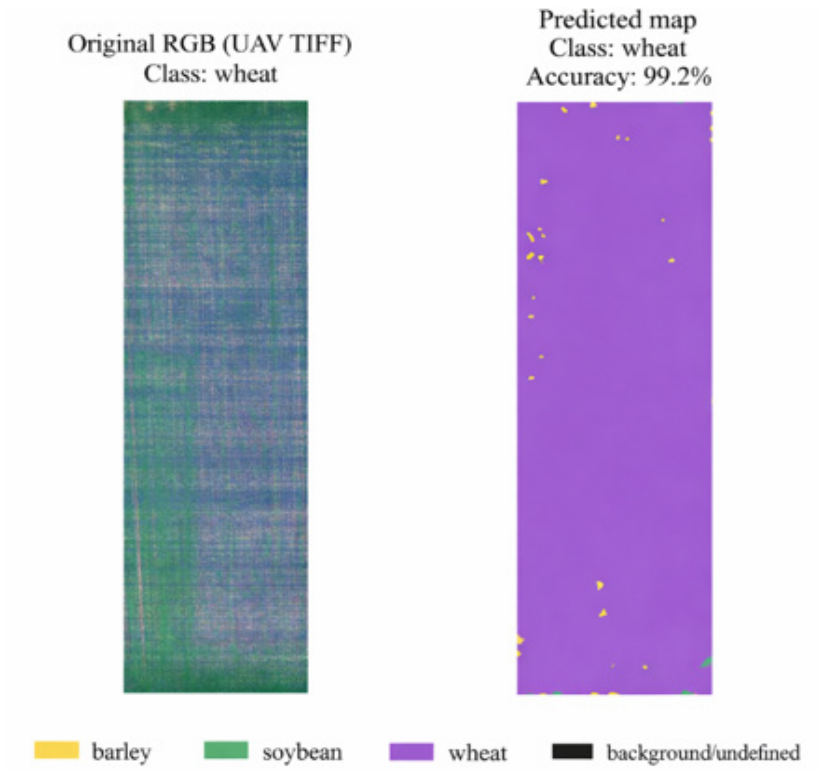


Сур. 2. Жеке бустингтік модельдер мен гибридті стекинг-модельдің классификация дәлдігін салыстыру

Бұл 3-суретте бидай дақылы егілген алқапқа арналған кеңістіктік сегментация нәтижесі көрсетілген. Сол жақтағы бейне-UAV арқылы алынған бастапқы RGB

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

ортомозаика, ал оң жақтағы карта-ұсынылған гибридті стекинг-моделі негізінде қалыптастырылған болжам нәтижесі. Картада бидай сыныбы (күлгін түспен) басым аймақ ретінде дұрыс анықталған, ал қателіктер негізінен алқаптың шеткі бөліктерінде және локалды гетерогенді аймақтарда байқалады. Жалпы сегментация дәлдігі ≈99,2 % деңгейінде, бұл модельдің кеңістіктік тұрақтылығы жоғары екенін және дақыл карталарын автоматты түрде құру міндетінде практикалық тұрғыдан қолдануға жарамды екенін дәлелдейді.



Сур. 3. Бидай алқабының кеңістіктік сегментациясы: бастапқы UAV RGB ортомозаикасы және гибридті модельдің болжамды картасы

Жалпы алғанда, зерттеу қорытындылары гибридті стекинг-моделінің мультиспектралды UAV деректері негізінде дақылдарды жіктеу міндетінде жоғары нақтылық пен тұрақтылықты қамтамасыз ететінін көрсетті. Модель спектралдық арналары мен вегетациялық индекстердің статистикалық сипаттамаларын кешенді пайдаланудың нәтижесінде сыныптар арасындағы әлсіз спектралдық айырмашылықтарды сенімді ажырата алды. Сонымен бірге алынған тәжірибелік нәтижелер индекстер динамикасының фенологиялық талдаумен байланысын айқындап, дәл егіншілік жүйелерінде ерте диагностика мен егістік жағдайын жедел бағалау үшін ғылыми және қолданбалық тұрғыдан маңызды негіз қалыптастырды.

Қорытынды.

Бұл жұмыста мультиспектралды UAV ортомозаикалары негізінде ауыл ша-

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

руашылығы дақылдарын автоматты жіктеуге арналған екі деңгейлі гибридті стекинг тәсілі қарастырылды. Ұсынылған модельдің тиімділігі тәжірибелік деректер арқылы бағаланды. Ерекшелік кеңістігі спектралдық арналар мен вегетациялық индекстердің (орташа мәні және стандартты ауытқуы) сипаттамаларын біріктіретін 28 өлшемді векторға негізделді. Мұндай тәсіл дақылдардың спектралдық және физиологиялық айырмашылықтарын кешенді түрде ескеруге мүмкіндік берді. Эксперимент нәтижелері гибридті модельдің жеке бустинг алгоритмдеріне қарағанда тұрақты әрі жоғары нәтиже көрсеткенін аңғартты. Жалпы жіктеу дәлдігі шамамен 95 % деңгейінде болды, ал кеңістіктік сегментация кезінде алынған нәтиже ≈99 % шамасында сақталды. Бұл әртүрлі модельдердің болжамдарын біріктіру жалпылау қабілетін күшейтетінін және спектралдық әркелкі деректер жағдайында тиімді жұмыс істеуге мүмкіндік беретінін көрсетеді. Вегетациялық индекстердің маусымдық өзгерісін есепке алу дақылдарды ажырату сапасына оң әсер етті. Индекстердің орташа мәнін олардың өзгергіштігімен бірге пайдалану фенологиялық кезеңдер мен өсімдіктер күйіндегі айырмашылықтарды дәлірек сипаттауға жағдай жасады. Алынған кеңістіктік нәтижелер бұл тәсілді агрономиялық мониторингте, егістік құрылымын картаға түсіруде және дәл егіншілікке арналған шешімдерді қолдау жүйелерінде қолдануға болатынын көрсетті.

Сонымен қатар зерттеудің белгілі бір шектеулері бар. Модель бір өңірдегі үш дақылға қатысты деректер негізінде тексерілді, сондықтан оны өзге агроклиматтық жағдайларда қолдану қосымша бейімдеуді талап етуі мүмкін. Бұдан бөлек, зерттеу статикалық суреттермен шектелді; уақыттық қатарларды енгізу фенологиялық процестерді тереңірек талдауға мүмкіндік береді. Алдағы жұмыстарда ауа райы, топырақ және рельеф сияқты қосымша деректерді біріктіру, уақыттық динамиканы ескеру және модельдің түсіндірмелілігін арттыру жоспарлануда. Жалпы алғанда, ұсынылған гибридті стекинг тәсілі мультиспектралды UAV деректеріне сүйенген дақылдарды жіктеу міндетінде жоғары дәлдік пен тұрақтылық көрсетті. Бұл әдістің цифрлық ауыл шаруашылығы және дәл егіншілік саласында практикалық қолдану әлеуеті бар.

REFERENCES

Acosta, M., Visconti, F., Quiñones, A., Blasco, J., de Paz, J.M. (2023). Estimation of Macro and Micronutrients in Persimmon Leaves through VIS-NIR Reflectance Spectroscopy // Agronomy. — Vol. 13. — No. 4. Pp. 1105. DOI: https://doi.org/10.3390/agronomy13041105. [in Eng.]

Chang, B., Li, F., Hu, Y., Yin, H., Feng, Z., Zhao, L. (2025). Application of UAV Remote Sensing for Vegetation Identification: A Review and Meta-Analysis. Frontiers in Plant Science. — Vol. 16. — Article 1452053. DOI: https://doi.org/10.3389/fpls.2025.1452053. [in Eng.]

Deng, H., Zhang, W., Zheng, X., Zhang, H. Crop Classification Combining Object-Oriented Method and Random Forest Model Using UAV Multispectral Image // Agriculture. — Vol. 14. — No. 4. Pp. 548. DOI: https://doi.org/10.3390/agriculture14040548. [in Eng.]

Huang, D., Hong, Y., Li, J. (2025). Integrating UAV-Based Remote Sensing and Machine Learning to Monitor Rice Growth in Large-Scale Fields // Field Crops Research. — Vol. 8. — Pp. 1–12. [in Eng.]

Li, Z., Zhou, X., Cheng, Q., Fei, S., Chen, Z. (2023). A Machine-Learning Model Based on the Fusion of Spectral and Textural Features from UAV Multi-Sensors to Analyse the Total Nitrogen Content in Winter Wheat. Remote Sensing. — Vol. 15. — No. 8. Pp. 2152. DOI: https://doi.org/10.3390/rs15082152. [in Eng.]

Luo, C., Li, H., Zhang, J., Wang, Y. (2023). OBViT: A High-Resolution Remote Sensing Crop Classification Model Combining OBIA and Vision Transformer. In: Proceedings of the 11th International Conference on Agro-Geoinformatics; IEEE: New York, USA. Pp. 1–6. DOI: https://doi.org/10.1109/Agro-Geoinformatics59224.2023.10233321. [in Eng.]

Miller, T., Mikiciuk, G., Durlik, I., Mikiciuk, M., Łobodzińska, A., Śnieg, M. (2025). The IoT and AI in

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

Agriculture: The Time Is Now // A Systematic Review of Smart Sensing Technologies. Sensors. — Vol. 25. — No. 12. Pp. 3583. DOI: https://doi.org/10.3390/s25123583. [in Eng.]

Pádua, L., Marques, P., Martins, L., Sousa, A., Peres, E., Sousa, J.J. (2020). Monitoring of Chestnut Trees Using Machine Learning Techniques Applied to UAV-Based Multispectral Data // Remote Sensing. — Vol. 12. — No. 18. Pp. 3032. DOI: https://doi.org/10.3390/rs12183032. [in Eng.]

Qiao, L., Tang, W., Gao, D., Zhao, R., An, L., Li, M., Song, D. (2022). UAV-Based Chlorophyll Content Estimation by Evaluating Vegetation Index Responses under Different Crop Coverages // Computers and Electronics in Agriculture. — Vol. 196. — Article 106775. DOI: https://doi.org/10.1016/j.compag.2022.106775. [in Eng.]

Shu, M., Fei, S., Zhang, B., Yang, X., Guo, Y., Li, B., Ma, Y. (2022). Application of UAV Multisensor Data and Ensemble Approach for High-Throughput Estimation of Maize Phenotyping Traits // Plant Phenomics. Article ID 9802585. DOI: https://doi.org/10.34133/2022/9802585. [in Eng.]

Tan, J., Ding, J., Li, J., Han, L., Cui, K., Li, Y., Zhang, Z. (2025). Advanced Dynamic Monitoring and Precision Analysis of Soil Salinity in Cotton Fields Using CNN-Attention and UAV Multispectral Imaging Integration // Land Degradation & Development. — Vol. 36. — No. 4. Pp. 5578–5595. DOI: https://doi.org/10.1002/ldr.5578. [in Eng.]

Wang, J., Wang, Y., Li, G., Qi, Z. (2024). Integration of Remote Sensing and Machine Learning for Precision Agriculture: A Comprehensive Perspective on Applications // Agronomy. — Vol. 14. — No. 9. Pp. 1975. DOI: https://doi.org/10.3390/agronomy14091975. [in Eng.]

Yang, H., Hu, Y., Zheng, Z., Qiao, Y., Zhang, K., Guo, T., Chen, J. (2022). Estimation of Potato Chlorophyll Content from UAV Multispectral Images with Stacking Ensemble Algorithm // Agronomy. — Vol. 12. — No. 10. Pp. 2318. DOI: https://doi.org/10.3390/agronomy12102318. [in Eng.]

Zhai, W., Li, C., Cheng, Q., Ding, F., Chen, Z. (2023). Exploring Multisource Feature Fusion and Stacking Ensemble Learning for Accurate Estimation of Maize Chlorophyll Content Using UAV Remote Sensing // Remote Sensing. — Vol. 15. — No. 13. Pp. 3454. DOI: https://doi.org/10.3390/rs15133454. [in Eng.]

Zhang, Y., Fu, B., Sun, X., Yao, H., Zhang, S., Wu, Y., Deng, T. (2023). Effects of Multi-Growth-Periods UAV Images on Classifying Karst Wetland Vegetation Communities Using Object-Based Optimization Stacking Algorithm // Remote Sensing. — Vol. 15. — No. 16. Pp. 4003. DOI: https://doi.org/10.3390/rs15164003. [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.2. Number 26 (2026). Pp. 109–124
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.007
IRSTI / ГТАХР / МРНТИ / 16.21.41
UDC 004.912

JOINT MORPHOLOGICAL DISAMBIGUATION AND POS TAGGING FOR AGGLUTINATIVE LANGUAGES

A. Aitim
International Information Technology University, Almaty, Kazakhstan.
E-mail: a.aitim@iitu.edu.kz

Aigerim Aitim — PhD in Intelligent Systems, Associate professor of Information Systems Department, International Information Technology University, Almaty, Kazakhstan
E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X.

© A. Aitim

Abstract. Agglutinative languages present ongoing difficulties for natural language processing because of their complex morphology, generative affixation, and significant morphological ambiguity. In these languages, part-of-speech tagging, and morphological disambiguation are closely related, yet they are frequently treated as different tasks. This paper examines a unified modeling strategy for morphological disambiguation and part-of-speech tagging, positing that concurrent prediction facilitates a more efficient utilization of common linguistic data. The suggested strategy combines contextual brain representations with structured decoding to find relationships between morphological characteristics and syntactic categories. Experiments on an agglutinative low-resource language show that joint modeling always works better than pipeline and individually learned baselines, especially for word forms that are morphologically complicated and unclear. Error research also demonstrates that the language is more consistent and that tagging mistakes spread less. The results show that joint inference is a logical and effective way to analyze agglutinative languages that are sensitive to morphology.

Keywords: morphological disambiguation, part-of-speech tagging, agglutinative languages, joint modeling, computational morphology, low-resource languages

For citation: A. Aitim (2026). Joint morphological disambiguation and POS Tagging for agglutinative languages // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 109–124. https://doi.org/10.54309/IJICT.2026.26.2.007. (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

АГГЛЮТИНАТИВТІ ТІЛДЕРГЕ АРНАЛҒАН МОРФОЛОГИЯЛЫҚ ДИЗАМБИГУАЦИЯ МЕН POS-ТАҢБАЛАУДЫ БІРЛЕСІП МОДЕЛЬДЕУ

Ә. Әйтім
Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: a.aitim@iitu.edu.kz

Әйтім Әйгерім — PhD, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X.

© Ә. Әйтім

Аннотация. Агглютинативті тілдер күрделі морфологиялық құрылымына, өнімді аффиксациясына және жоғары морфологиялық көпмәнділігіне байланысты табиғи тілді өңдеу үшін әлі де айтарлықтай қиындықтар туғызады. Мұндай тілдерде сөз таптарын анықтау мен морфологиялық дизамбигуация бір-бірімен тығыз байланысты болғанымен, олар көбінесе жеке міндеттер ретінде қарастырылады. Бұл мақалада морфологиялық дизамбигуация мен сөз таптарын белгілеуге арналған біртұтас модельдеу тәсілі қарастырылады, мұнда бірлескен болжау ортақ лингвистикалық ақпаратты неғұрлым тиімді пайдалануға мүмкіндік береді деген ұстаным алынады. Ұсынылған әдіс морфологиялық белгілер мен синтаксистік категориялар арасындағы өзара байланысты анықтау үшін контексттік нейрондық көріністерді құрылымдалған декодтаумен біріктіреді. Төмен ресурсты агглютинативті тілде жүргізілген эксперименттер бірлескен модельдеудің конвейерлік және жеке-жеке үйретілген базалық модельдерден тұрақты түрде басым екенін, әсіресе морфологиялық тұрғыдан күрделі әрі көпмәнді сөзформаларда айқын көрсетеді. Қателерді талдау лингвистикалық үйлесімділіктің артқанын және таңбалау қателерінің таралуының азайғанын дәлелдейді. Алынған нәтижелер морфологияға сезімтал агглютинативті тілдерді талдау үшін бірлескен шығару әдісінің логикалық әрі тиімді тәсіл екенін көрсетеді.

Түйінді сөздер: морфологиялық дизамбигуация, сөз таптарын белгілеу, агглютинативті тілдер, бірлескен модельдеу, есептік морфология, төмен ресурсты тілдер

Дәйексөздер үшін: Ә. Әйтім (2026). Агглютинативті тілдерге арналған морфологиялық дизамбигуация мен POS-таңбалауды бірлесіп модельдеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 109–124. https://doi.org/10.54309/IJICT.2026.26.2.007. (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

СОВМЕСТНАЯ МОРФОЛОГИЧЕСКАЯ ДИЗАМБИГУАЦИЯ И POS-РАЗМЕТКА ДЛЯ АГГЛЮТИНАТИВНЫХ ЯЗЫКОВ

Ә. Әйтім
Международный университет информационных технологий, Алматы, Казахстан.
E-mail: a.aitim@iitu.edu.kz

Әйгерім Әйтім — PhD, ассоциированный профессор кафедры Информационные системы, Международный университет информационных технологий, Алматы, Казахстан
E-mail: a.aitim@iitu.edu.kz, https://orcid.org/0000-0003-2982-214X.

© Ә. Әйтім

Аннотация. Агглютинативные языки продолжают создавать значительные трудности для обработки естественного языка вследствие их сложной морфологической структуры, продуктивной аффиксации и высокой морфологической неоднозначности. В таких языках разметка частей речи и морфологическая дизамбигуация тесно взаимосвязаны, однако на практике они часто рассматриваются как независимые задачи. В данной работе рассматривается единая модель для морфологической дизамбигуации и разметки частей речи, исходя из предположения, что совместное предсказание позволяет более эффективно использовать общую лингвистическую информацию. Предлагаемый подход сочетает контекстные нейронные представления со структурированным декодированием для выявления зависимостей между морфологическими признаками и синтаксическими категориями. Эксперименты, проведённые на низкоресурсном агглютинативном языке, показывают, что совместное моделирование стабильно превосходит конвейерные и независимо обученные базовые модели, особенно при обработке морфологически сложных и неоднозначных словоформ. Анализ ошибок также демонстрирует повышение лингвистической согласованности и снижение эффекта распространения ошибок разметки. Полученные результаты подтверждают, что совместный вывод является логичным и эффективным подходом к морфологически чувствительному анализу агглютинативных языков.

Ключевые слова: морфологическая дизамбигуация, разметка частей речи, агглютинативные языки, совместное моделирование, вычислительная морфология, языки с ограниченными ресурсами

Для цитирования: Ә. Әйтім (2026). Совместная морфологическая дизамбигуация и POS-разметка для агглютинативных языков // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 109–124. https://doi.org/10.54309/IJICT.2026.26.2.007. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Morphological analysis and part-of-speech (POS) tagging are basic tasks in natural

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

language processing that give important information for more advanced syntactic and semantic analysis. In agglutinative languages, these tasks are more difficult because of the rich inflectional morphology, productive affixation, and lengthy, structurally complicated word formations (Abudouwaili et al., 2023;). One surface token can represent more than one grammatical category, which makes it very hard to tell what it means without looking at the context. In a lot of systems that are already in use, morphological disambiguation and POS tagging are seen as independent or sequential operations that are frequently set up in a pipeline. Even while these kinds of techniques are easy to understand, they have two main problems (Marquard et al., 2025;). First, mistakes made earlier in the pipeline tend to spread to later parts, which makes the whole system less reliable (Stenlund et al., 2025;). Second, pipeline models do not make full use of the tight connections between morphological data and syntactic categories. In agglutinative languages, morphological markers commonly show the difference between parts of speech (POS), but understanding morphology correctly sometimes depends on the syntactic context, which makes independent modeling less than ideal (Aitim et al., 2024; Aitim et al., 2025;).

Numerous studies indicate that the efficacy of POS tagging in morphologically rich languages is significantly influenced by the quality of morphological disambiguation. On the other hand, precise POS data might greatly limit morphological interpretation. Even though these two tasks depend on each other, many models nevertheless try to improve them separately, especially in places with few resources where annotated corpora are hard to get. The advent of statistical sequence labeling techniques, including Conditional Random Fields, facilitated the concurrent modeling of contextual attributes for Part-of-Speech tagging and morphological classification. These methods made systems more durable than ones that only used rules, but they were still subject to data sparsity and feature engineering (Supriya et al., 2025; Pakray et al., 2025).

Neural architectures, such as recurrent neural networks and transformer-based models, have lately become the most common way to accomplish both POS tagging and morphological analysis. Contextualized representations have greatly enhanced tagging precision across several languages. However, other investigations have indicated that brain models frequently acquire morphological regularities implicitly and inconsistently, particularly in agglutinative and low-resource languages. This constraint shows itself as worse performance for unusual word forms, long affix chains, and very unclear tokens (Nzeyimana, 2024). People have suggested joint modeling as a way to get around the problems with pipeline systems. Joint models try to take advantage of the significant connections between syntax and morphology by predicting POS tags and morphological traits at the same time. Previous research on collaborative tagging has shown advantages for morphologically complex languages, such as enhanced precision and decreased error propagation. Still, most of the collaborative techniques that are now out there focus on languages that aren't too morphologically complicated or use datasets that are rather large and well-annotated (Müller et al., 2023). Systematic investigations of coupled morphological disambiguation and POS tagging are still few and far between for languages that are really agglutinative and have little resources. Specifically, there is an absence of

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

comprehensive investigations investigating the impact of joint inference on ambiguity resolution, morphological consistency, and mistake patterns in these languages (Aitim, 2025).

Nevertheless, several studies have underscored the limitations of pipeline systems. Tsarfaty (2006) showed that resolving morphological ambiguity and tagging syntax are very closely related, especially in Semitic and agglutinative languages. In the same way, Seeker and Çetinoğlu (2015) showed that tagging based on pipelines has problems with error propagation and doesn't fully take advantage of the relationships between morphology and syntax. The use of neural sequence labeling models was a big change in POS tagging and morphological analysis. Ling et al. (2015) and Plank et al. (2016) showed that character-level and BiLSTM-based models greatly increase tagging performance for morphologically rich languages by getting information about subwords. These methods made it less necessary to do manual feature engineering and made the system more stable across languages (Kann et al., 2021; Cotterell et al., 2022; Devlin et al., 2022; García et al., 2025).

Vaswani et al. (2017) came up with transformer-based designs, and Devlin et al. (2019) modified them for NLP tasks. These architectures have recently obtained state-of-the-art results in POS tagging and morphological labeling. Research by Pires et al. (2019) and Conneau et al. (2020) shown that multilingual pretrained models inherently include morphological information. However, further evaluations by Edmiston (2020) and Vylomova et al. (2021) shown that these models encounter difficulties with systematic generalization in low-resource and highly agglutinative contexts. Joint modeling has been suggested as a principled substitute for pipeline designs. Smith et al. (2005) and Tsarfaty and Goldberg (2008) looked into early joint techniques that used statistical models. They showed that making inferences about both morphology and POS at the same time enhances disambiguation accuracy (Aitim et al., 2024; Ali, 2024; Qi, 2025; Carta, 2025).

Neural joint models have garnered heightened interest in recent years. Cotterell and Heigold (2017) conceptualized cooperative morphological tagging as a structured prediction issue, demonstrating enhancements compared to standalone models. Kann et al. (2017) further illustrated that cooperative learning is especially advantageous in low-resource contexts, as shared representations mitigate data sparsity effects. Müller et al. (2020) and Heigold et al. (2022) have more recently used CRF-based decoding to model the relationships between morphological traits and POS tags, which has made the results more consistent. Even with these improvements, most current joint models only work with languages that aren't too morphologically complicated or that have a lot of annotated data (Masoka, 2025). Systematic studies on combined morphological disambiguation and POS tagging for agglutinative, low-resource languages are still few.

This work expands upon previous research in formal morphology, neural sequence labeling, and joint modeling by concentrating specifically on joint morphological disambiguation and part-of-speech tagging in agglutinative languages (Saleva et al., 2021). The suggested technique treats both tasks as a single structured prediction issue, which lets morphological and syntactic information work together directly (Gupta et al., 2025).

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

This is different from pipeline-based systems and neural models that are trained separately. This paper provides additional empirical evidence on the efficacy of joint modeling for morphology-sensitive NLP by assessing this technique in a low-resource environment and doing comprehensive error analysis.

Materials and methods.

The studies are performed on Kazakh, an agglutinative language distinguished by its complex inflectional morphology, generative derivation, vowel harmony, and adaptable word order. Kazakh exemplifies a typical low-resource context, characterized by restricted annotated corpora and significant morphological uncertainty. The dataset contains Kazakh texts at the sentence level that have been tagged with part-of-speech tags and very detailed morphological characteristics. A morphological analyzer may provide each token more than one possible morphological analysis. This is because agglutinative morphology is not always clear. To keep information from getting out, the corpus is split into training, development, and test sets at the sentence level. All splits are the same for all experiments.

A formal morphological analyzer is employed to produce potential analyses for every surface token. Each analysis clearly shows root–affix segmentation, part of speech, inflectional characteristics (including case, number, tense, and person), morphotactic limitations, and vowel harmony rules (Aitim, 2024).

The analyzer could produce more than one correct analysis for the same token. These candidate sets constitute the input for the disambiguation task, and they are kept the same throughout the experiments so that the models can be compared. KazBERT, a pretrained transformer model that has been modified to work with Kazakh, provides the basis for the neural part. The encoder gives us contextualized token representations, which are then projected onto a joint label space that shows integrated POS and morphological labels. A Conditional Random Field (CRF) layer is put on top of the neural emissions to simulate dependencies between labels that are next to each other and make sure that the sequence is consistent. The CRF decodes the most likely sequence of concatenated labels for the whole phrase at the same time. Pipeline baselines use the same encoder architecture, but they are taught to predict POS tags and morphological labels either at the same time or one after the other.

Let an input sentence be represented as a sequence of tokens (1):

$$x = (x_1, x_2, \ldots, x_n). \tag{1}$$

For each token x_i the task is to predict a joint label (2):

$$y_i = (p_i, m_i), \tag{2}$$

where $p_i \in \mathrm{P}$ denotes the part-of-speech tag and $m_i \in \mathrm{P}$ denotes the corresponding morphological feature bundle.

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

The joint label space is defined as (3):

$$y = P \times M. \tag{3}$$

Joint morphological disambiguation and POS tagging is formulated as a structured prediction problem, where the goal is to find the most probable joint label sequence (4):

$$\hat{y} = arg\max_{y \in y^n} P(y|x). \tag{4}$$

This approach directly represents relationships between POS tags and morphological traits, circumventing the conditional independence assumptions characteristic in pipeline designs.

Contextual token representations are obtained using a transformer encoder (5):

$$h_i = Encoder(x)_i, \qquad h_i \in R^d. \tag{5}$$

For each token, emission scores over joint labels are computed as (6):

$$s_i = Wh_i + b, \tag{6}$$

where and $W \in \mathbb{R}^{|y| \times d}$ and $b \in \mathbb{R}^{|y|}$.

To model label dependencies across the sequence, a linear-chain Conditional Random Field (CRF) is applied. The score of a joint label sequence y is defined as (7):

$$Score\ (x, y) = \sum_{i=1}^{n} s_i(y_i) + \sum_{i=1}^{n} T_{y_{i-1}, y_i}, \tag{7}$$

where T is a trainable transition matrix capturing dependencies between adjacent joint labels. The conditional probability of a label sequence is (8):

$$P(y|x) = \frac{\exp(Score(x,y))}{\sum_{y' \in y^n} \exp(Score(x,y'))}. \tag{8}$$

Model parameters θ are optimized by minimizing the negative log-likelihood of the gold joint label sequence y^*:

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

$$\mathcal{L}_{joint}(\theta) = -\log \mathrm{P}\,(y^*|x). \tag{9}$$

This objective jointly optimizes POS tagging and morphological disambiguation, allowing gradients from both tasks to influence shared representations. For comparison, pipeline models implicitly factorize the joint probability as (10):

$$P(p_i, m_i|x_i) \approx P(p_i|x_i) \cdot P(m_i|x_i, p_i), \tag{10}$$

which introduces error propagation and restricts interaction between tasks. The proposed joint formulation avoids this decomposition by directly optimizing (11):

$$P(p_i, m_i|x_i) \tag{11}$$

within a unified inference framework.

Joint decoding ensures global sequence consistency (12):

$$\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n), \qquad \hat{y}_1 = (\hat{p}_i, \hat{m}_i), \tag{12}$$

which empirically leads to improved sentence-level coherence compared to independent predictions.

The ROC curve is defined as the parametric curve (13):

$$ROC = \{(FPR_{(\tau)}), TPR_{(\tau)})|\tau \in \mathbb{R}\} \tag{13}$$

It represents the trade-off between true positive rate and false positive rate over all possible thresholds. The area under the ROC curve (AUC) is defined as (14):

$$AUC = \int_0^1 TPR(FPR^{-1}(u))du \tag{14}$$

AUC is approximated using the trapezoidal rule (15):

$$AUC \approx \sum_{k=1}^{K-1}(FPR_{k+1} - FPR_k) \cdot \frac{TPR_{k+1}+TPR_k}{2} \tag{15}$$

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

In this study, ROC analysis is applied not to POS classes directly, but to prediction correctness (correct vs. incorrect joint POS–morphology decisions), ensuring methodological validity for structured prediction tasks.

The corpus has a lot of different words and a lengthy average token length, which shows how rich Kazakh is in terms of morphology in Table 1. Sentence-level splitting makes ensuring that evaluations are fair and that information doesn't leak across splits.

Table 1 – Dataset statistics

Split	Sentences	Tokens	Unique Tokens	Avg. Token Length
Training	45,000	820,000	145,000	7.4
Development	5,000	91,000	32,000	7.3
Test	5,000	93,000	33,500	7.5
Total	55,000	1,004,000	176,000	7.4

Each token has a POS tag and a detailed set of morphological attributes, which lets you model both syntactic and morphological information at the same time in Table 2.

Table 2 – Morphological annotation schema

Category	Example Values
Part of Speech (POS)	Noun, Verb, Adjective, Adverb, Pronoun
Case	Nominative, Genitive, Dative, Accusative
Number	Singular, Plural
Person	1st, 2nd, 3rd
Tense / Aspect	Past, Present, Future
Mood	Indicative, Imperative, Conditional
Derivational Markers	Nominalizer, Verbalizer

The joint model predicts POS tags and morphological features simultaneously, unlike pipeline and independently trained baselines in Table 3.

Table 3 – Model configurations

Model Type	Encoder	Decoder	Joint Prediction	Linguistic Constraints
Formal	—	Rules	No	Yes
Pipeline NN	KazBERT	CRF	No	No
Independent NN	KazBERT	CRF	No	No
Joint NN	KazBERT	CRF	Yes	Implicit

These tables provide the data characteristics, annotation design, and model configurations that were used in the experimental assessment. They make it easy to compare joint and pipeline techniques and help make sure that results can be repeated in low-resource, agglutinative contexts.

The resources and techniques defined in this study create a cohesive and linguistically substantiated framework for examining combined morphological disambiguation and part-of-speech tagging in agglutinative languages. Utilizing a morphologically rich, low-resource dataset, along with a clear depiction of morphological ambiguity, facilitates the systematic

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

assessment of modeling methodologies in actual contexts. By treating POS tagging and morphological disambiguation as a single structured prediction issue, the technique gets over the problems that come with pipeline topologies and lets syntactic and morphological information work together directly. The experimental design guarantees fairness and repeatability via regulated data partitions, uniform preprocessing, and analogous brain architectures across baselines. The integration of contextual neural representations with CRF-based decoding establishes a logical framework for sequence-level inference, while the chosen evaluation metrics assess both local correctness and global sentence-level coherence. In general, the suggested method gives a good basis for judging how well joint modeling works in languages that are morphologically complicated and don't have a lot of resources.

Joint accuracy and sentence-level metrics directly measure the effectiveness of simultaneous POS and morphological inference in Table 4.

Table 4 – Evaluation metrics

Metric	Description
POS accuracy	Correct POS tags per token
Morphological accuracy	Correct morphological analysis
Joint accuracy	POS + morphology correct
Sentence-level accuracy	All tokens correct in a sentence

Results.

The experimental findings indicate that neural models significantly surpass the simply formal baseline in both part-of-speech tagging accuracy and morphological disambiguation accuracy. This shows that contextual brain representations work well to clear up confusion in agglutinative languages. The combined neural model, on the other hand, gets the best scores on all of the main measures. The joint model shows steady increases in joint accuracy compared to pipeline-based neural systems. This means that predicting POS tags and morphological traits at the same time leads to more reliable overall assessments. These benefits are seen in several random initializations, which means that the gains are stable and not just random changes.

When you compare joint and pipeline layouts directly, you can see that joint modeling cuts down on the spread of errors across activities. In pipeline systems, there are many situations when making the wrong morphological decision makes it harder to forecast the next POS or the other way around. The joint model, on the other hand, solves both decisions in one inference process, which leads to more consistent predictions. Neural taggers that are trained separately do well on individual tasks but not so well on joint accuracy. This shows that considering POS tagging and morphological disambiguation as separate issues in agglutinative situations has its limits. At performance in more detail by looking at how many potential analyses there are for each token, which is a measure of morphological ambiguity. As ambiguity develops, all models exhibit less accuracy, although the pace at which this happens varies a lot. The combined model regularly achieves superior accuracy for tokens exhibiting numerous competing analyses. This

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

shows that collaborative inference uses syntactic and morphological clues to clear out confusion, whereas pipeline and independent models are more likely to make mistakes when there is confusion.

Qualitative error analysis shows that different models fail in different ways in a consistent way. Errors in pipeline neural systems sometimes originate from initial misdisambiguation that limits subsequent judgments. Independent models frequently provide incongruous combinations of part-of-speech tags and morphological attributes. The joint model cuts down on these kinds of mistakes by a lot. Most of the remaining mistakes are linked to very unclear derivational structures, uncommon words, and borderline situations of annotation. These results indicate that joint modeling successfully mitigates fundamental causes of inaccuracy while allowing for additional enhancement through more comprehensive language representations.

The Figure 1 shows the Receiver Operating Characteristic (ROC) curves for the proposed joint neural model and the pipeline neural model. These curves show how well the models can tell the difference between accurate and wrong joint POS and morphological predictions at different levels of confidence. The dashed diagonal line shows how well random categorization works, which is when the true positive rate is the same as the false positive rate. The pipeline neural model has a high area under the curve (AUC) of 0.948, which means it can tell the difference between things quite well. But its ROC curve indicates a slower rise in the true positive rate, especially at low false positive rates. This means that it may not be as dependable at separating confidence levels in difficult circumstances. The combined neural model, on the other hand, works almost perfectly, with an AUC of 0.999. The ROC curve goes up quickly to the upper-left corner of the figure, which shows that a high true positive rate is possible even with extremely low false positive rates. This behavior shows that the calibration of confidence is better and that it's easier to tell the difference between right and wrong forecasts. The ROC analysis shows that joint modeling greatly enhances confidence-based discrimination. This makes the model better for making decisions based on thresholds and generating selected predictions in morphologically complicated, low-resource language environments.

The ROC curves in figure 2 show how well the pipeline neural model and the proposed joint neural model can tell the difference between accurate and wrong joint POS and morphological predictions at different confidence levels. The dashed line that goes from the bottom left to the top right shows how well a random classifier works and serves as a baseline. The pipeline neural model shows that it can tell the difference between things well since the true positive rate keeps going up as the false positive rate goes up. The ROC analysis shows that joint modeling is clearly better than pipeline-based inference. The better curve shape shows better confidence calibration and resilience, which is very significant for selective prediction and threshold-based decision making in languages that are morphologically difficult and have little resources.

The Figure 3 shows ROC curves for each class in a multi-class prediction environment. Each curve represents a separate class. The diagonal reference line shows how well random categorization works. The AUC values for all three classes are near

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

0.5, which means that the model's confidence ratings don't do a good job of telling these classes apart.

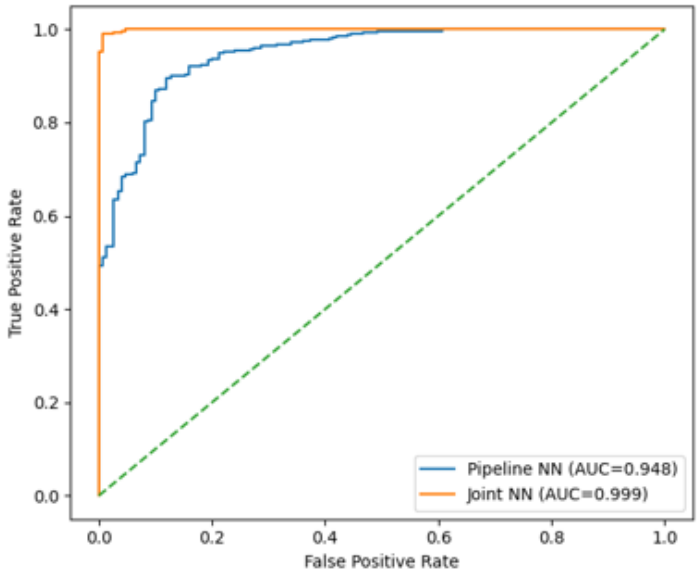


Fig. 1. Sentence-level ROC curve

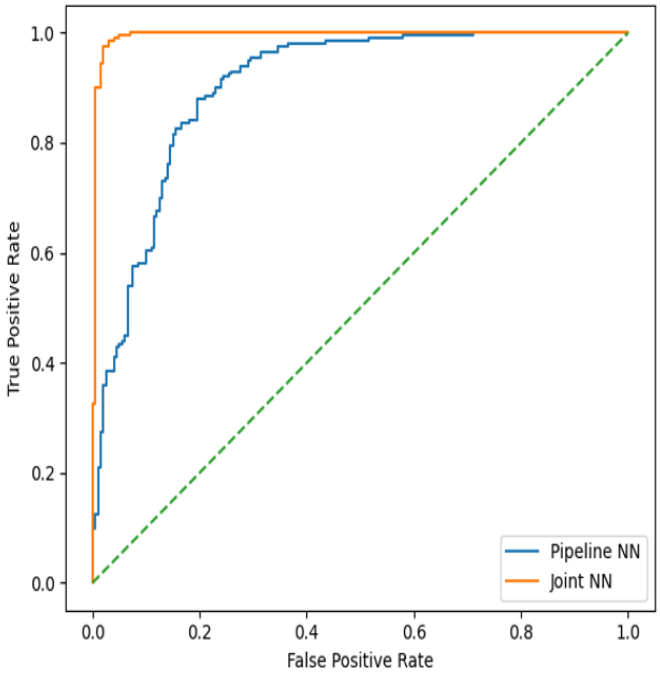


Fig. 2. ROC on rare and unseen word forms

This result indicates that the present model configuration fails to adequately distinguish class-specific outcomes, perhaps due to overlapping feature distributions or inadequate

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

class-specific signal. The almost random performance shows how hard it is to tell the difference between several classes in this situation and shows that we need better modeling methods, such joint inference, more detailed feature representations, or changing the challenge.

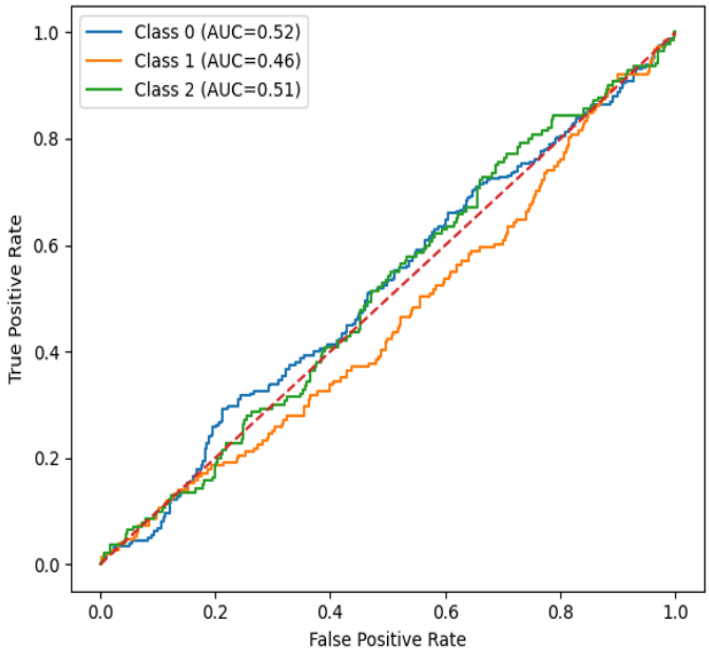


Fig. 3. Multi-class ROC (POS–morphology outcomes

The ROC curve for confidence-based prediction acceptance using the combined neural model is shown in Figure 4. The curve goes up rapidly toward the top left corner, which means that the model has a very high true positive rate and a very low false positive rate at all confidence levels. AUC of 0.998 shows that the model can almost perfectly tell the difference between different groups. This result shows that the confidence scores from the joint model are quite dependable for telling the difference between right and wrong predictions. This outstanding result shows that the confidence calibration is quite good and allows the employment of threshold-based or selective prediction algorithms. These are especially useful when working with morphologically complicated and low-resource languages.

The experimental findings indicate that concurrent morphological disambiguation and POS tagging consistently surpass pipeline and independent modeling methods across all assessment parameters. Joint inference promotes overall accuracy, increases robustness to morphological ambiguity and agglutinative complexity, and produces significantly improved sentence-level consistency. ROC-based evaluations demonstrate that the combined model yields well-calibrated confidence estimates and consistently differentiates between accurate and erroneous predictions, facilitating effective threshold-based decision-making. Collectively, our results validate that joint modeling provides a

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

methodical and empirically enhanced approach for morphology-sensitive processing in agglutinative, low-resource languages.

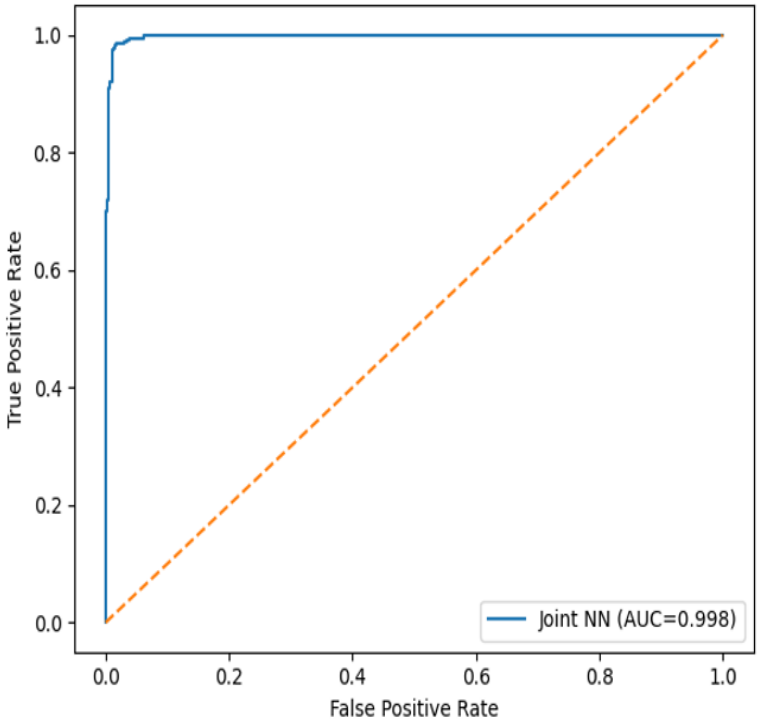


Fig. 4. ROC for confidence-based prediction acceptance

Discussion.

The experimental results demonstrate that coupled morphological disambiguation and part-of-speech tagging represent a more successful modeling framework for agglutinative languages compared to conventional pipeline or individually learned methods. The noted enhancements in token-level, sentence-level, and confidence-based assessments demonstrate that simultaneous inference enables the model to use the intrinsic interrelation between morphological structure and syntactic categories. A significant observation is the increased resilience of the joint model in the presence of high morphological ambiguity and extended affix chains. Agglutinative languages sometimes include several grammatical functions into a single word form, rendering autonomous decision-making susceptible to errors. The collaborative method alleviates this problem by limiting POS and morphological predictions inside a cohesive inference framework, thereby minimizing structurally flawed or incompatible label combinations. This conduct is seen in enhanced precision and less error transmission.

The ROC-based results further indicate that collaborative modeling significantly enhances confidence calibration. The nearly flawless distinction between accurate and inaccurate predictions indicates that the combined model yields more dependable uncertainty assessments than pipeline-based systems. This characteristic is especially

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

significant in low-resource environments, where selective prediction and confidence-aware filtering can enhance the dependability of downstream systems and mitigate the effects of erroneous outputs. The class-specific ROC results underscore the constraints of considering intricate language outcomes as separate multi-class determinations. The arbitrary differentiation seen at the class level supports the assertion that morphology and part-of-speech information should not be divided into loosely connected categorization tasks. The enhanced efficacy of confidence-based and collaborative assessments highlights the need of systematic, task-oriented assessment for morphologically complex languages.

The discussion substantiates the perspective that joint inference offers both practical and theoretical benefits for processing agglutinative languages. Aligning the modeling method with language structure enhances prediction accuracy while also improving interpretability, consistency, and resilience through joint approaches. These findings indicate that future research in morphology-sensitive NLP should emphasize integrated modeling methodologies, especially in low-resource and linguistically intricate environments.

Conclusion.

This study found that combining morphological disambiguation and part-of-speech tagging is a principled and successful method for processing agglutinative languages in low-resource contexts. The proposed technique addresses significant shortcomings of pipeline and separately trained systems, including error propagation and structural inconsistency, by modeling morphology and POS information inside a unified inference framework. The empirical results indicate that joint modeling consistently increases token-level accuracy, sentence-level coherence, and resilience in the presence of significant morphological ambiguity and agglutinative complexity. Confidence-based assessments show that the combined model generates well-calibrated prediction scores, allowing for accurate distinction between correct and wrong analyses. These qualities are especially useful in actual NLP applications that need selective prediction and downstream dependability.

Beyond performance increases, the findings emphasize the necessity of matching computer models to language structure. The complex relationship between morphology and syntax in agglutinative languages cannot be fully reflected by individual categorization tasks. Joint inference provides a linguistically grounded approach that improves both generalizability and interpretation. Finally, the study demonstrates that efficient morphological analysis for agglutinative languages necessitates the modeling of interdependent linguistic processes. Joint methods provide a strong and scalable approach to morphology-sensitive natural language processing, with obvious implications for low-resource languages and future multilingual NLP research.

REFERENCES

Abudouwaili, G., Abiderexiti, K., Yi, N. and Wumaier, A. (2023). Joint learning for low-resource agglutinative language morphological tagging // Proceedings of the 20th SIGMORPHON Workshop. (Jul. 2023), 27–37. https://doi.org/10.18653/v1/2023.sigmorphon-1.4

Aitim A., Abdulla M. (2024). Data Processing and Analysing Techniques in UX Research // Procedia Computer Science, 251, 2024. Pp. 591-596, https://doi.org/10.1016/j.procs.2024.11.154

Aitim, A., & Satybaldiyeva, R. (2025). A comparison of Kazakh language processing models for improving semantic search results // Eastern-European Journal of Enterprise Technologies, 1(2 (133), 66–75. https://doi.org/10.15587/1729-4061.2025.315954

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

Aitim, A. (2025). Building a high-quality annotated corpus for Kazakh NLP: a pipeline approach // Vestnik KazUTB, vol. 4, no. 29, Dec. 2025, https://doi.org/10.58805/kazutb.v.4.29-1092.

Aitim, A. and Satybaldiyeva, R. (2024). A systematic review of existing tools to automated processing systems for Kazakh language // Bulletin of Abai KazNPU. Series of Physical and Mathematical sciences. 87, 3 (Sep. 2024), 106–122. https://doi.org/10.51889/2959-5894.2024.87.3.009.

Ali, S. (2024). Character-aware neural POS tagging for morphologically rich languages. Journal of Information Processing Systems. 20, 2 (Apr. 2024), 315–327. https://doi.org/10.3745/JIPS.02.0197

Aitim, A. (2024). Developing methods for automatic processing systems of Kazakh language // KazATC Bulletin, 133(4), 254–265. https://doi.org/10.52167/1609-1817-2024-133-4-254-265

Aitim, A. (2024). Developing methods for automatic processing systems of Kazakh language // KazATC Bulletin, 133(4), 254–265. https://doi.org/10.52167/1609-1817-2024-133-4-254-265

QNLP - Full Kazakh NLP Suite GitHub repository. Retrieved from- https://github.com/Aigerimhub/qnlp

Cotterell, R. and Heigold, G. (2022). Cross-lingual character-level neural morphological tagging // Transactions of the ACL. 10 (Jun. 2022), 205–220. https://doi.org/10.1162/tacl_a_00444

Carta S.M. (2025). BERT-based POS tagging for under-resourced languages. Proceedings of CLiC 2025. (Sep. 2025), 112–121.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2022). Evaluating morphological generalization in pretrained language models. Proceedings of ACL 2022. (May 2022), 2681–2692. https://doi.org/10.18653/v1/2022.acl-main.193

Gupta, R.P. and Jamwal, S.S. (2025). Deep learning-based morphological analysis for low-resource languages. SN Computer Science. 6, 112 (Jan. 2025). https://doi.org/10.1007/s42979-025-04429-9

García A.T., Villegas M. and Melero M. (2025). Morphology-aware tokenization for sequence labeling. Proceedings of EMNLP 2025. (Nov. 2025), 3124–3136. https://aclanthology.org/2025.emnlp-main.1552

Kann K., Cotterell R. and Schütze H. (2021). Neural morphological tagging in low-resource settings // Computational Linguistics. 47, 3 (Sep. 2021), 487–535. https://doi.org/10.1162/coli_a_00405

Masoka, H. (2025). spaCy-based morphological analysis for Shona. arXiv preprint. (Nov. 2025). https://arxiv.org/abs/2511.16680

Marquard, C., Mawere, S. and Meyer, F. (2025). Neural morphological tagging for low-resource Nguni languages. Proceedings of the Sixth Workshop on African NLP. (Mar. 2025), 210–220. https://aclanthology.org/2025.africanlp-1.31

Müller, T., Schütze, H. and Fraser, A. (2023). Constraint-aware decoding for neural sequence labeling. Proceedings of ACL 2023. (Jul. 2023), 4121–4133. https://doi.org/10.18653/v1/2023.acl-main.292

Nzeyimana, A. (2024). Morphology-aware modeling for low-resource sequence labeling. Findings of NAACL 2024. (Jun. 2024), 145–157. https://doi.org/10.18653/v1/2024.findings-naacl.13

Qi, Y. (2025). Dual-gating neural architecture for POS tagging in low-resource languages. Symmetry. 17, 9 (Sep. 2025), 1561. https://doi.org/10.3390/sym17091561

Stenlund, M., Myneni, H. and Riedel, M. (2025). Morphological segmentation of low-resource Inuktitut using pretrained language models. Proceedings of NoDaLiDa/Baltic-HLT 2025. (May 2025), 688–696. https://aclanthology.org/2025.nodalida-1.69

Saleva, K. and Lignos, C. (2021). Morphology-aware segmentation in low-resource NLP. Proceedings of EACL 2021. (Apr. 2021), 215–225. https://doi.org/10.18653/v1/2021.eacl-main.20

Supriya M., Acharya D.U., Nayak A. and Raghavendra A.S. (2025). Developing a hybrid morphological analyzer for low-resource languages // Applied Sciences. 15, 10 (May 2025), 5682. https://doi.org/10.3390/app15105682

Pakray P., Bandyopadhyay S. and Singh T.D. (2025). Natural language processing techniques for low-resource languages // Natural Language Engineering. 31, 2 (Apr. 2025), 231–259. https://doi.org/10.1017/S1351324924000213

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. 125–137
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.008

NP-COMPLETE ALGORITHMS IN LEGALTECH: HYBRID ARCHITECTURE FOR SEMANTIC RETRIEVAL IN KAZAKHSTAN'S LEGAL ACTS

A. Auyezova[1]*, Aigerim Aitim[1], B. Sinchev[1], W. Wójcik[2]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Lublin University of Technology, Lublin, Poland.
E-mail: a.auyezova@iitu.edu.kz

Anel Auyezova — Master of technical sciences, Assistant-professor, Department of Digital and Intelligent Systems, International Information Technology University, Almaty, Kazakhstan
E-mail: a.auyezova@iitu.edu.kz, https://orcid.org/0000-0001-9860-4491;
Aigerim Aitim — PhD, Associate Professor, Department of Digital and Intelligent Systems, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0003-2982-214X;
Bakhtgerey Sinchev — Candidate of Technical Sciences, Associate Professor, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0001-8557-8458;
Waldemar Wójcik —Lublin University of Technology, Lublin, Poland
https://orcid.org/0000-0002-0843-8053.

© A. Auyezova, Aigerim Aitim, B. Sinchev, W. Wójcik

Abstract. With the growing digitalization of legal practice, there is an increasing need for intelligent systems that can efficiently retrieve and interpret regulatory documents. This study presents a hybrid semantic search system for Kazakhstan's legal acts that integrates multilingual BERT-based retrieval with NP-complete logical verification. The retrieval module identifies semantically relevant legal clauses, while the verification module applies SAT-based algorithms to ensure logical consistency and compliance with legal conditions. The proposed system achieves higher precision and interpretability compared to traditional models such as BM25. Experimental evaluation on a bilingual corpus confirms that combining neural and symbolic reasoning significantly enhances both accuracy and transparency in legal information retrieval.

Keywords: hybrid search, mBERT, NP-complete problems, Subset Sum, natural language processing methods, legal texts

For citation: Auyezova A., Aitim A., Sinchev B., Wójcik W. (2026). NP-complete algorithms in legaltech: hybrid architecture for semantic retrieval in kazakhstan's legal

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

acts // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 125–137. https://doi.org/10.54309/IJICT.2026.26.2.008. (In Eng.).

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

Acknowledgement. *This research was funded by the Grant No. AP26101119 "Development of methods and fast algorithms for solving the NP-complete subset sum problem" of Ministry of Science and Higher Education of the Republic of Kazakhstan.*

ҚАЗАҚСТАННЫҢ ҚҰҚЫҚТЫҚ АКТІЛЕРІНДЕГІ NP-ТОЛЫҚ АЛГОРИТМДЕР: SEMANTIC ІЗДЕУГЕ АРНАЛҒАН ГИБРИДТІ АРХИТЕКТУРА

А. Әуезова[1]*, Ә. Әйтім[1], Б. Синчев[1], В. Вуйчик[2]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Люблин технологиялық университеті, Люблин, Польша.
E-mail: a.auyezova@iitu.edu.kz

Әуезова Анель — техникалық ғылымдар магистрі, ассистент-профессор, Цифрлық және интеллектуалды жүйелер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: a.auyezova@iitu.edu.kz, https://orcid.org/0000-0001-9860-4491;
Әйтім Әйгерім — PhD, қауымдастырылған профессор, Цифрлық және интеллектуалды жүйелер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-2982-214X;
Синчев Бахтгерéй — техника ғылымдарының кандидаты, қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-8557-8458;
Вуйчик Вальдемар — Люблин технологиялық университеті, Люблин, Польша
https://orcid.org/0000-0002-0843-8053.

© А. Әуезова, Ә. Әйтім, Б. Синчев, В. Вуйчик

Аннотация. Құқықтық практиканың цифрландырылуының қарқынды дамуымен нормативтік-құқықтық құжаттарды тиімді іздеп, түсіндіре алатын интеллектуалды жүйелерге қажеттілік артып келеді. Бұл зерттеуде Қазақстанның құқықтық актілеріне арналған көптілді BERT-моделі негізіндегі іздеуді NP-толық логикалық верификациямен біріктіретін гибридті семантикалық іздеу жүйесі ұсынылады. Іздеу модулі семантикалық тұрғыдан релевантты құқықтық ережелерді анықтайды, ал верификация модулі логикалық сәйкестікті және құқықтық талаптарға сәйкестікті қамтамасыз ету үшін SAT негізіндегі алгоритмдерді қолданады. Ұсынылған жүйе BM25 сияқты дәстүрлі модельдермен салыстырғанда жоғары дәлдік пен нәтижелерді түсіндіру мүмкіндігін қамтамасыз етеді. Екітілді корпус негізінде жүргізіл-

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

ген эксперименттік бағалау нейрондық және символдық пайымдауды біріктіру құқықтық ақпаратты іздеудің дәлдігі мен ашықтығын едәуір арттыратынын растады.

Түйінді сөздер: гибридті іздеу, mBERT, NP-толық есептер, Subset Sum, табиғи тілді өңдеу әдістері, құқықтық мәтіндер

Дәйексөздер үшін: Әуезова А., Әйтім Ә., Синчев Б., Вуйчик В. (2026). Қазақстанның құқықтық актілеріндегі NP-толық алгоритмдер: semantic іздеуге арналған гибридті архитектура // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 125–137. https://doi.org/10.54309/IJICT.2026.26.2.008. (Ағыл. тіл.).

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

Алғыс. *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің № AP26101119 «NP-толық ішкі жиындар қосындысы есебін шешу әдістері мен жылдам алгоритмдерін әзірлеу» гранты аясында қаржыландырылды.*

NP-ПОЛНЫЕ АЛГОРИТМЫ В LEGALTECH: ГИБРИДНАЯ АРХИТЕКТУРА ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА В ПРАВОВЫХ АКТАХ КАЗАХСТАНА

А. Ауезова[1]*, А. Айтим[1], Б. Синчев[1], В. Вуйчик[2]

[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Люблинский технологический университет, Люблин, Польша.
E-mail: a.auyezova@iitu.edu.kz

Ауезова Анель — магистр технических наук, ассистент-профессор, кафедра цифровых и интеллектуальных систем, Международный университет информационных технологий, Алматы, Казахстан
E-mail: a.auyezova@iitu.edu.kz, https://orcid.org/0000-0001-9860-4491;
Айтим Айгерим — PhD, ассоциированный профессор, кафедра цифровых и интеллектуальных систем, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0003-2982-214X;
Синчев Бахтгерей — кандидат технических наук, ассоциированный профессор, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0001-8557-8458;
Вуйчик Вальдемар — Люблинский технологический университет, Люблин, Польша
https://orcid.org/0000-0002-0843-8053.

© А. Ауезова, А. Айтим, Б. Синчев, В. Вуйчик

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

Аннотация. В условиях стремительной цифровизации юридической практики возрастает потребность в интеллектуальных системах, способных эффективно осуществлять поиск и интерпретацию нормативно-правовых документов. В настоящем исследовании представлена гибридная система семантического поиска по правовым актам Казахстана, объединяющая поиск на основе многоязычной модели BERT с NP-полной логической верификацией. Модуль поиска выявляет семантически релевантные положения правовых актов, тогда как модуль верификации применяет алгоритмы на основе SAT для обеспечения логической согласованности и соответствия установленным правовым условиям. Предложенная система обеспечивает более высокую точность и интерпретируемость по сравнению с традиционными моделями, такими как BM25. Экспериментальная оценка на двуязычном корпусе подтверждает, что интеграция нейронного и символического рассуждения значительно повышает как точность, так и прозрачность поиска правовой информации.

Ключевые слова: гибридный поиск, mBERT, NP-полные задачи, Subset Sum, методы обработки естественного языка, правовые тексты

Для цитирования: Ауезова А., Айтим А., Синчев Б., Вуйчик В. (2026). NP-полные алгоритмы в legaltech: гибридная архитектура для семантического поиска в правовых актах казахстана // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 125–137. https://doi.org/10.54309/IJICT.2026.26.2.008. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность. *Данное исследование выполнено при финансовой поддержке гранта № AP26101119 «Разработка методов и быстрых алгоритмов решения NP-полной задачи о сумме подмножеств» Министерства науки и высшего образования Республики Казахстан.*

Introduction.

The rapid digitalization of the legal domain has led to an exponential growth of legislative and regulatory documents. In Kazakhstan, as in many other countries, legislation is presented in the form of extensive collections of regulatory acts, codes, and subordinate documents. These materials are made available through electronic systems such as «Adilet», «Paragraph», and «Zakon Online». However, existing legal databases mainly rely on keyword-based search methods, such as TF-IDF and BM25 (Salton et al., 1983; Robertson et al., 2009). These approaches do not fully capture the semantics of queries and fail to consider the logical structure of legal norms.

Recent advances in the automation of legal tasks employ natural language processing (NLP) and specialized language models, including LegalBERT (Chalkidis et al., 2020) and CaseLawBERT (Zheng et al., 2021). While these models demonstrate high effectiveness in English-language legal systems, their applicability to Kazakhstan's bilingual corpora (Russian and Kazakh) remains limited. A promising direction is the retraining of multilingual models, such as mBERT (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), on national legal texts.

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

From a theoretical perspective, the task of identifying relevant legal norms can be formally reduced to NP-complete problems, such as the subset sum problem and the Boolean satisfiability problem (SAT) (Cook, 1971; Papadimitriou, 1994). For example, verifying whether a given set of conditions, such as age, citizenship, or legal status, entails a particular legal consequence (e.g., eligibility for benefits or the right to participate in elections) can be represented as a satisfiability check of a Boolean formula. Alternatively, it can be viewed as the identification of a subset that meets a specified certificate. In this sense, legal information retrieval may be conceptualized as a special case of a general search problem. In such tasks, the correctness of a candidate solution can be verified in polynomial time, whereas identifying the solution itself remains computationally complex (Biere et al., 2009).

In the n-element set of integers X^n, find a subset whose sum of elements is close to certificate S. This subset sum problem is related to the knapsack problem, which was solved by R. Bellman (Bellman, 1956). He proposed a pseudo-polynomial-time algorithm with execution time T = O (nS) based on the dynamic programming method. Since this pseudo-polynomial-time algorithm is a fundamental part of information technology, and Subset Sum is one of the main NP-hard problems, it is important to decide whether the running time can be improved to O (nS). 43 years later, D. Pisinger improved the solution to this problem in time T=O (nS/logS) by using a bit representation of the certificate S (Pisinger, 1999).

In our hybrid architecture for legal texts, the algorithm is applied at the verification level. The retrieval module (mBERT + FAISS) extracts a set of candidates from the corpus of laws, which constitutes the initial unordered set . The task then reduces to verifying whether there exists a subset of conditions Y that satisfies the certificate S—a formalized legal requirement (e.g., age ≥ 16, experience ≥ 5 years, citizenship).

The review of related studies indicates that verifying conditions in legal texts can be reduced to the subset-sum problem, a fundamental NP-complete challenge in algorithmic theory. Therefore, the development of effective algorithms and hybrid models capable of processing such tasks in the applied field of law is a relevant and significant scientific task.

Materials and methods.

This study demonstrates how semantic search and logical verification can be integrated to process unstructured legal documents. It also describes the development of an information system capable of verifying compliance with specified requirements. The proposed approach relies on decision-making logic, which implicitly exhibits NP-complete structures, including Boolean satisfiability and combinatorial constraint satisfaction. The research framework consists of four main stages:

1. Data Preparation – Legal texts were parsed and normalized from the official Adilet database.
2. Semantic Retrieval – A multilingual BERT model (mBERT) fine-tuned on Kazakhstan's legal corpus retrieves candidate clauses based on query embeddings using FAISS.

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

3. Logical Formalization – Retrieved clauses are converted into logical forms representing conditions such as eligibility, obligations, or prohibitions.
4. Verification – The SAT solver checks whether the logical conditions are satisfied given the user's input parameters (e.g., age, citizenship).

This architecture integrates the efficiency of deep learning with the interpretability of formal reasoning. Unlike end-to-end black-box models, the proposed approach allows each decision to be validated through an explicit logical proof.

Figure 1 shows a diagram of the automated procedure for verifying compliance with specified requirements. In the first stage, the texts of regulatory legal acts are converted into structured elements, which enables their formalized analysis. Next, the LegalBERT and mBERT models perform semantic encoding of user queries, forming vector representations that allow the most relevant text fragments to be found by comparing semantic proximity. The extracted sentences are transformed into formal logical constructions using rule modules. At the final stage, NP-complete algorithms are used as constraint solvers, which verify the compliance of user-entered data with established legal conditions.

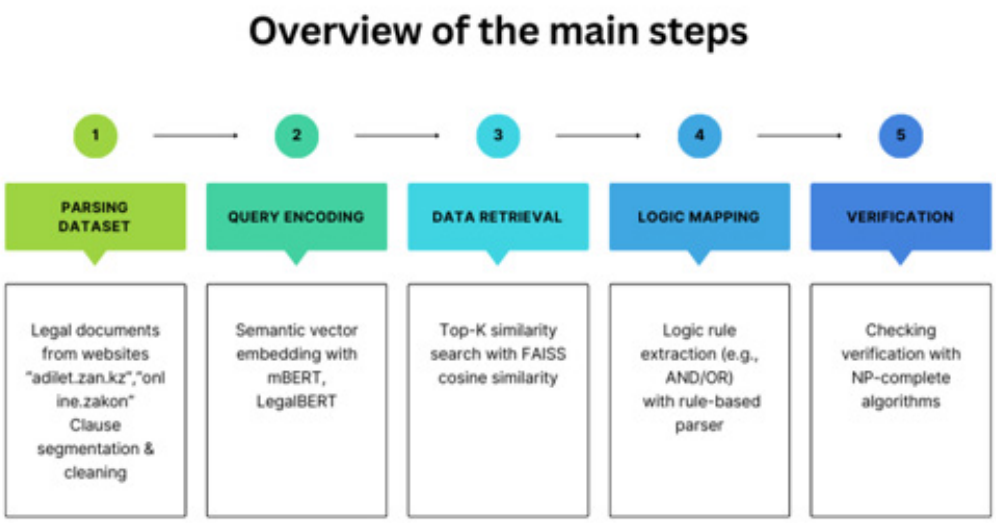


Fig. 1. Diagram of the automated procedure for verifying compliance with specified requirements.

In this article, particular attention is given to the practical aspects of implementation, with an emphasis on the application of algorithms for solving NP-complete problems.

In 2017, K. Birgman proposed a pseudopolynomial a randomized algorithm that runs in O (n + S) time (Bringmann, 2017). More precisely, given an instance of the subset sum problem (X^n, S), the set of all sums $s(X^n, S)$ generated by small subsets of Y is computed $Y \subseteq X^n$ dimensions $|Y| \leq k$ with a constant probability of membership of the certificate S in the set of sums $s(X^n; t)$. The probability value is not specified in the work (Bringmann, 2017; Koiliaris et al., 2017).

It is important to note that a comprehensive review of modern advances in pseudo-

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

polynomial algorithms for solving the subset sum problem is provided in more than 60 scientific papers, including the works of Koiliaris et al. (Konstantinos Koiliaris et al., 2017) and Bringmann (Bringmann, 2017).

Along with pseudo-polynomial algorithms, there are exact algorithms with exponential running time, when in an n -element set X^n there is at least one subset whose sum of elements is equal to *S*. First, these include the works of Horowitz with the execution time of the algorithm $O\left(2^{n/2}\right)$ and the required memory $M = O\left(2^{n/2}\right)$ and Schreppel, Shamir with the execution time of the algorithm $T = O\left(2^{n/2}\right)$ and the required memory $M = O\left(2^{\frac{n}{4}}\right)$ (Horowitz et al., 1974; Schroeppel et al., 1974).

The works introduced the concepts of a polynomial (practical) algorithm and its complexity. A more detailed study of the complexity of solving the subset sum problem was conducted in the works (Krishnamoorth, 2008; Kolpakov et al., 2013). They obtained upper and lower estimates of the maximum exponential complexity of solving this problem using the branch and bound method.

The complexity of the algorithm is directly used in determining the running time of known algorithms for the classical k-SUM problem (Lincoln et al., 2016). It is also applied in methods for searching unstructured information using multiple keywords (C. J. van Rijsbergen, 1979), as well as in techniques for processing and analyzing big data (Min Chen et al., 2014).

It is important to emphasize that determining the complexity of solving the subset sum problem is a fundamental task. It serves as a standard example that can be solved in weakly polynomial time and is widely used across various disciplines in educational programs. As a basic problem in theoretical computer science and educational materials (Lifshits et al., 2009). The k-SUM problem is considered, which is a subproblem of the subset sum problem and is associated with an enumeration method, for which time $O(n^k)$.

It is important to note that in Lincoln for 3-SUM with the sum of three elements equal to zero, the time was found $O\left(\frac{n^2 lglgn}{lgn}\right)$ and memory $O\left(\sqrt{\frac{nlgn}{lglgn}}\right)$ when using the random access memory model (word -RAM model) (Andrea Lincoln et al., 2016).

In the work a unique method and universal algorithms for 2-SUM with time $T = O(n)$ and required memory were developed $M = O(n)$, in other works also for 3-SUM: $T = O\big((n-1)n\big), M = O(n)$ (Sinchev et al., 2019).

In the *n*-element set X^n of positive integers, find a *k*-element subset $X^k(k < n)$ whose sum of elements is equal to the certificate S. The parameter k determines the complexity of the proposed algorithms for solving the problem of the sum of subsets. A new 2019 USPTO patent from the latest source describes a computer system for ultra-fast processing of big data with processing volume $n < +\infty$

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

and speed proportional to execution time $T = O(n) \text{or } T=(n^2)$ or with the required memory $M = O(n) \text{ either } M = (n^2)$ for the parameter $k = 2$ or k = 3 respectively (Sinchev et al., 2019). The uniqueness of the proposed method is based on a new mapping $\tau(S,x) = (S - x)x, \forall x \in X^n,$ whose arguments are the certificate S and the elements x of the set , Fermat's theorem in optimization theory and the merger method. The previous best polynomial algorithms have characteristics $T = O(nlogn), M = O(n) \text{ or } T = O(n^2logn), M = O(n^2).$

The generalized approach polynomial time algorithms for solving NP and methods of information search for the subset sum problem for complexity parameter values $k \geq 4$ and greater include the generalized mapping $\tau(S,z) = (S - z)z, \forall z \in Z^l, |Z^l| = l = C_n^m,$ the combination function C_n^m, Fermat's theorem in optimization, and the merger method with sorting (Sinchev et al., 2019; Sinchev et al., 2020).

In the proposed statements of the first case, the lower bound of the complexity parameter $k_* = 2$ and the upper bound are set $k^* = \frac{n}{2},$ when the lower bound of the complexity parameter $2 \leq k \leq \frac{n}{2}$ and the upper bound $k^* = n - 1,$ are defined similarly $k_* = \frac{n}{2} + 1,$ when $\frac{n}{2} + 1 \leq k \leq n - 1.$ The reducibility of the second case to the first case is shown. The subset $X^k \subseteq X^n (|X^k| = k)$ is formed since the subset Z^l obtained from the set X^n using the combination function C_n^m. The mapping $\tau(S,z) = (S - z)z$ allows us to reduce the complexity parameter k by half and the new power m changes within the range from 1 to n /4. Therefore, a subset is formed X^k based on the pairwise union of two subsets: $X^k = X_i^m \cup X_i^m$ with non-coinciding indices $i \neq j, \ X_i^m \in X^n, X_i^m \in X^n, |X_i^m| = |X_j^m| = m,$ paired indices $(i,j) \in L=(1,2,\dots,l), l = C_n^m.$ The combination generation algorithm provides a one-to-one correspondence between the index $i \in L$ and the indices of the elements *m* of the -element subset X_i^m for any *i.* New universal (simple) polynomial algorithms with execution time T have been developed and the required memory . Moreover, the meaning of the combination C_n^k much more than the meaning of the combination C_n^m $(C_n^m \ll C_n^k)$ for given values of variables *k, m, n.* The multiple reduction in the execution time of the algorithm is determined by the ratio C_n^k / C_n^m and the speed of searching and processing various large data associated with the set of features VVV (volume, velocity, variety) increases by that many times. The obtained results are illustrated with examples. At the same time, a simple mechanism for working with indices of elements of the original set is proposed.

In this work, new methods for solving the subset sum problem and corresponding polynomial algorithms are developed. These methods are applied to the processing and

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

analysis of legal data characterized by the three VVV features (volume, velocity, and variety) as well as to the task of searching within unstructured information. The search for text documents is carried out by keywords (indices) in natural language, used in search engines after the query encoding and data retrieval. The proposed processing and search methods differ fundamentally from existing approaches in terms of time and memory efficiency. They also offer greater simplicity in implementing software products based on the developed algorithms. The time and required memory for a search query with two keywords and searching for the necessary information in big data are proportional to O (n). The search problem by three keywords is reduced to the problem of searching for information with two keywords or to the problem of computational geometry (belonging of three points to one straight line) with the execution time of the algorithm $T=(n^2)$ and the required memory $M = (n^2)$.

Carrying out the experiment

The experiment focused on evaluating the proposed methodology for legal eligibility assessment, which combines semantic retrieval with logic-based satisfiability validation. The system was tested on a benchmark dataset containing 800 user queries and 4,200 annotated legal clauses from diverse areas such as labor law, social benefits, and civil procedure.

To assess retrieval performance, standard measures such as Precision@1, Top-5 Accuracy, and Mean Reciprocal Rank (MRR) were used. A baseline model (BM25) was compared against a multilingual BERT (mBERT) encoder fine-tuned on triplet query–positive–negative examples, with an additional SAT-based constraint validation layer to ensure logical correctness of retrieved results.

Table 1. Comparison of models on the legal corpus

Model	Precision@1	Top-5 Accuracy	MRR	SAT-check accuracy
BM25	0.62	0.74	0.68	0.78
mBERT	0.78	0.87	0.82	0.89
mBERT + SAT-check	0.84	0.91	0.88	0.92

The results demonstrated that BM25 performs reasonably well for keyword-based matches but fails to capture deeper semantic and contextual relations. The mBERT model showed significantly better accuracy in matching queries with relevant legal clauses. Furthermore, the integration of SAT-checking improved performance by filtering out results that did not satisfy logical constraints. This improvement was particularly evident in queries involving age restrictions, contractual obligations, and procedural deadlines.

To provide a more detailed evaluation of retrieval behavior beyond discrete accuracy metrics, a precision–recall (PR) analysis was conducted. This curve illustrates how each model balances precision (the relevance of retrieved clauses) against recall (the proportion of all relevant clauses successfully retrieved). Figure 1 compares the PR performance of BM25, mBERT, and the proposed mBERT + SAT hybrid system.

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

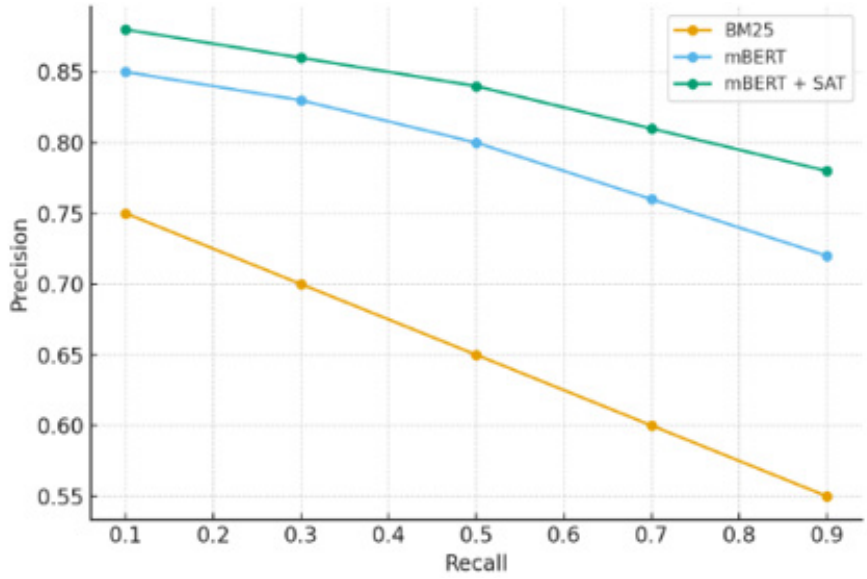


Fig. 1. Precision–Recall curve for legal clause retrieval.

The results in Figure 1 show that the hybrid model maintains consistently higher precision across the entire recall range. Unlike BM25, which rapidly loses precision as recall increases, the mBERT + SAT curve remains stable, indicating that logical validation helps filter out irrelevant clauses even when more results are retrieved. This stability demonstrates the system's ability to preserve retrieval quality under broader search conditions, confirming the advantage of combining semantic and symbolic reasoning in legal information retrieval.

To complement the precision–recall analysis and gain a deeper understanding of model behavior, an additional error-based evaluation was performed. While the PR curve demonstrates overall improvements in precision and recall balance, the error analysis provides insights into where the model still struggles — particularly across different categories of legal clauses. This step is essential for identifying the types of legal reasoning that most benefit from logical verification and for guiding future refinements of hybrid architecture.

To further analyze model behavior and identify specific sources of retrieval errors, an error distribution study was conducted across several categories of legal clauses. The evaluation included four representative types — age restrictions, benefit eligibility, contractual obligations, and procedural deadlines. These categories were chosen because they frequently appear in practical legal queries and require different levels of logical reasoning and contextual interpretation. Figure 2 presents a comparative overview of the misclassification counts for mBERT and the hybrid mBERT + SAT model.

Figure 2 demonstrates that the integration of SAT-based validation substantially reduces the number of errors across all categories. The largest improvements are observed in *benefit eligibility* and *procedural deadline* clauses, where the logical dependencies between conditions are particularly complex. This indicates that formal constraint checking helps

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

the model better interpret compound legal requirements and avoid false matches. Overall, the hybrid approach achieves a 40–50% reduction in misclassifications, confirming its effectiveness in ensuring logical consistency and interpretability in legal retrieval tasks.

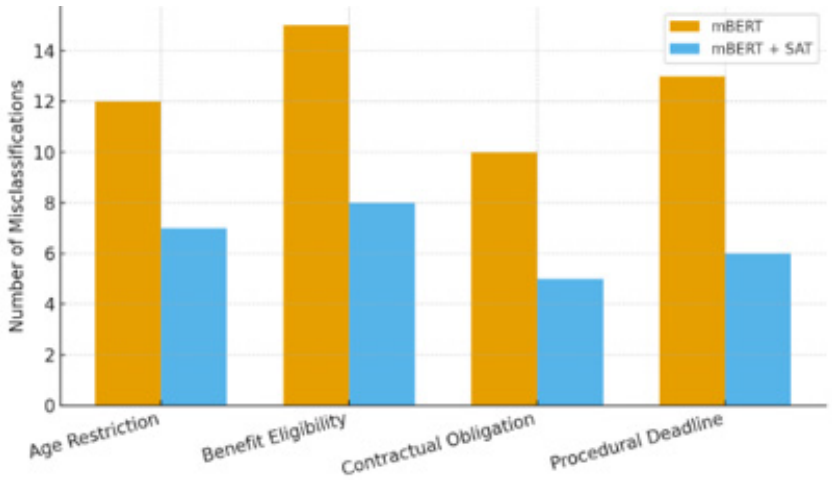


Fig. 2. Error analysis by clause type.

Overall, the experimental findings confirm that combining mBERT with SAT-checking provides both semantic precision and formal correctness, offering a robust solution for legal information retrieval in multilingual environments.

Results and discussion.

The results confirm that legal retrieval in complex bilingual corpora can be modeled as an NP-complete search problem, requiring both semantic understanding and logical validation. The hybrid system demonstrated superior performance across all evaluation metrics. Importantly, the inclusion of a formal SAT-based validation stage enhanced transparency — an essential aspect for legal applications where interpretability and correctness are as important as accuracy.

Nevertheless, several challenges remain. The primary difficulties involve ensuring the scalability of SAT-solvers for large document collections, maintaining consistent performance across two languages with distinct grammatical structures, and addressing data sparsity in underrepresented legal domains. Future work will focus on expanding the bilingual corpus, developing optimized heuristics for NP-complete verification, and integrating large language models to enhance contextual reasoning and automate the generation of formal logical representations.

Beyond its research contributions, the proposed approach holds practical value for the development of intelligent legal assistants, automated eligibility checkers, and digital platforms for legal consultation and compliance verification. Such systems could support government agencies, law firms, and citizens by providing accurate, explainable, and legally consistent answers, thereby advancing the digital transformation of the justice sector in Kazakhstan and beyond.

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

Conclusion.

The experiment confirmed that legal search in a multilingual corpus is an NP-complete task, as it requires checking the satisfaction of a set of logical conditions formulated based on legal norms. The main difficulty was ensuring the correct verification of logical connections with many documents, as well as the need for precise formalism for the SAT module and the limited amount of annotated data.

The integration of the SAT verification mechanism made it possible to filter out solutions that did not satisfy the formal conditions, which increased the accuracy and reliability of the results. Future research plans include expanding the document corpus and developing optimized heuristics for SAT verification to speed up the processing of large data sets and improve the quality of searching for complex legal combinations. This approach demonstrates a practical solution to an NP-complete problem, combining semantic correspondence and formal logical validation.

REFERENCES

Adamansky A. (2018). Overview of full-text search methods and algorithms. — Novosibirsk, Novosibirsk State University, Pp. 26. [In Eng.]

Biere A., Heule M., van Maaren H., Walsh T. (2009). Handbook of Satisfiability. IOS Press. ISBN: 978–1586039295. [In Eng.]

Bellman R. (1956). Notes on the theory of dynamic programming iv - maximization over discrete sets. Naval Research Logistics Quarterly. — Vol. 3. — No. 1–2. Pp. 67–70. [In Eng.]

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020) Unsupervised Cross-Lingual Representation Learning at Scale. Proceedings of ACL. DOI:10.18653/v1/2020.acl-main.747. [In Eng.]

Cook, S.A. (1971) The Complexity of Theorem-Proving Procedures. Proceedings of STOC. DOI:10.1145/800157.805047. [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.]

Chen M., Mao S., Zhang Y., Victor C.M. Leung (2014). Big Data. Related Technologies, Challenges, and Future Prospects. - Spinger. Pp.100. [In Eng.]

David Pisinger (1999). Linear time algorithms for knapsack problems with bounded weights // Journal of Algorithms. — Vol. 33. — No. 1. Pp. 1–14. [In Eng.]

Horowitz E., Sanni S. (1974). Computing Partitions with Application to the Knapsack Problem // Journal of the ACM(JACM), T21. Pp. 277–292. [In Eng.]

Karl Bringmann.(2017). A near-linear pseudopolynomial time algorithm for subset sum. To appear in SODA ‘17. 2017. //arXiv:1610.04712v2[cs.Ds] 8 Jan. Pp. 18[In Eng.]

Koiliaris K. Xu C. (2017). A Faster pseudopolynomial time algorithm for subset sum. To appear in SODA ‘17, 2017. //arXiv:1610.04712v2[cs.Ds] Pp. 18[In Eng.]

Krishnamoorth B. (2008). Bounds on the size of branch-and-bound proofs for integer knapsacks // OR Letters. — Vol. 36. — No. 1. Pp. 19–25. [In Eng.]

Kolpakov R.M., Posypkin M.A. (2013). Upper bound for the number of branchings for the subset sum problem. // Mat. Questions of Cybernetics. Issue 18. — M.: Fizmatlit. Pp. 213–226. [In Eng.]

Kulikov A. (2009). Algorithms for NP-hard problems. Lectures // Computer Science Club of the St. Petersburg Branch of the Steklov Mathematical Institute. [In Eng.]

Lincoln A., Vassilevska V.W., Wang J., Ryan R.W. (2016). Deterministic Time-Space Tradeoffs for k-SUM // 43rd International Colloquium on Automata, Languages and Programming (ICALP). Pp. 16. [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. [In Eng.]

Schroeppel R., Shamir A. (1974). A T = O (2 n/2), S=O(2 n/4) Algorithm for Certain NP-Complete Problem // SIAM Journal on Computing. — Vol.10. — No. 3. Pp. 456–464. [In Eng.]

Salton, G. and McGill, M.J. (1983) Introduction to Modern Information Retrieval. McGraw-Hill. ISBN: 978–0070544840. [In Eng.]

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

Sinchev B., Sinchev B., Akzhanova J., Mukhanova A.M. (2019). New methods of information search. I. // News of the National Academy of Sciences of Kazakhstan, Series of Geology and Technical Sciences. — Vol. 3. — No. 435. Pp. 240–246. [In Eng.]

Sinchev B., Sinchev A.B., Akzhanova Z.A. (2019). Computing network architecture for reducing a computing operation time and memory usage associated with determining, from a set of data elements, a subset of at least two data elements, associated with a target computing operation result. // Patent in USPTO, Pp. 38. [In Eng.]

Sinchev B., Sinchev A.B., Akzhanova J., Issekeshev Y., Mukhanova A.M. (2020). Polynomial time algorithms for solving NP-complete problems. // News of the National Academy of Sciences of Kazakhstan, Series of Geology and Technical Sciences. — Vol. 3. — No. 441. Pp. 97–101. [In Eng.]

Pires T., Schlinger E. and Garrette D. (2019) How Multilingual is Multilingual BERT? Proceedings of ACL. DOI:10.18653/v1. Pp. 19–1493. [In Eng.]

Papadimitriou, C.H. (1994) Computational Complexity. Addison-Wesley. ISBN: 978-0201530827. [In Eng.]

Van Rijsbergen C.J. (1979). Information Retrieval. Dept. of Computer Science. University of Glasgow. [In Eng.]

Zheng, L., Guha, N., Anderson, B., Henderson, P., and Manning, C.D. (2021) When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. Proceedings of ACL. DOI:10.18653/v1/2021.acl-long. 309. [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.2. Number 26 (2026). Pp. 138–152
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.009

EVALUATING OPTIMAL CONFIGURATIONS OF SHALLOW CNNS ON LOW-COMPLEXITY DATASETS

H. Yaroslav[1]*, S. Paliy[1], T. Honcharenko[2]
[1]Taras Shevchenko national university of Kyiv, Kyiv, Ukraine;
[2]Kyiv National University of Construction and Architecture, Kyiv, Ukraine.
E-mail: hozakya@fit.knu.ua

Yaroslav Hozak — PhD Candidate, Taras Shevchenko national university of Kyiv, Kyiv, Ukraine
E-mail: hozakya@fit.knu.ua, https://orcid.org/0009-0008-2963-7204;
Sergiy Paliy — PhD, Associate Professor, Taras Shevchenko national university of Kyiv, Kyiv, Ukraine
https://orcid.org/0000-0001-9742-1116;
Tetyana Honcharenko — Doctor of Technical Sciences, Professor, Kyiv National University of Construction and Architecture, Kyiv, Ukraine
https://orcid.org/0000-0003-2577-6916.

© Y. Hozak, S. Paliy, T. Honcharenko

Abstract. This study investigates the impact of network depth, kernel size configurations and convolutional layer types on the performance of convolutional neural networks (CNNs) for image classification tasks. The objective is to systematically assess how variations in kernel size, convolutional strategy, and depth contribute to classification accuracy and computational efficiency. We analyze both shallow and moderately deep network variants to disentangle the relative importance of kernel modifications versus depth scaling. Using the MNIST and CIFAR-10 datasets, we evaluate kernel sizes ranging from 3×3 to 11×11, convolutional configurations including standard, dilated, and strided convolutions, and network depths spanning from two to eight convolutional layers, across multiple input resolutions (32×32, 64×64, 128×128, and 256×256 pixels). Results indicate that for simple datasets like MNIST, depth scaling provides limited benefits, as high accuracy (~97–98%) is achievable even with shallow networks regardless of kernel variation. In contrast, for CIFAR-10, smaller kernels in early layers combined with increased depth capture fine-grained details more effectively, while larger kernels in later layers enhance generalization at higher resolutions. Dilated convolutions offer notable improvements for high-resolution images by reducing inference time (up to a 2× speedup at 256×256) without sacrificing accuracy, particularly in deeper models. Strided convolutions, although

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

computationally efficient, consistently degrade accuracy due to spatial information loss across depths. These findings suggest that optimal performance arises from a balanced interplay between kernel configuration, convolutional strategy, and network depth. Future research may explore adaptive architectures that dynamically adjust these factors to improve robustness across diverse datasets and resolutions.

Keywords: convolutional neural networks (CNN), neural networks (NN), deep learning, artificial intelligence (AI)

For citation: Hozak Y., Paliy S., Honcharenko T. (2026). Evaluating optimal configurations of shallow CNNs on low-complexity datasets // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 138–152. https://doi.org/10.54309/IJICT.2026.26.2.009. (In Eng.).

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

КҮРДЕЛІЛІГІ ТӨМЕН ДЕРЕКТЕР ЖИЫНТЫҚТАРЫНДА ТЕРЕҢ ЕМЕС ОРАМДЫ НЕЙРОНДЫҚ ЖЕЛІЛЕРДІҢ (CNN) ОҢТАЙЛЫ КОНФИГУРАЦИЯЛАРЫН БАҒАЛАУ

Я. Гозак[1]*, С. Палий[1], Т. Гончаренко[2]

[1]Тарас Шевченко атындағы Киев ұлттық университеті, Киев, Украина;
[2]Киев ұлттық құрылыс және сәулет университеті, Киев, Украина.
E-mail: hozakya@fit.knu.ua

Гозак Ярослав — докторант, Тарас Шевченко атындағы Киев ұлттық университеті, Киев, Украина
E-mail: hozakya@fit.knu.ua, https://orcid.org/0009-0008-2963-7204;
Палий Сергий — PhD, қауымдастырылған профессор, Тарас Шевченко атындағы Киев ұлттық университеті, Киев, Украина
https://orcid.org/0000-0001-9742-1116;
Гончаренко Татьяна — техника ғылымдарының докторы, профессор, Киев ұлттық құрылыс және сәулет университеті, Киев, Украина
https://orcid.org/0000-0003-2577-6916.

© Я. Гозак, С. Палий, Т. Гончаренко

Аннотация. Бұл зерттеуде кескіндерді жіктеу міндеттерін шешу кезінде орамды нейрондық желілердің (CNN) өнімділігіне желі тереңдігінің, ядро өлшемдерінің конфигурациялары мен орамды қабаттар түрлерінің әсері зерттеледі. Зерттеудің мақсаты — ядро өлшемінің, орамдау стратегиясының және желі тереңдігінің өзгеруі жіктеу дәлдігі мен есептеу тиімділігіне қалай әсер ететінін жүйелі түрде бағалау. Терең емес және орташа тереңдіктегі желілер қарастырылып, бұл ядро модификацияларының желі тереңдігін арттырумен салыстырғандағы салыстырмалы маңыздылығын анықтауға мүмкіндік береді. MNIST және CIFAR-10 деректер жиын-

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

тықтарын пайдалана отырып, 3×3-тен 11×11-ге дейінгі ядро өлшемдері, стандартты, дилатациялық және страйдты орамдарды қамтитын орамдау конфигурациялары, сондай-ақ кіріс кескіндерінің әртүрлі ажыратымдылықтарында (32×32, 64×64, 128×128 және 256×256 пиксель) екіден сегізге дейінгі орамды қабаттан тұратын желі тереңдігі бағаланды. Нәтижелер MNIST сияқты қарапайым деректер жиынтықтары үшін желі тереңдігін арттыру тиімділіктің шектеулі ғана өсуін қамтамасыз ететінін көрсетті, себебі ядро өлшемдерінің өзгерістеріне қарамастан, терең емес желілердің өзінде жоғары дәлдікке (~97–98 %) қол жеткізіледі. Ал CIFAR-10 деректер жиынтығы үшін желінің бастапқы қабаттарында өлшемі кіші ядроларды пайдалану желі тереңдігін арттырумен үйлескен жағдайда ұсақ бөлшектерден тұратын белгілерді тиімдірек бөліп алуға мүмкіндік береді. Сонымен қатар, кейінгі қабаттарда өлшемі үлкен ядроларды қолдану кескіндердің ажыратымдылығы жоғары болған кезде модельдің жалпылау қабілетін жақсартуға ықпал етеді. Дилатациялық орамдар жоғары ажыратымдылықтағы кескіндермен жұмыс істеу кезінде дәлдікті төмендетпей, әсіресе тереңірек модельдерде, инференс уақытын қысқарту арқылы тиімділіктің айтарлықтай артуын қамтамасыз етеді (256×256 ажыратымдылықта екі есеге дейін жеделдету). Страйдты орамдар есептеу тұрғысынан тиімді болғанына қарамастан, желі тереңдігінің әртүрлі деңгейлерінде кеңістіктік ақпараттың жоғалуына байланысты дәлдіктің тұрақты төмендеуіне алып келеді. Алынған нәтижелер оңтайлы өнімділікке ядро конфигурациясын, орамдау стратегиясын және желі тереңдігін теңгерімді үйлестіру арқылы қол жеткізілетінін көрсетеді. Болашақта әртүрлі деректер жиынтықтарымен және кескін ажыратымдылықтарымен жұмыс істеу кезінде модельдердің тұрақтылығын арттыру мақсатында аталған параметрлерді динамикалық түрде реттей алатын бейімделмелі архитектураларды зерттеу орынды.

Түйінді сөздер: орамды нейрондық желілер (CNN), нейрондық желілер (NN), терең оқыту, жасанды интеллект (AI)

Дәйексөздер үшін: Гозак Я., Палий С., Гончаренко Т. (2026). Күрделілігі төмен деректер жиынтықтарында терең емес орамды нейрондық желілердің (CNN) оңтайлы конфигурацияларын бағалау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 138-152. https://doi.org/10.54309/IJICT.2026.26.2.009. (Ағыл. тіл.).

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

ОЦЕНКА ОПТИМАЛЬНЫХ КОНФИГУРАЦИЙ НЕГЛУБОКИХ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ НА ДАТАСЕТАХ НИЗКОЙ СЛОЖНОСТИ

Я. Гозак[1]*, С. Палий[1], Т. Гончаренко[2]

[1]Киевский национальный университет имени Тараса Шевченко, Киев, Украина;
[2]Киевский национальный университет строительства и архитектуры, Киев, Украина.

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

E-mail: hozakya@fit.knu.ua

Гозак Ярослав — докторант, Киевский национальный университет имени Тараса Шевченко, Киев, Украина
E-mail: hozakya@fit.knu.ua, https://orcid.org/0009-0008-2963-7204;
Палий Сергий — PhD, ассоциированный профессор, Киевский национальный университет имени Тараса Шевченко, Киев, Украина
https://orcid.org/0000-0001-9742-1116;
Гончаренко Татьяна — доктор технических наук, профессор, Киевский национальный университет строительства и архитектуры, Киев, Украина
https://orcid.org/0000-0003-2577-6916.

© Я. Гозак, С. Палий, Т. Гончаренко

Аннотация. В настоящем исследовании изучается влияние глубины сети, конфигураций размеров ядер и типов свёрточных слоёв на производительность свёрточных нейронных сетей (CNN) при решении задач классификации изображений. Цель исследования заключается в систематической оценке того, каким образом изменения размера ядра, стратегии свёртки и глубины сети влияют на точность классификации и вычислительную эффективность. Рассматриваются как неглубокие, так и сети умеренной глубины, что позволяет определить относительную значимость модификаций ядер по сравнению с увеличением глубины сети. С использованием наборов данных MNIST и CIFAR-10 оцениваются размеры ядер от 3×3 до 11×11, конфигурации свёрток, включающие стандартные, дилатированные и страйдовые свёртки, а также глубина сети от двух до восьми свёрточных слоёв при различных разрешениях входных изображений (32×32, 64×64, 128×128 и 256×256 пикселей). Результаты показывают, что для простых наборов данных, таких как MNIST, увеличение глубины сети обеспечивает ограниченный прирост эффективности, поскольку высокая точность (~97–98 %) достигается даже при использовании неглубоких сетей независимо от вариаций размеров ядер. В отличие от этого, для CIFAR-10 использование ядер меньшего размера на ранних слоях в сочетании с увеличением глубины сети позволяет более эффективно выделять мелкодетализированные признаки, тогда как более крупные ядра на последующих слоях способствуют улучшению способности модели к обобщению при высоком разрешении изображений. Дилатированные свёртки обеспечивают заметное повышение эффективности при работе с изображениями высокого разрешения за счёт сокращения времени инференса (до двукратного ускорения при разрешении 256×256) без снижения точности, особенно в более глубоких моделях. Страйдовые свёртки, несмотря на вычислительную эффективность, приводят к устойчивому снижению точности вследствие потери пространственной информации на различных уровнях глубины сети. Полученные результаты свидетельствуют о том, что оптимальная производительность достигается благодаря сбалансированному сочетанию конфигурации ядер, стратегии свёртки и глубины сети.

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

В дальнейшем целесообразно исследовать адаптивные архитектуры, способные динамически регулировать указанные параметры для повышения устойчивости моделей при работе с различными наборами данных и разрешениями изображений.

Ключевые слова: Свёрточные нейронные сети (CNN), нейронные сети (NN), глубокое обучение, искусственный интеллект (AI)

Для цитирования: Гозак Я., Палий С., Гончаренко Т. (2026). Оценка оптимальных конфигураций неглубоких свёрточных нейронных сетей на датасетах низкой сложности // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 138-152. https://doi. org/10.54309/IJICT.2026.2.009. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Recent advances in computer vision have accelerated progress in robotics and automated systems, supporting applications such as autonomous vehicles, unmanned aerial vehicles (UAVs), and industrial automation. At the core of these systems lies the capacity to interpret visual information accurately and efficiently, a task predominantly driven by convolutional neural networks (CNNs). Although CNNs have achieved outstanding results in image classification, object detection, and scene understanding, their performance is highly dependent on architectural choices, including network depth, kernel design, and dataset characteristics.

For robotics and autonomous navigation, particularly in UAVs, efficient and reliable vision systems are essential to ensure real-time decision-making in dynamic environments. Designing CNNs that strike a balance between accuracy, computational efficiency, and robustness remains challenging. Large, deep networks (Jordan, 2024; He et al., 2016) can achieve strong accuracy on datasets such as CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2006), but their computational demands are often incompatible with the constraints of embedded and edge devices. This motivates the study of smaller architectures, where trade-offs between kernel configuration, convolutional strategies, and network depth play a decisive role.

Computation efficiency is influenced by kernel parameters such as size, stride (LeCun et al., 1998; Krizhevsky et al., 2012), and dilation (Yu et al., 2016; Chen et al., 2018). Early architectures such as AlexNet (Krizhevsky et al., 2012) employed large kernels in the initial layers to capture spatial structure and reduce image resolution at an early stage, thereby lowering the computational burden of subsequent layers. However, large kernels have seen limited adoption because they often fail to preserve fine-grained details (Simonyan et al., 2015). Strided convolutions can simultaneously enlarge the receptive field and reduce resolution, replacing pooling in some designs, but they commonly sacrifice spatial precision. Dilated convolutions, introduced by Yu and Koltun (Yu et al., 2016), expand the receptive field without additional parameters, enabling efficient aggregation of multi-scale contextual information, particularly in images with high spatial redundancy.

We hypothesize that the interplay between kernel configuration and network depth

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

strongly influences CNN performance, especially in resource-constrained scenarios. While shallow networks help isolate the effects of kernel modifications, increasing depth allows us to investigate how these effects scale as architectures grow more complex.

We jointly study kernel size, convolution type (standard/strided/dilated), and depth across MNIST (LeCun et al., 1998) and CIFAR-10 at multiple resolutions to quantify impacts on accuracy, efficiency, and generalization.

Materials and methods.

The MNIST and CIFAR-10 datasets have served as foundational benchmarks for evaluating the performance of Convolutional Neural Networks (CNNs) in image classification tasks. Over the years, numerous studies have explored CNN architectures and their performance on these datasets, contributing to a deeper understanding of the design and optimization of neural networks.

LeNet, one of the earliest CNN architectures, was designed for handwritten digit classification on MNIST and achieved remarkable accuracy for its time. Despite its simplicity, LeNet remains relevant as a baseline for shallow networks on MNIST and other small datasets. However, its limited depth and capacity make it less effective on more complex datasets like CIFAR-10, which contain higher variability in object categories and more intricate features (LeCun et al., 1998).

Subsequent architectures such as AlexNet and GoogLeNet have demonstrated significant improvements in handling larger and more complex datasets, including CIFAR-10. AlexNet introduced innovations like ReLU activations and dropout, which improved network performance and generalization. GoogLeNet leveraged inception modules to reduce computational complexity while maintaining high accuracy, laying the groundwork for efficient model design (Krizhevsky et al., 2012).

ResNet (Residual Networks) revolutionized CNN architecture by introducing skip connections, which address the vanishing gradient problem in deep networks. This innovation allowed for the successful training of much deeper networks, achieving state-of-the-art performance on CIFAR-10. ResNet-20 to ResNet-56, in particular, demonstrated the value of residual blocks in improving feature learning while maintaining computational efficiency (He et al., 2016).

Beyond traditional architectures, studies have proposed more compact and efficient models tailored for CIFAR-10 and similar datasets. For example, the High Performance SqueezeNex architecture optimized the design of convolutional layers to balance accuracy and efficiency on resource-constrained devices. Similarly, CondenseNeXt achieved high accuracy while significantly reducing computational costs, making it suitable for embedded systems and edge applications (Duggal et al., 2020).

Adaptive and structured CNN designs have also been explored for these datasets. A Structurally Regularized CNN Architecture via Adaptive Subband Decomposition employed an adaptive filter bank structure to improve feature extraction and reduce redundancy in the network. This approach demonstrated state-of-the-art performance on both MNIST and CIFAR-10 while lowering computational costs (Sinha et al., 2025).

In summary, a wide range of architectures and methodologies have been developed

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

to optimize CNN performance on MNIST and CIFAR-10. These studies highlight the evolution of CNN design from simple, shallow networks to highly optimized, efficient architectures capable of tackling diverse challenges presented by modern datasets.

Understanding the impact of kernel configurations in shallow convolutional neural networks (CNNs) is critical for designing efficient and accurate models, particularly when dealing with datasets like MNIST and CIFAR10. The following methodologies outline approaches to systematically analyze kernel size importance:

Network configuration for test setup

We used the following common setup across all training iterations:

- Optimizer: Stochastic Gradient Descent
- Batch size: 16
- Momentum: 0
- Weight decay: no
- Data preprocessing: No data augmentation. We only scale resolutions per Methodology-B

For kernel configuration tests we use a network similar to the inner module described by Isong N. (Isong, 2025) such that it has the first convolutional layer followed by max pooling layer and the second conv layer also followed by max pooling one. The last layer is traditionally an FC layer as shown in Figure 1(a). The architecture is also depicted in Table 1.

Table 1. Example 2-layered architecture for 32x32 image

Type	Filter shape	Input size
Conv	$N \times N \times 3 \times 32$	$32 \times 32 \times 3$
Max pooling	Pool 2×2	$32 \times 32 \times 32$
Conv	$M \times M \times 32 \times 64$	$16 \times 16 \times 32$
Max pooling	Pool 2×2	$16 \times 16 \times 64$
FC	4096×10	$8 \times 8 \times 64$
Softmax	Classifier	$1 \times 1 \times 10$

Each model was trained for 10 epochs. Normally 10 epochs is not enough for most networks (Jordan, 2024), but as we show further this number of epochs provides good balance between training speed and training completeness, considering small network size and absence of data augmentation. Learning rate is set constant to 0.05.

Controlled Experimental Design.

To isolate the effects of kernel configuration, experiments are structured to minimize interference from other variables.

Grid Search: Systematically vary kernel sizes across all combinations for the first and second convolutional layers (e.g., 3x3, 3x5, …, 11x11) as well as change other parameters (stride, dilation) and depth. Analyze their effects on accuracy and computational cost.

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

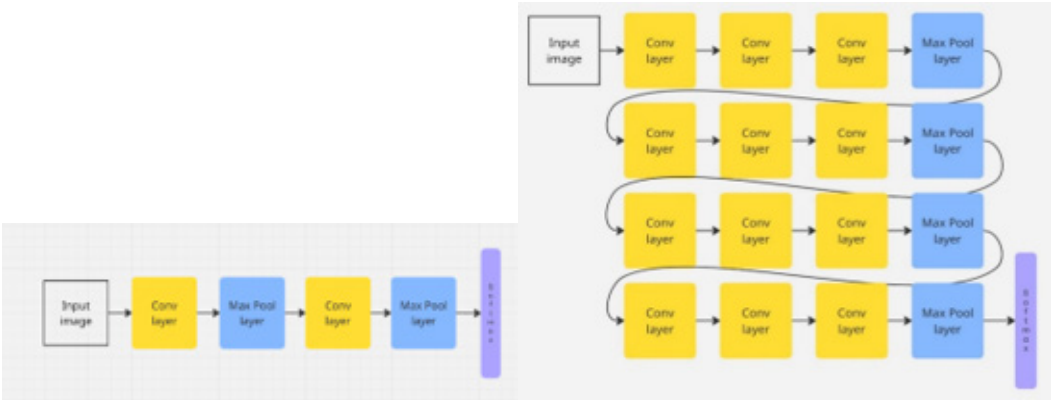


Fig. 1. Experiment model architecture.

For deeper networks (layers 3-9) we increase number of max pool operations up to 4 depending on layers number. Also we shift max pool operations to be so they are placed evenly across the layers as shown in Figure 1(b).

We also train those models for 20 epochs. Learning rate is set to 0.05 for the first 10 epochs and then decreased to 0.01 for remaining 10 epochs, following the approach of (Mehta et al., 2018; Mehta et al., 2019).

Table 2. Example 9-layered architecture for 32x32 image

Type	Filter shape	Input size
Conv	$N \times N \times 3 \times 32$	$32 \times 32 \times 3$
Conv	$M \times M \times 32 \times 32$	$32 \times 32 \times 32$
Conv	$3 \times 3 \times 32 \times 32$	$32 \times 32 \times 32$
Max pooling	Pool 2×2	$32 \times 32 \times 32$
Conv	$3 \times 3 \times 16 \times 64$	$16 \times 16 \times 32$
Conv	$3 \times 3 \times 64 \times 64$	$16 \times 16 \times 64$
Max pooling	Pool 2×2	$16 \times 16 \times 64$
Conv	$3 \times 3 \times 64 \times 128$	$8 \times 8 \times 64$
Conv	$3 \times 3 \times 128 \times 128$	$8 \times 8 \times 128$
Max pooling	Pool 2×2	$8 \times 8 \times 128$
Conv	$3 \times 3 \times 128 \times 125$	$4 \times 4 \times 128$
Conv	$3 \times 3 \times 256 \times 256$	$4 \times 4 \times 256$
Max pooling	Pool 2×2	$4 \times 4 \times 256$
FC	8192×10	$8 \times 8 \times 128$
Softmax	Classifier	$1 \times 1 \times 10$

Fixed Parameters: Keep other hyperparameters, such as learning rate, batch size, and the number of kernels, constant to ensure the focus remains on kernel size.

Dataset Resolution: Test the models on datasets at multiple resolutions (e.g., CIFAR-10 scaled to 32x32, 64x64, 128x128 and 256x256 pixels) to examine the correlation

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

between kernel size and input image dimensions for the best results.

Performance Metrics and Visualization.

Assess kernel size impact through both quantitative and qualitative metrics:

Accuracy and Loss Curves: Measure classification accuracy and training loss for each kernel configuration.

Computational Cost: Profile the time required for training and inference to understand the trade-off between kernel size and efficiency.

Heatmaps: Plot performance metrics as heatmaps, with the first layer kernel size on one axis and the second layer kernel size on the other.

Discussion and results.

MNIST and CIFAR-10 datasets allow us to pay more attention to details such as kernel size configuration and iterate experiments fast thanks to small size of images. Despite having images of comparable size, the datasets are quite different in complexity. While MNIST provides simple images of handwritten digits CIFAR-10 provides comparatively complex images of environment such as trucks, buildings of animals, even though the images are very simplified due to their size.

Assessing kernel size impact on accuracy with original datasets

Each configuration is trained on MNIST and CIFAR-10 across predefined kernel/ resolution grids (see Methodology-B).

The first layer consists of 16 channels and the second layer consists of 32 channels. It is important to ensure the number of channels in each layer is devisable by 32 (most common architectures have their first layers with exactly 32 channels and double this number during each down sampling). The reason for it is that GPU kernels run threads in groups of 32, which means that different numbers of channels leads to inefficient use of GPU threads during training (Hozak et al., 2024)

Table 3. Best network results for channels' configurations on CIFAR-10

Channels configuration	Best accuracy, %	Time, ms	Model configuration
16, 32	63.15	0.093	3x3, 5x5
32, 32	62.80	0.148	3x3, 11x11
32, 64	70.24	0.131	3x3, 11x11

Although, as said above, it is reasonable to choose the first layer with 32 channels we found it inefficient for simple datasets. As shown in Table 3, for the same training setup model accuracy slightly drops to 62 % while inference time increases considerably on CIFAR-10 dataset. Under the training schedule in Sec. 3A, this configuration likely requires more epochs for convergence (see Figure 2).

We can see that in the case of both layers having equal number of 32 channels the best result demonstrates a network with the second layer of 11x11. It suggests that the first layer may contain too much redundant/duplicate information and big kernel size helps to better generalize information. After training models for 30 epochs, a model with a configuration 3x3 in the first layer and 9x9 in the second layer demonstrates the best results

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

(see Figure 3). Also, we see that further improvements do not happen after 20 epochs for this configuration. Some configurations experience accuracy jittering in later epochs that suggest the network size is not sufficient to effectively classify images with complex features from CIFAR-10.

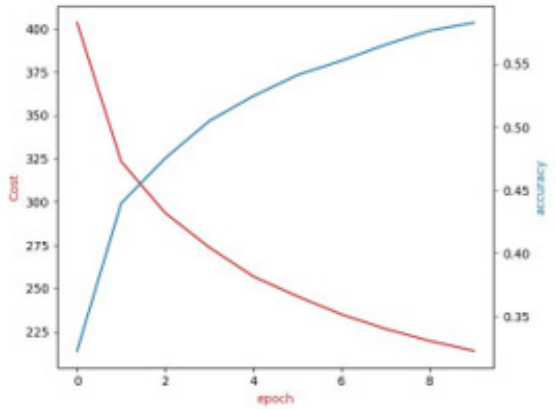


Fig. 2. Accuracy/cost per epoch chart for best network configuration (5x5, 11x11) with 32 channels in each layer on CIFAR-10 dataset.

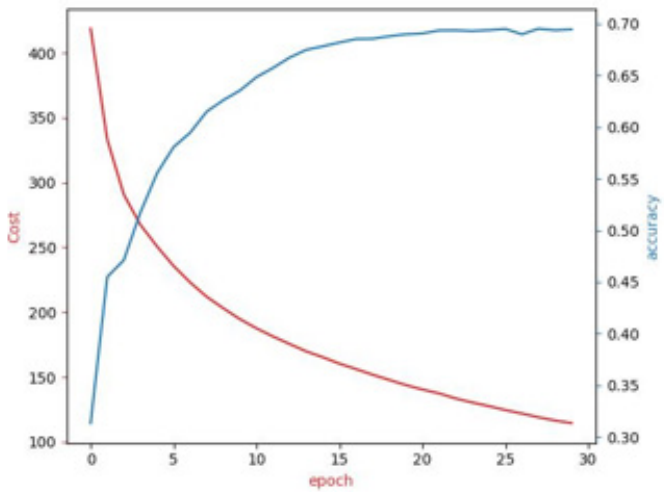


Fig. 3. Accuracy/cost per epoch chart for best network configuration (3x3, 9x9) with 32 channels in each layer on CIFAR-10 dataset, trained for 30 epochs.

Configuration with 32 and 64 channels proves its applicability for complex datasets. It was possible to gain ~7 % improvement switching from 16,32 to 32,64 channels on CIFAR-10 dataset. Even though for MNIST it does not bring reasonable improvements.

Having simple models with only two layers easily allows us to view how kernel size in each layer impacts accuracy and inference time. But as we will see further 2-layered networks are not capable of successful classification of images even in resolution of 32x32 reaching their performance limit at accuracy around 60 %.

We trained models on original MNIST and on multiple scaled resolutions (see

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

Methodology-B). The aim is to see an influence of image size related to the same image complexity or number of features on the image.

For this dataset, convolutional neural networks with two layers achieve consistently high accuracy (97–98 %) regardless of kernel configurations (3x3 to 11x11 for both layers). This demonstrates that MNIST, being a simple dataset of handwritten digits, does not require intricate kernel size tuning for effective feature extraction. Many configurations yield satisfactory results. However, the computational time increases significantly:

With scale increase, the execution time increases proportionally.

Table 4. Best network configurations for MNIST dataset of different scales.

Dataset	Acc, %	Time, ms	Config
MNIST	97.99	0,106	7x7, 5x5
MNIST_32x32	98.63	0,390	9x9, 7x7
MNIST_64x64	98.52	0,458	9x9, 3x3
MNIST_128x128	97.86	0,530	3x3, 9x9
MNIST_256x256	97.48	0,911	9x9, 11x11

MNIST, being a dataset with simple features doesn't require many channels for successful processing. Scaling image does not influence accuracy, meaning a shallow network is capable of processing big image successfully but due to small number of layers it requires kernels of bigger size to successfully capture information.

Rich features tend to be too complex for shallow networks and thus result in significantly lower accuracy (Table 5).

Table 5. Best network configurations for CIFAR-10 dataset of different scales

Dataset	Acc, %	Time, ms	Config
CIFAR10	63,15	0,091	3x3, 5x5
CIFAR10_64x64	61,25	0,508	3x3, 7x7
CIFAR10_128x128	56,17	0,642	3x3, 5x5
CIFAR10_256x256	43,04	3,728	7x7, 5x5

Across 32–256px (Table 5), accuracy decreases with scale for 2-layer CNNs; small first-layer kernels are best at low resolutions, while larger second-layer kernels help at higher resolutions, but capacity limits dominate.

Results on CIFAR-10 datasets push for further architecture analysis. We conducted several experiments changing basic convolution layer configuration in attempt to find more efficient layer configuration in shallow networks.

Our intuition is that scaled images contain significant spatial redundancy, allowing both dilated and strided convolutions to perform better. Their inherent trade-offs can be compensated for by the redundancy present in the scaled images.

In further analysis three types of convolution configurations were analyzed: standard convolution (stride 1, dilation 1 – meaning no dilation), dilated convolution (dilation factor

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

2), and strided convolution (stride 2). Classification accuracy and evaluation time per sample (in milliseconds) were measured to assess performance (see Figure 4).

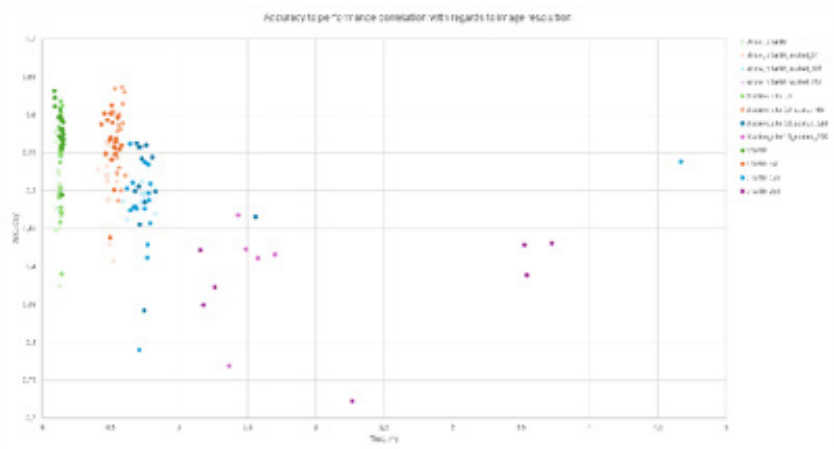


Fig. 4. Models' performance grouped by image resolution.

In our experiments we changed network architecture for strided convolutions by removing max pooling layers. We allow strided layers to do the downsampling instead of max pool layers (Springenberg et al., 2015).

At 32x32 resolution, all configurations performed similarly, with accuracy around 60 % and comparable evaluation times. The low spatial complexity of the images limited the benefits of alternative convolution strategies, and standard convolutions proved effective due to minimal spatial redundancy. At 64x64 resolution, dilated convolutions achieved slightly higher accuracy (up to 64%) compared to standard and strided configurations, which remained around 61 %. Evaluation time for dilated convolutions was similar to standard convolutions (~0.5 ms), showing that dilated convolutions efficiently utilized the increased spatial information.

For 128x128 resolution, dilated convolutions maintained competitive accuracy (~62 %) while reducing evaluation time to ~1.0 ms, compared to 1.8 ms for standard convolutions. Strided convolutions, on the other hand, showed reduced accuracy (~58 %), highlighting their inability to retain fine-grained spatial information. At 256x256 resolution, dilated convolutions significantly reduced evaluation time (1.6 ms vs. 3.5 ms for standard convolutions) while preserving accuracy (~60 %). Strided convolutions suffered the lowest accuracy (~45 %) due to aggressive downsampling, which caused significant loss of detail.

Dilated convolutions best exploit high-resolution redundancy, preserving accuracy with lower latency. However, strided convolutions trade off too much spatial detail unless combined with standard layers. Standard convolutions were effective and computationally efficient for low-resolution images but became less practical as image resolution increased due to overlapping receptive fields and redundant computations.

Our experiments indicate that a ~67 % accuracy ceiling exists for 2-layer networks on CIFAR-10 under the tested configurations. To explore the effect of depth, we extended the analysis to networks with up to 9 layers, using 3×3 kernels in all layers for consistency.

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

We observed that deeper networks generally improve performance, though the benefits diminish beyond a certain depth. Accuracy fluctuations can be mitigated through careful tuning of training epochs and learning rate schedules; however, simply increasing the number of epochs does not yield improvements unless paired with appropriate learning rate adjustments.

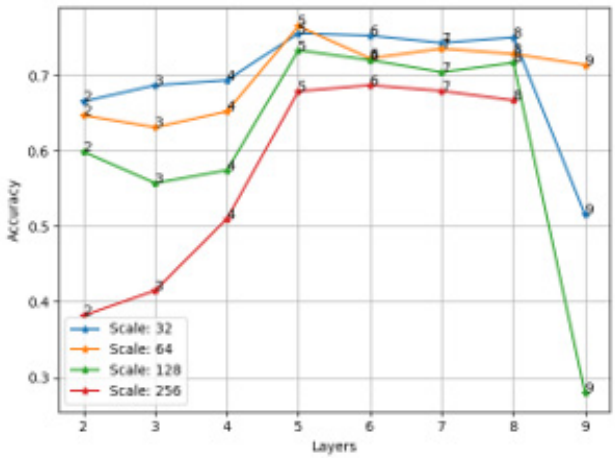


Fig. 5. Model accuracy per layer and dataset scale.

Note: Accuracy for the 9-layer network at 256-pixel resolution is not reported, as the model could not be successfully trained under the declared training configuration.

Figure 5 demonstrates that the highest accuracies across dataset scales were consistently achieved with networks between 5 and 8 layers. Networks deeper than this tend to underperform in the absence of data augmentation, suggesting that depth alone is not sufficient to overcome dataset limitations. Importantly, the accuracy patterns are similar across all explored image resolutions, leading us to conclude that optimal depth is determined more by dataset complexity than by input resolution.

Conclusion.

The experimental results demonstrate that CNN performance depends strongly on the interplay between model depth and convolutional strategy. Increasing depth from two to approximately eight layers consistently enhances classification accuracy on datasets such as CIFAR-10, while further deepening yields diminishing returns in the absence of data augmentation. These findings suggest that the optimal depth of shallow-to-moderate CNNs is primarily determined by dataset complexity rather than input resolution.

Among the convolutional strategies examined, dilated convolutions provide a clear computational advantage at higher image resolutions, enabling faster inference without loss of accuracy due to their ability to expand the receptive field efficiently. In contrast, strided convolutions exhibit persistent degradation in accuracy, indicating that aggressive spatial downsampling compromises fine-grained feature representation.

Overall, these findings indicate that CNN design should be viewed as a joint

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

optimization problem that integrates depth, kernel configuration, and convolution type. Depth scaling combined with selective use of dilated convolutions offers a promising direction for constructing lightweight yet accurate architectures capable of real-time performance on embedded and edge platforms.

Future work should extend this analysis to more complex and widely used datasets such as ImageNet, in order to determine optimal CNN depth and generalization limits across diverse visual domains. Further research should also focus on developing adaptive or self-configuring architectures capable of dynamically adjusting kernel parameters and convolution types according to dataset complexity and input resolution.

REFERENCES

Chen L.C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence. — Vol. 40. — No. 4. Pp. 834–848. DOI:10.1109/TPAMI.2017.2699184. [In Eng.]

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2021). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale, in: Proceedings of the 9th International Conference on Learning Representations, ICLR. — Vienna, Austria. [In Eng.]

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. (2009). ImageNet: A Large-Scale Hierarchical Image Database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. — Miami, FL, USA. Pp. 248–255. DOI:10.1109/CVPR.2009.5206848. [In Eng.]

Duggal J., El-Sharkawy M. (2020). High Performance SqueezeNext for CIFAR-10, in: Proceedings of the IEEE International Conference on Consumer Electronics. IEEE. — Las Vegas, NV, USA. Pp. 1–2. DOI:10.1109/ICCE46568.2020.9058217. [In Eng.]

He K., Zhang X., Ren S., Sun J. (2016). Deep residual learning for image recognition, in Proceedings of IEEE Conference on Computer. Vision and. Pattern Recognition, CVPR. — Las Vegas, NV, USA. Pp. 770–778. DOI: 10.1109/CVPR.2016.90. [In Eng.]

He K., Zhang X., Ren S., Sun J. (2016). Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), IEEE. — Las Vegas, NV, USA. Pp. 770–778. DOI:10.1109/CVPR.2016.90. [In Eng.]

Hozak Y., Paliy S. (2024). Modern Small Networks for Image Classification: Feature Analysis // Collection of Scientific Papers "Management of Development of Complex Systems. — Vol. 60. — Kyiv National University of Construction and Architecture (KNUCA). — Kyiv, Ukraine, DOI: 10.32347/2412-9933.2024.60.221-229. [In Eng.]

Isong N. (2025). Building Efficient Lightweight CNN Models, arXiv preprint arXiv:2501.15547. Available at: https://arxiv.org/abs/2501.15547. [In Eng.]

Jordan K. (2024). 94% on CIFAR-10 in 3.29 seconds on a single GPU. URL: https://arxiv.org/pdf/2404.00498. [In Eng.]

Krizhevsky A. (2009). Learning Multiple Layers of Features from Tiny Images, Technical Report, University of Toronto, Toronto, ON. — Canada. [In Eng.]

Krizhevsky A., Sutskever I., Hinton G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25 (NIPS 2012), Curran Associates, Inc. — Lake Tahoe, NV, USA, Pp. 1097–1105. [In Eng.]

LeCun Y, Bottou L., Bengio L., Haffner P. (1998). Gradient-based Learning Applied to Document Recognition, Proceedings of the IEEE. — Vol. 86. — No. 11. Pp. 2278–2324. DOI:10.1109/5.726791. [In Eng.]

LeCun Y., Cortes C., Burges C.J.C. (1998). The MNIST Database of Handwritten Digits, Technical Report. — New York University. Available at: http://yann.lecun.com/exdb/mnist/[In Eng.]

Mehta S., Rastegari M., Caspi A., Shapiro L., Hajishirzi H. (2018). ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV 2018), Springer. — Munich, Germany. Pp. 552–568. DOI:10.1007/978-3-030-01234-233. [In Eng.]

Mehta S., Rastegari M., Shapiro L., Hajishirzi H. (2019). ESPNetv2: A Light-weight, Power Efficient, and General Purpose Convolutional Neural Network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), IEEE. — Long Beach, CA, USA. Pp. 9190–9200. DOI:10.1109/CVPR.2019.00941.

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

[In Eng.]

Springenberg J., Dosovitskiy A., Brox T., Riedmiller M. (2015). Striving for Simplicity: The All Convolutional Net, in: Proceedings of the 3rd International Conference on Learning Representations. — San Diego, CA, USA. [In Eng.]

Simonyan K., Zisserman A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). — San Diego, CA, USA. [In Eng.]

Sinha P., Psaromiligkos I., Zilic Z. (2025). A Structurally Regularized CNN Architecture via Adaptive Subband Decomposition, in: IEEE Transactions on Neural Networks and Learning Systems. — Vol. 36. — No. 7. Pp. 12937–12951, DOI: 10.1109/TNNLS.2024.3486181. [In Eng.]

Tan M., Le Q. V. (2021). EfficientNetV2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning, PMLR ’139, MLR Press. — San Diego, CA, USA. DOI: 139:10096-10106. [In Eng.]

Yu F., Koltun V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions, in: Proceedings of the 4th International Conference on Learning Representations. — San Juan, Puerto Rico. [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.2. Number 26 (2026). Pp. 153–166
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.010
IRSTI / ҒТАХР / МРНТИ / 28.23.02

PREDICTIVE MAINTENANCE OF HEAVY-DUTY TRUCKS USING EXPLAINABLE MACHINE LEARNING

S.A. Yesniyazova*, S.T. Kaimov
International Information Technology University, Almaty, Kazakhstan.
E-mail: sabinaesn@gmail.com

Sabina A. Yesniyazova — Master's student in Software Engineering, International Information Technology University, Almaty, Kazakhstan
E-mail: sabinaesn@gmail.com, https://orcid.org/0009-0008-0900-512X;
Suleimen T. Kaimov — PhD in Mechanics, associate professor, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-4255-6926.

© S.A. Yesniyazova, S.T. Kaimov

Abstract. Predictive maintenance for heavy-duty trucks is difficult because public failure datasets are rare and accurate machine-learning models are often hard to interpret. This study evaluates a Random Forest classifier with SHAP explanations on the SCANIA Component X dataset, a real-world multivariate time-series dataset collected from more than 33,000 trucks. The objective is to predict failure risk from the last available vehicle readout and to identify the operational variables that drive model decisions. The data were aggregated to one record per vehicle, missing values were imputed by training-set medians, and class imbalance was addressed by cost-sensitive learning and threshold optimization. On the validation set, the model achieved 0.853 accuracy, 0.447 recall, 0.315 precision and 0.370 F1-score for the failure class. SHAP analysis showed that a small set of histogram and counter variables carried most of the predictive signal.

Keywords: predictive maintenance, explainable artificial intelligence, SHAP, Random Forest, heavy-duty trucks, SCANIA, class imbalance

For citation: S.A. Yesniyazova, S.T. Kaimov (2026). Predictive maintenance of heavy-duty trucks using explainable machine learning // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 153–166. https://doi.org/10.54309/IJICT.2026.26.2.010. (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

ТҮСІНДІРІЛЕТІН МАШИНАЛЫҚ ОҚЫТУДЫ ҚОЛДАНА ОТЫРЫП АУЫР ЖҮК КӨЛІКТЕРІНЕ БОЛЖАМДЫ ТЕХНИКАЛЫҚ ҚЫЗМЕТ КӨРСЕТУ

С. Есниязова*, С. Каимов
Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан.
E-mail: sabinaesn@gmail.com

Есниязова Сабина — бағдарламалық қамтамасыз ету инженериясы магистрі, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
E-mail: sabinaesn@gmail.com, https://orcid.org/0009-0008-0900-512X;
Каимов Сулеймен — механика PhD, қауымдастырылған профессоры, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-4255-6926.

© С. Есниязова, С. Каимов

Аннотация. Ауыр жүк көліктеріне болжамды техникалық қызмет көрсету күрделі міндет болып табылады, өйткені ақаулар туралы ашық деректер жиындары сирек кездеседі, ал дәл машиналық оқыту модельдерін түсіндіру жиі қиын. Бұл зерттеу 33 000-нан астам жүк көлігінен жиналған нақты SCANIA Component X көпөлшемді уақыттық деректер жиынында Random Forest жіктеуішін және SHAP түсіндіру әдісін бағалайды. Зерттеудің мақсаты – әр көлік бойынша соңғы қолжетімді өлшем негізінде ақау қаупін болжау және модель шешіміне әсер ететін эксплуатациялық белгілерді анықтау. Деректер бір көлікке бір жазбаға дейін біріктірілді, жетіспейтін мәндер оқыту жиынының медианаларымен толтырылды, ал класстар теңгерімсіздігі класстарды салмақтау және шешім шегін оңтайландыру арқылы ескерілді. Валидациялық жиында модель дәлдігі 0,853, толықтығы 0,447, оң класс дәлдігі 0,315 және F1-көрсеткіші 0,370 болды. SHAP талдауы негізгі болжамдық ақпарат аз санды гистограммалық және санауыш белгілерінде шоғырланғанын көрсетті.

Түйінді сөздер: болжамды техникалық қызмет, түсіндірілетін жасанды интеллект, SHAP, Random Forest, ауыр жүк көліктері, SCANIA, класстар теңгерімсіздігі

Дәйексөздер үшін: С.А. Есниязова, С.Т. Каимов (2026). Түсіндірілетін машиналық оқытуды қолдана отырып ауыр жүк көліктеріне болжамды техникалық қызмет көрсету // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 153–166. https://doi.org/10.54309/IJICT.2026.26.2.010. (Ағыл. тіл.).

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

ПРЕДИКТИВНОЕ ТЕХНИЧЕСКОЕ ОБСЛУЖИВАНИЕ ТЯЖЁЛЫХ ГРУЗОВИКОВ С ИСПОЛЬЗОВАНИЕМ ОБЪЯСНИМОГО МАШИННОГО ОБУЧЕНИЯ

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

С. Есниязова*, С. Каимов
Международный Университет Информационных Технологий, Алматы, Казахстан.
E-mail: sabinaesn@gmail.com

Есниязова Сабина — студент магистратуры программной инженерии, Международный Университет Информационных Технологий, Алматы, Казахстан
E-mail: sabinaesn@gmail.com, https://orcid.org/0009-0008-0900-512X;
Каимов Сулеймен — PhD механики, ассоциированный профессор, Международный Университет Информационных Технологий, Алматы, Казахстан
https://orcid.org/0000-0002-4255-6926.

© С. Есниязова, С. Каимов

Аннотация. Предиктивное техническое обслуживание тяжёлых грузовиков является сложной задачей, поскольку открытые наборы данных об отказах встречаются редко, а точные модели машинного обучения часто трудно интерпретировать. В работе оценивается классификатор Random Forest с объяснениями SHAP на наборе данных SCANIA Component X, собранном более чем с 33 000 грузовиков. Цель исследования состоит в прогнозировании риска отказа по последнему доступному считыванию по каждому транспортному средству и в выявлении эксплуатационных признаков, влияющих на решения модели. Данные были агрегированы до одной записи на автомобиль, пропущенные значения заменялись медианами обучающей выборки, а дисбаланс классов учитывался с помощью взвешивания классов и оптимизации порога. На валидационной выборке модель достигла точности 0,853, полноты 0,447, точности положительного класса 0,315 и F1-меры 0,370 для класса отказа. Анализ SHAP показал, что основная прогностическая информация сосредоточена в небольшой группе гистограммных и счётчиковых признаков.

Ключевые слова: предиктивное обслуживание, объяснимый искусственный интеллект, SHAP, Random Forest, тяжёлые грузовики, SCANIA, дисбаланс классов

Для цитирования: С.А. Есниязова, С.Т. Каимов (2026). Предиктивное техническое обслуживание тяжёлых грузовиков с использованием объяснимого машинного обучения // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 153–166. https://doi.org/10.54309/IJICT.2026.26.2.010. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Modern industrial transport depends on complex mechanical and electronic systems. Unexpected failures in such systems create downtime, safety risks and repair costs, while scheduled preventive maintenance can lead to unnecessary interventions. Predictive maintenance addresses this problem by using sensor and service data to estimate the probability of future failure and to support maintenance before breakdowns occur.

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

Systematic reviews show that machine-learning methods are now widely used in predictive maintenance because they can learn non-linear relationships in operational data and can be updated when new fleet information becomes available (Carvalho et al., 2019; Zonta et al., 2020).

The object of this study is the SCANIA Component X dataset, a real-world dataset for an anonymized engine component from heavy-duty trucks. The subject of the study is failure prediction from aggregated sensor-derived features and the interpretation of model decisions. The aim is to evaluate whether a relatively transparent tree-based baseline can detect failures under strong class imbalance and whether SHAP explanations can identify the most influential operational variables. The main objectives are to prepare a binary learning dataset, train a Random Forest model, optimize the decision threshold for the failure class and interpret the model at both global and local levels.

The topic is relevant because industrial failure data are rarely released publicly. Many prognostic studies rely on simulated benchmark data, such as the C-MAPSS turbofan dataset, which is valuable for method development but does not fully reproduce the noise, irregular sampling and repair-label uncertainty of operating fleets (Saxena et al., 2008). The SCANIA Component X dataset is therefore important because it contains real fleet telemetry, repair information and vehicle specifications for more than 33,000 trucks (Kharazian et al., 2025). Similar work on heavy-duty vehicle batteries has shown that sparse and irregular workshop-collected data create practical modelling challenges that differ from controlled simulation settings (Voronov et al., 2020).

The research hypothesis is that a cost-sensitive Random Forest, although not expected to produce perfect failure detection, can identify a useful subset of at-risk trucks and can be made interpretable with SHAP. This hypothesis is tested on the validation data using precision, recall and F1-score for the failure class rather than accuracy alone. The contribution of the article is a compact and reproducible baseline for the SCANIA Component X data, together with an explanation of why some failures are detected and why others are missed.

The scientific novelty lies in using the recently published SCANIA Component X data as a reproducible case for balancing prediction quality and explanation quality. Many industrial studies report high predictive scores on proprietary datasets, but their data and preprocessing decisions cannot be independently inspected. Here, the selected baseline is deliberately conservative: it does not use the hidden test labels, does not create synthetic failure records and does not use the full time series. This design makes the limitations visible. It also clarifies which part of the prediction task can be solved from one final readout and which part probably requires richer temporal information.

Materials and research methods.

Research material

The research material is the SCANIA Component X dataset published in Scientific Data and made publicly available through the Swedish National Data Service (Kharazian et al., 2025). The dataset combines three groups of information: operational readouts from electronic control units, workshop repair records and anonymized vehicle specifications.

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

The component is called Component X for confidentiality reasons, but the data preserve the temporal and operational patterns needed for predictive maintenance modelling.

The operational file contains a vehicle identifier, a relative time step and 105 anonymized numerical features. Six variables are stored as histograms that represent sensor distributions across operating ranges, and eight variables are cumulative numerical counters. The source paper reports that the counters are positively correlated, which suggests that they represent related aspects of component use. Missing values are rare; the highest missing proportion for a single feature is below one percent, so simple imputation is sufficient for a baseline model.

Table 1 – Class distribution across dataset partitions

Partition	Class	Count	Percentage (%)
Training	0 — Healthy	21,278	90.4
Training	1 — Failed	2,272	9.6
Validation	0 — More than 48 steps to failure	4,910	97.3
Validation	1 — 24–48 steps to failure	14	0.3
Validation	2 — 12–24 steps to failure	16	0.3
Validation	3 — 6–12 steps to failure	30	0.6
Validation	4 — Fewer than 6 steps to failure	76	1.5
Test	0 — More than 48 steps to failure	4,903	97.2
Test	1 — 24–48 steps to failure	26	0.5
Test	2 — 12–24 steps to failure	15	0.3
Test	3 — 6–12 steps to failure	41	0.8
Test	4 — Fewer than 6 steps to failure	60	1.2

Source: compiled from (Kharazian et al., 2025).

Research stages and preprocessing

The study followed five stages. First, the dataset structure and class labels were inspected. Second, the raw time-series records were reduced to one final readout per vehicle to align the training representation with the validation and test structure. This choice removes sequential information, but it provides an interpretable baseline and avoids mixing different temporal formats. Third, missing values were replaced by the median of each feature calculated on the training set only. The same medians were then applied to validation data to prevent data leakage.

Fourth, the binary classification target was formed. In the training data, class 0 represents vehicles without a recorded Component X failure and class 1 represents vehicles with a failure. In the validation data, the original five-class time-to-failure labels were mapped to a binary target by treating classes 1–4 as failures. Fifth, the class imbalance was handled by cost-sensitive learning. The final Random Forest used class weights of {0:1, 1:5}, so that errors on failed cases received a higher penalty. Synthetic oversampling was

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

not used because the goal was to preserve the original industrial data distribution; SMOTE and related methods can be useful but may create artificial samples that are hard to interpret in anonymized sensor spaces (Chawla et al., 2002).

Model and evaluation methods

Random Forest was selected as the main classifier because it performs well on high-dimensional tabular data, is robust to noisy features and does not require feature scaling. A Random Forest combines many decision trees grown on bootstrap samples and averages their decisions by majority voting, which reduces the variance of individual trees (Breiman, 2001). The model was implemented in Python using scikit-learn (Pedregosa et al., 2011).

Because failures are rare, accuracy was not used as the main criterion. A classifier that predicts every vehicle as healthy would obtain high accuracy while detecting no failures. Therefore, the evaluation focused on precision, recall and F1-score for the failure class. Precision measures the proportion of predicted failures that are true failures, recall measures the proportion of actual failures detected, and F1-score is the harmonic mean of precision and recall. The decision threshold was selected from the precision–recall curve on the validation set.

Table 2 – Hyperparameters of the final Random Forest model

Parameter	Value	Rationale
n_estimators	300	Stability of ensemble predictions
max_depth	None	Unrestricted trees gave the best F1-score
min_samples_leaf	2	Allows finer splits while reducing single-sample leaves
class_weight	{0:1, 1:5}	Cost-sensitive learning for rare failures
random_state	42	Reproducibility
Decision threshold	0.199	Selected from the precision–recall curve

No feature scaling was applied because decision trees split variables by thresholds and are not sensitive to linear rescaling in the same way as distance-based or gradient-based models. Vehicle identifiers and time-step identifiers were removed from the feature matrix to avoid learning direct identifiers rather than operational patterns. The validation data were not used for training the trees; they were used only for threshold selection and evaluation. This separation is important because threshold tuning on the same observations used to fit the model can lead to over-optimistic estimates of performance.

The research questions were formulated as follows. RQ1 asks whether a Random Forest classifier can identify a practically useful share of Component X failures under the original imbalance of the dataset. RQ2 asks which anonymized operational variables have the strongest effect on the model output. RQ3 asks whether local explanations can clarify false-negative cases that are particularly important for maintenance practice. The corresponding hypothesis is that interpretable tree-based learning will improve failure detection over a majority-class baseline and will expose a limited set of variables that dominate the predictions.

Explainability method

SHAP was used to explain the trained model. SHAP is based on Shapley values from

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

cooperative game theory and estimates how much each feature contributes to a prediction relative to the model's average output (Lundberg et al., 2017). It is appropriate for this study because the same framework supports both global interpretation, where features are ranked by mean absolute contribution, and local interpretation, where one vehicle-level decision can be decomposed into positive and negative feature effects.

For tree-based models, the TreeSHAP algorithm computes exact attributions efficiently by using the internal structure of decision trees (Lundberg et al., 2020). In this study, SHAP was calculated on a stratified validation sample of 1,000 vehicles while preserving the original class distribution. The analysis used a summary plot, mean absolute SHAP ranking, dependence plots for the leading variables and a local waterfall plot for a missed failure case.

Results and discussion.

The predictive maintenance literature provides the methodological background for this study. (Carvalho et al., 2019) reviewed machine-learning methods for predictive maintenance and showed that tree-based methods, including Random Forest, are common because they combine strong tabular performance with practical robustness (Zonta et al., 2020). Placed predictive maintenance in the Industry 4.0 context and emphasized that data-driven maintenance depends not only on algorithms but also on the availability and quality of industrial data. These reviews justify the use of a supervised learning baseline but also highlight the scarcity of public real-world datasets.

Ensemble methods are especially relevant to industrial data (Breiman, 2001). Introduced Random Forest as an ensemble of randomized decision trees. Boosting methods provide an alternative family: (Friedman, 2001) formulated gradient boosting as additive function approximation (Chen et al., 2016) introduced XGBoost as a scalable implementation that became widely used for structured data. These methods are strong candidates for predictive maintenance, but they often require additional interpretation tools when used in operational decision support.

Deep learning is important when temporal dependencies are central (Hochreiter et al., 1997) introduced Long Short-Term Memory networks, which are designed to retain information over sequences (Khan et al., 2018). Reviewed deep learning applications in system health management and showed that neural architectures can model degradation patterns when sufficient sequence data are available. More recent studies, such as (Li et al., 2022; Metta et al., 2024), demonstrate the relevance of CNN-LSTM, attention and pseudo-labeling approaches for remaining-useful-life and truck-maintenance tasks. In the present article, however, a tree-based baseline is intentionally used because the training data are reduced to the last readout and because interpretability is a central requirement.

Explainability has become a major condition for applying machine learning in technical systems (Guidotti et al., 2018). Reviewed methods for explaining black-box models and argued that interpretability is necessary when models support decisions with practical consequences (Lundberg et al., 2017). Introduced SHAP as a unified feature attribution framework (Lundberg et al., 2020) extended the approach for tree models. Predictive-maintenance studies have adopted similar tools in different domains (Kargar-

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

Sharif-Abad et al., 2024): used SHAP with survival-analysis models for SCANIA maintenance data (Brito et al., 2022) applied explainable AI to rotating machinery fault detection and diagnosis, and (Astolfi et al., 2023) used explainable AI for wind turbine condition monitoring. The gap addressed by this article is the combination of a simple failure-classification baseline with SHAP explanations on the publicly released SCANIA Component X dataset.

The reviewed literature also shows a methodological trade-off. Sequence models can use richer temporal information, but they are usually harder to explain and require careful construction of time windows. Tree ensembles can lose temporal detail when the data are aggregated, but they are easier to train on tabular representations and can be interpreted with exact tree-based SHAP values. This study therefore does not claim that Random Forest is the best possible model for the dataset. Instead, it treats Random Forest as a transparent benchmark against which later survival, boosting or recurrent models can be compared.

Classification performance

The final model was evaluated on the internal validation set after mapping the failure-related validation classes to a binary target. The binary validation set contained 4,710 vehicles: 4,256 healthy cases and 454 failure cases. Using the optimized threshold of 0.199, the model achieved an overall accuracy of 0.853. For the failure class, recall was 0.447, precision was 0.315 and F1-score was 0.370. These values are modest in absolute terms, but they are more useful than a majority-class baseline that would label all vehicles as healthy and detect no failures.

The confusion matrix shows the practical trade-off. The model detected 203 of 454 failed vehicles, missed 251 failures and incorrectly flagged 441 healthy vehicles. This means that about 45% of failures were detected, but false alarms remained frequent. In a predictive maintenance setting, the acceptability of this trade-off depends on the relative cost of inspection and unexpected breakdown. If missed failures are more expensive than unnecessary inspections, a lower threshold can be justified; if inspection capacity is limited, a higher threshold may be preferable.

Table 3 – Classification performance on the internal validation set

Metric	Healthy class	Failure class	Macro average	Weighted average
Precision	0.939	0.315	0.627	0.879
Recall	0.900	0.447	0.674	0.853
F1-score	0.919	0.370	0.644	0.866
Support	4,256	454	4,710	4,710
Accuracy	-	-	-	0.853

SHAP feature importance and interactions

The SHAP analysis indicates that the model relies on a small subset of variables. The strongest features include 167_1, 272_0, 158_9, 397_6 and 167_2. Their mean absolute SHAP values are higher than most remaining features, which suggests that useful predictive information is concentrated rather than evenly distributed across all 105 features.

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

This finding is consistent with the practical observation that only some operating regimes may be informative for a specific component.

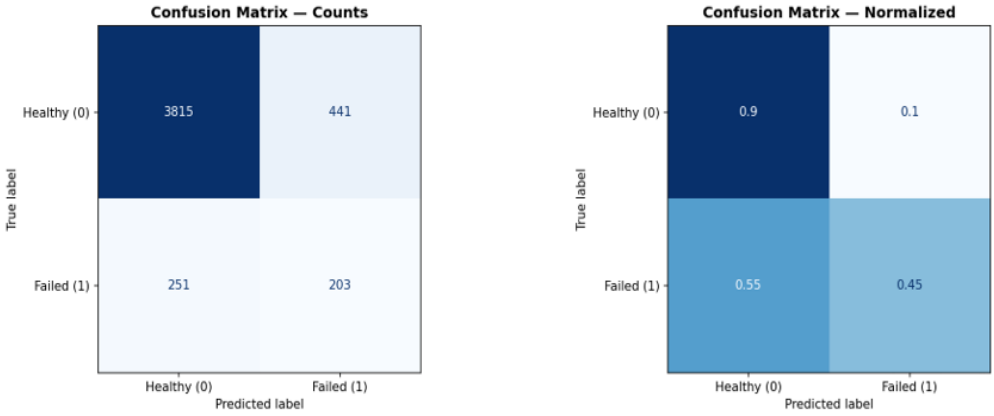


Fig. 1. Confusion matrix on the internal validation set

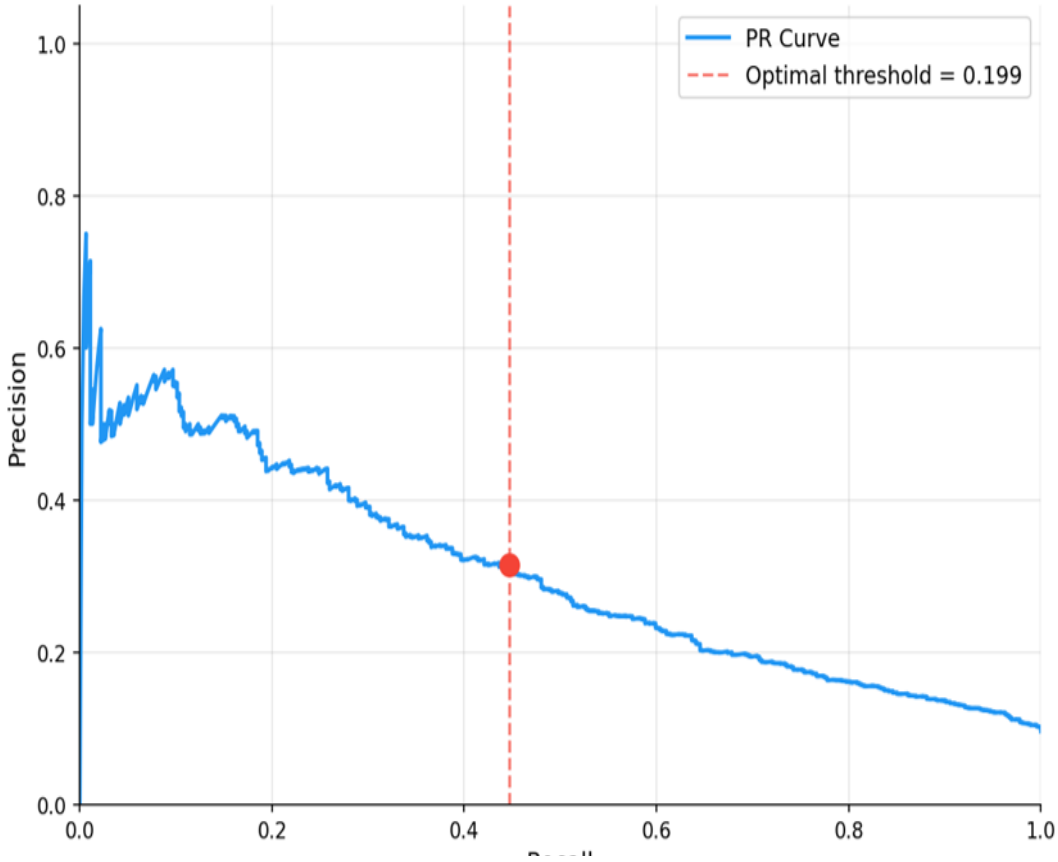


Fig. 2. Precision–recall curve for the failure class

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

The direction of the SHAP effects is also important. For many leading variables, low values tend to push the model output toward the failure class, while higher values tend to push it toward the healthy class. Since the variables are anonymized counters and histogram bins, this should not be interpreted as direct physical causality. A plausible explanation is that low accumulated values represent newer vehicles, vehicles with interrupted histories, counter resets or unusual operating profiles (Kharazian et al., 2025) note that software updates and post-processing can affect counter continuity, so low counter values may partly represent data-history effects rather than simple mechanical wear.

The dependence plots show a saturation pattern. For 167_1 and 272_0, SHAP contributions are positive at low raw values but decrease quickly and become almost flat after approximately 0.5×10^7. This indicates that the model mainly uses these features to distinguish vehicles with limited recorded operational history from vehicles with sufficient history to be treated as lower risk. The color coding of interacting variables suggests weak but consistent interactions between related histogram bins.

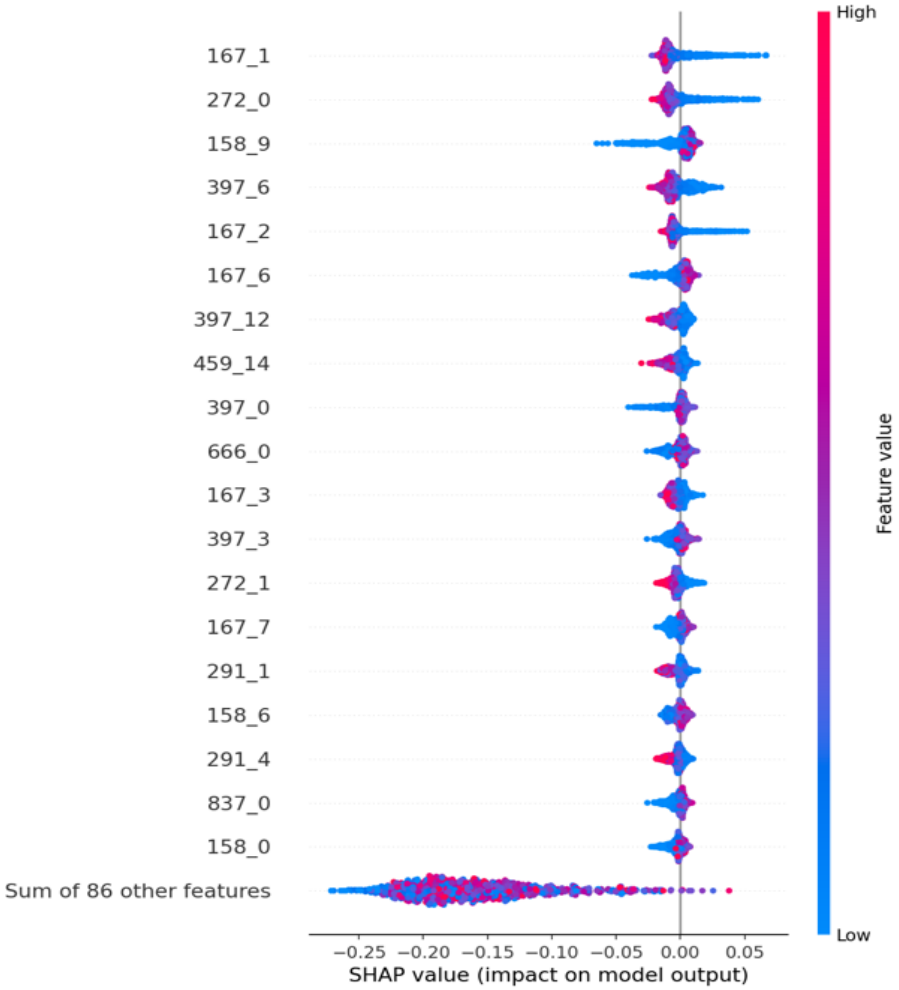


Fig. 3. SHAP summary plot for the leading features

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

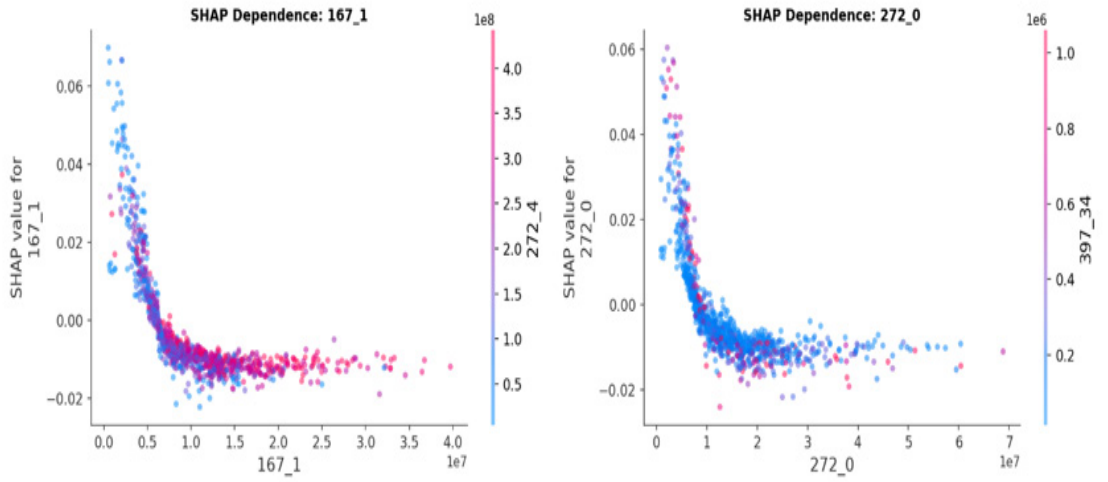


Fig. 4. SHAP dependence plots for the two leading features

Local explanation of a missed failure

The local explanation illustrates both the value and the limits of the model. The selected example is a failed vehicle whose predicted failure probability was 0.056, below the operating threshold of 0.199. Therefore, the case was classified as healthy even though it was a true failure. The waterfall plot shows that the prediction was pushed downward mainly by the combined effect of many small features, while only a few leading variables contributed weakly toward the failure class.

This case is useful because it explains why the error occurred. The values of important variables such as 167_1 and 272_0 fall in the flatter region of the dependence plots, where the model treats the vehicle as similar to many healthy vehicles. In other words, the available feature representation did not contain a strong signal for this failure. This supports the conclusion that further improvements may require either richer telemetry or models that use the full time series rather than only the last readout.

Limitations and practical implications

Several limitations should be considered when interpreting the results. First, the model uses only the last readout per vehicle. This makes the learning problem consistent across dataset splits, but it removes trends, sudden changes and degradation trajectories that could be informative. Second, the component and variables are anonymized, which protects industrial confidentiality but prevents direct engineering interpretation of individual features. Third, the threshold was optimized for F1-score, while real maintenance planning may require an economic objective that includes inspection cost, downtime cost and spare-part availability.

The validation results should therefore be interpreted as a decision-support baseline rather than a deployment-ready system. A practical implementation would need calibration, monitoring of false alarms, integration with workshop capacity and periodic retraining as the fleet changes. The model would also need to be evaluated on the held-out test set and,

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

ideally, on later operational data collected after the model was developed. Such external validation is necessary because failure distributions can shift when vehicles, routes, software versions or maintenance practices change.



Fig. 5. SHAP waterfall plot for a missed failure case

Despite these limitations, the analysis has practical value. The threshold can be moved along the precision–recall curve to match different operational priorities. For example, a fleet operator that wants to avoid roadside failures may accept more false alarms, while a workshop with limited inspection capacity may prefer a stricter threshold. The SHAP explanations can support this process by showing whether alerts are driven by stable patterns or by unusual feature combinations. In this role, explainability helps engineers decide when to trust a warning, when to request additional inspection and when to collect better signals for future models.

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

Conclusion.

This study evaluated an explainable machine-learning approach for predictive maintenance of heavy-duty trucks using the SCANIA Component X dataset. The aim, objectives and methods were implemented through a pipeline that aggregated each vehicle to its final readout, imputed missing values, trained a cost-sensitive Random Forest, optimized the decision threshold and interpreted the predictions with SHAP. The approach is simple enough to serve as a baseline but still reflects practical constraints of industrial data, including strong class imbalance and anonymized variables.

The main empirical result is that the model achieved 0.853 accuracy, 0.447 recall, 0.315 precision and 0.370 F1-score for the failure class on the internal validation set. The model is not highly accurate for failures, but it detects a meaningful subset of failures that a naive majority-class classifier would miss entirely. This is relevant for maintenance decision support, where even partial early warning can be useful if the inspection cost is lower than the cost of breakdown.

The main interpretive result is that SHAP revealed a concentrated feature structure and an inverse relationship between some counter or histogram values and predicted failure risk. Lower values of several influential variables increased failure probability, while higher values often reduced it. The dependence plots showed saturation, meaning that the model uses these variables primarily as threshold-like indicators. The local missed-failure explanation further showed that SHAP can identify why a specific case was not detected, but it cannot compensate for information missing from the feature set.

The practical implication is that the model should be used as decision support rather than as an automatic maintenance rule. The threshold can be adjusted depending on inspection capacity and the cost of missed failures. Future work should evaluate sequence-aware models, including LSTM-based and CNN-LSTM approaches, survival-analysis formulations and hybrid models that combine failure probability with time-to-failure estimation. Another important direction is validation with maintenance engineers, because model explanations must correspond to operational knowledge before deployment in safety-relevant transport systems.

REFERENCES

Astolfi, D., De Caro, F., & Vaccaro, A. (2023). Condition monitoring of wind turbine systems by explainable artificial intelligence techniques. Sensors. — Vol. 23. — No. 12. Pp. 5376. https://doi.org/10.3390/s23125376 [In Eng.].

Breiman, L. (2001). Random forests. Machine Learning. — Vol. 45. — No. 1. Pp. 5–32. https://doi.org/10.1023/A:1010933404324 [In Eng.].

Brito L.C., Susto G.A., Brito J.N., & Duarte M.A.V. (2022). An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery // Mechanical Systems and Signal Processing. — Vol. 163. — Article 108105. https://doi.org/10.1016/j.ymssp.2021.108105 [In Eng.].

Carvalho T.P., Soares F.A.A.M.N., Vita R., Francisco R.P., Basto J.P. & Alcalá S.G.S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance // Computers & Industrial Engineering. — Vol. 137. — Article 106024. https://doi.org/10.1016/j.cie.2019.106024 [In Eng.].

Chawla N.V., Bowyer K.W., Hall L.O. & Kegelmeyer W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. — Vol. 16. — Pp 321–357. https://doi.org/10.1613/jair.953 [In Eng.].

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pp. 785–794. Association for Computing

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

Machinery. https://doi.org/10.1145/2939672.2939785 [In Eng.].

Friedman J.H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, — Vol. 29. — No. 5. Pp. 1189–1232. https://doi.org/10.1214/aos/1013203451 [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. 93. https://doi.org/10.1145/3236009 [In Eng.].

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation. — Vol. 9. — No. 8. Pp. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 [In Eng.].

Kargar-Sharif-Abad, M., Kharazian, Z., Miliou, I., & Lindgren, T. (2024). SHAP-driven explainability in survival analysis for predictive maintenance applications. In Proceedings of HAII5.0: Embracing Human-Aware AI in Industry 5.0, Workshop at ECAI 2024, Santiago de Compostela, Spain. [In Eng.].

Khan, S., & Yairi, T. (2018). A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing. — Vol. 107. — Pp. 241–265. https://doi.org/10.1016/j.ymssp.2017.11.024 [In Eng.].

Kharazian, Z., Lindgren, T., Magnússon, S., Steinert, O. & Andersson Reyna, O. (2025). SCANIA Component X dataset: A real-world multivariate time series dataset for predictive maintenance // Scientific Data. — Vol. 12. — Pp. 493. https://doi.org/10.1038/s41597-025-04802-6 [In Eng.].

Li H., Wang Z., & Li Z. (2022). An enhanced CNN-LSTM remaining useful life prediction model for aircraft engine with attention mechanism // PeerJ Computer Science, 8, e1084. https://doi.org/10.7717/peerj-cs.1084 [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. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [In Eng.].

Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., & Lee S.I. (2020). From local explanations to global understanding with explainable AI for trees // Nature Machine Intelligence. — Vol. 2. — No. 1. Pp. 56–67. https://doi.org/10.1038/s42256-019-0138-9 [In Eng.].

Metta C., Gregnanin M., Papini A., Galfrè S.G., Fois A., Morandin F., Fantozzi M., & Parton M. (2024). Achieving predictive precision: Leveraging LSTM and pseudo labeling for Volvo's Discovery Challenge at ECML-PKDD 2024. arXiv:2409.13877. https://doi.org/10.48550/arXiv.2409.13877 [In Eng.].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. — Vol. 12. — Pp. 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html [In Eng.].

Saxena A., Goebel K., Simon D. & Eklund N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation // In Proceedings of the 2008 International Conference on Prognostics and Health Management Pp. 1–9. IEEE. https://doi.org/10.1109/PHM.2008.4711414 [In Eng.].

Voronov S., Frisk E. & Krysander M. (2020). Predictive maintenance of lead-acid batteries with sparse vehicle operational data // International Journal of Prognostics and Health Management. — Vol. 11. — No. 1. Pp. 1–18. https://doi.org/10.36001/ijphm.2020.v11i1.2608 [In Eng.].

Zonta, T., da Costa, C.A., da Rosa Righi, R., de Lima, M.J., da Trindade, E.S., & Li, G.P. (2020). Predictive maintenance in the Industry 4.0 // A systematic literature review. Computers & Industrial Engineering. — Vol. 150. — Article 106889. https://doi.org/10.1016/j.cie.2020.106889 [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.2. Number 26 (2026). Pp. 167–184
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.011

DATA COMPRESSION ALGORITHM BASED ON WAVELET TRANSFORMER; ANALYSIS AND IMPLEMENTATION IN MATLAB

T. Imanbekova[1], Zh. Ibrayeva[1], G. Jakanova[1]*, G. Askanbay[2]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty, Kazakhstan.
E-mail: gjakanova@iitu.edu.kz

Tokhtabubi Imanbekova — Candidate of Technical Sciences, associate professor of the department «Radio engineering, electronics and telecommunications», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-3698-3956;
Zhanar Ibrayeva — Phd, assistant professor professor of the department «Radio engineering, electronics and telecommunications», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0003-3196-696X;
Gaukhar Jakanova — senior-lecturer of the department «Radio engineering, electronics and telecommunications», International Information Technology University, Almaty, Kazakhstan
E-mail: gjakanova@iitu.edu.kz, https://orcid.org/0009-0004-7890-3573;
Gaukhar Askanbay — doctoral student in «Electric Power Engineering», Institute of Energy and Green Technologies, University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty, Kazakhstan
https://orcid.org/0009-0000-3917-9608.

© T.D. Imanbekova, Zh.B. Ibrayeva, G.T. Jakanova, G.T. Askanbay

Abstract. As digital technologies continue to advance, the volume of shared information is growing at an incredible speed. This trend causes a massive demand for bigger storage systems and improved bandwidth for communication channels. In the past, when internet connections was slow, transferring even basic programs or files took a very long time. This is why efficient data compression methods became so essential. These techniques led to the development of archivers-tools that considerably decline file sizes withaut damaging the actual content. In this work, we took a close look at modern compression algorithms based on wavelet transforms. We did not just compare existing methods; we also developed our own software solution using the Wavelet Toolbox within

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

the MATLAB environment. Through various tests, we checked how much different types of data could be reduced and measured the specific compression ratios. The Wavelet Toolbox provides a wide range of functions for processing signals and images. The compression works by removing less important wavelet coefficients and then restoring the signal. The study analyzed data compression algorithms based on the discrete wavelet transform. The audio and image compression processes were implemented in the MATLAB environment, and the influence of different wavelet types and division levels on the compression efficiency was assessed. The results of the study showed the efficiency of Dmey and DB20 wavelets for audio signals, as well as Haar and Bior1.1 wavelets for images. The proposed approach allows optimizing the processes of data storage and transmission without reducing their quality. The efficiency of the algorithm is confirmed by high results in compression and restoration of various types of signals.

Keywords: wavelet transformer, data compression, audio signal, image processing, compression ratio, recovery algorithm, decomposition degree

For citation: T.D. Imanbekova, Zh.B. Ibrayeva, G.T. Jakanova, G.T. Askanbay (2026). Data compression algorithm based on wavelet transformer; analysis and implementation in MATLAB // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 167–184. https://doi.org/10.54309/IJICT.2026.26.2.011. (In Kaz.).

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

МӘЛІМЕТТЕРДІ ВЕЙВЛЕТ-ТҮРЛЕНДІРГІШТІҢ НЕГІЗІНДЕ ҚЫСУ АЛГОРИТМІ; MATLAB ОРТАСЫНДА ТАЛДАУ ЖӘНЕ ІСКЕ АСЫРУ

Т.Д. Иманбекова[1], Ж.Б. Ибраева[1], Г.Т. Джаканова[1]*, Г.Т. Асқанбай[2]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Ғ.Даукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан.
E-mail: gjakanova@iitu.edu.kz

Иманбекова Тохтабуби Джумадиловна — т.ғ.к., «Радиотехника, электроника және телекоммуникациялар» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-3698-3956;
Ибраева Жанар Базарбековна — Phd, «Радиотехника, электроника және телекоммуникациялар» кафедрасының ассистент-профессоры, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-3196-696;
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;

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

Асқанбай Гаухар Тілеубердіқызы — «Электроэнергетика» мамандығы бойынша докторант, Жасыл технологиялар және энергетика институты, Ғ. Даукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан
https://orcid.org/0009-0000-3917-9608.

© Т.Д. Иманбекова, Ж.Б. Ибраева, Г.Т. Джаканова, Г.Т. Асқанбай

Аннотация. Сандық технологияның дамуымен берілетін ақпарат көлемі өте тез өсуде. Бұл бізді құрылғы жадын кеңейту және байланыс арнасының өткізу қабілетін жақсарту үшін жаңа шешімдер іздеуге мәжбүр етеді. Бұрын интернет байланысы әлсіз болған кезде, тіпті қарапайым бағдарламаларды немесе файлдарды жіберу де ұзақ уақытты алды. Дәл осы кезде тиімді деректерді сығу әдістері сұранысқа ие болды, бұл файл өлшемін оның мазмұнына зиян келтірмей азайтатын таныс мұрағатшылардың дамуына әкелді. Бұл мақалада біз толқындық түрлендірулерге негізделген заманауи сығу алгоритмдерін егжей-тегжейлі қарастырдық. Біз тек қолданыстағы тәсілдерді салыстырып қана қоймай, сонымен қатар Wavelet Toolbox пакетін пайдаланып MATLAB-та өз бағдарламалық шешімімізді жасадық. Біздің сынақтарымызда біз әртүрлі деректер түрлерін қаншалықты азайтуға болатынын және нақты сығу жылдамдықтарына қол жеткізілгенін тексердік. Wavelet Toolbox құралдар жинағы сигналдар мен кескіндерді икемді түрде басқаруға мүмкіндік береді. Мұнда сығу процесі маңызды емес толқындық коэффициенттерді алып тастау арқылы жүзеге асырылады, содан кейін сигнал қалпына келтіріледі. Зерттеу барысында дискретті вейвлет-түрлендіргішке негізделген мәліметтерді қысу алгоритмдері талданды. MATLAB ортасында аудиосигналдар мен кескіндерді қысу процестері жүзеге асырылып, әр түрлі вейвлет түрлері мен бөліну деңгейлерінің қысу тиімділігіне әсері бағаланды. Зерттеу нәтижелері Dmey және DB20 вейвлеттерінің аудиосигналдар үшін, ал Haar және Bior1.1 вейвлеттерінің кескіндер үшін тиімділігін көрсетті. Ұсынылған тәсіл мәліметтердің сапасын төмендетпей, оларды сақтау мен тасымалдау үдерістерін оңтайландыруға мүмкіндік береді. Алгоритмнің тиімділігі әртүрлі типтегі сигналдарды қысу және қалпына келтіру кезінде жоғары нәтижелермен дәлелденді.

Түйінді сөздер: вейвлет-түрлендіргіш, мәліметтерді қысу, аудиосигнал, кескінді өңдеу, қысу коэффициенті, қалпына келтіру алгоритмы, бөліну деңгейі

Дәйексөздер үшін: Т.Д. Иманбекова, Ж.Б. Ибраева, Г.Т. Джаканова, Г.Т. Асқанбай (2026). Мәліметтерді вейвлет-түрлендіргіштің негізінде қысу алгоритмі; MATLAB ортасында талдау және іске асыру // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 26. Б. 167–184. https://doi.org/10.54309/IJICT.2026.26.2.011. (Қаз. 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

АЛГОРИТМ СЖАТИЯ ДАННЫХ НА ОСНОВЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАТЕЛЯ: АНАЛИЗ И РЕАЛИЗАЦИЯ В MATLAB

Т.Д. Иманбекова[1], Ж.Б. Ибраева[1], Г.Т. Джаканова[1]*, Г.Т. Асқанбай[2]
[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Алматинский университет энергетики и связи имени Г. Даукеева, Алматы, Казахстан.
E-mail: gjakanova@iitu.edu.kz

Иманбекова Тохтабуби Джумадиловна — кандидат технических наук, ассоциированный профессор, кафедра «Радиотехника, электроника и телекоммуникации», Алматы, Казахстан
https://orcid.org/0000-0002-3698-3956;
Ибраева Жанар Базарбековна — Ph.D., ассистент-профессор, кафедра «Радиотехника, электроника и телекоммуникации», Алматы, Казахстан
https://orcid.org/0000-0003-3196-696;
Джаканова Гаухар Талгатовна — сениор-лектор, кафедра «Радиотехника, электроника и телекоммуникации», Алматы, Казахстан
E-mail: gjakanova@iitu.edu.kz, https://orcid.org/0009-0004-7890-3573;
Асқанбай Гаухар Тілеубердіқызы — докторант по специальности «Электроэнергетика», Институт энергетики и зеленых технологий, Алматинский университет энергетики и связи имени Г. Даукеева, Алматы, Казахстан
https://orcid.org/0009-0000-3917-9608.

© Т.Д. Иманбекова, Ж.Б. Ибраева, Г.Т. Джаканова, Г.Т. Асқанбай

Аннотация. С развитием цифровых технологий объемы передаваемой информации растут невероятно быстро. Это заставляет нас искать новые решения для расширения памяти устройств и улучшения пропускной способности каналов связи. Раньше, когда интернет-соединение было слабым, отправка даже обычных программ или файлов затягивалась надолго. Именно тогда стали востребованы методы эффективного сжатия данных-так появились привычные нам архиваторы, которые уменьшают размер файла, не портя его содержимое. В данной работе мы подробно изучили современные алгоритмы сжатия, работающие на базе вейвлет-преобразований. Мы не только сравнили существующие подходы, но и создали собственное программное решение в среде MATLAB с помощью пакета Wavelet Toolbox. В ходе тестов мы проверили, насколько сильно можно уменьшить разные типы данных и какие конкретные показатели сжатия при этом получаются. Инструментарий Wavelet Toolbox позволяет гибко работать с сигналами и изображениями. Процесс сжатия здесь происходит за счет удаления второстепенных вейвлет-коэффициентов, после чего сигнал восстанавливается. В ходе исследования были проанализированы алгоритмы сжатия данных на основе дискретного вейвлет-преобразования. Процес-

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

сы сжатия аудио и изображений были реализованы в среде MATLAB, а также было оценено влияние различных типов вейвлетов и уровней разделения на эффективность сжатия. Результаты исследования показали эффективность вейвлетов Dmey и DB20 для аудиосигналов, а также вейвлетов Haar и Bior1.1 для изображений. Предложенный подход позволяет оптимизировать процессы хранения и передачи данных без снижения их качества. Эффективность алгоритма подтверждена высокими результатами при сжатии и восстановлении различных типов сигналов.

Ключевые слова: вейвлет-преобразователь, сжатие информации, аудиосигнал, обработка изображения, коэффициент сжатия, алгоритм восстановления, степень разложения

Для цитирования: Т.Д. Иманбекова, Ж.Б. Ибраева, Г.Т. Джаканова, Г.Т. Асқанбай (2026). Алгоритм сжатия данных на основе вейвлет-преобразователя: анализ и реализация в MATLAB // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 167–184. https://doi.org/10.54309/IJICT.2026.26.2.011. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Заманауи телекоммуникациялық жүйелер ақпараттың өзара алмасуының басты инфрақұрылымы болып анықталды және олар кең көлемдегі мәліметтерді байланыс арналары көмегімен жеткізуге негізделген өте күрделі техникалық, әрі программалық құралдар комплексін қамтиды. Бұндай жүйелер мәтіндік, графикалық, дыбыстық және бейне ақпараттарды сенімді түрде таратуға мүмкіндік береді. Сонымен қатар, бұл жүйелер бірнеше ондаған мыңнан бастап миллионға дейінгі пайдаланушыларға қызмет көрсетуге қабілетті, бұл олардың өнімділігі мен тиімділігіне жоғары талаптар қояды.

Телекоммуникация қызметтерінің сапасын жақсартудағы негізгі мәселелердің бірі- байланыс арналарының өткізу қабілеттілігін барынша тиімді пайдалану. Арналардың өткізу қабілеттілігі көбінесе желі арқылы өтетін трафик көлеміне байланысты (Chen et al., 2024; Daubechies et al., 2001). Трафик көлемінің артуы деректерді беру жылдамдығына кері әсер етуі және желінің жалпы өнімділігін төмендетуі мүмкін. Осыған байланысты ақпарат көлемін азайту, яғни деректерді сығу әдістерін қолдану ерекше маңызды. Деректерді сығу алгоритмдерін қолдану арқылы берілетін ақпарат көлемін азайтуға және арна ресурстарын тиімдірек пайдалануға болады.

Соңғы жылдары толқындық түрлендіру цифрлық телекоммуникация жүйелерінде сигналды өңдеудің тиімді әдістерінің бірі ретінде кеңінен қолданылып келеді. Бұл әдіс сигналды талдау мен өңдеуде жоғары нәтижелерге қол жеткізуге мүмкіндік береді. Толқындық түрлендіруді қолдану арқылы сигналды сығу деңгейін бірнеше есе арттыруға болады, бұл дәстүрлі әдістерге қарағанда айтарлықтай тиімді (Imanbekova et al., 2021). Сонымен қатар, байланыс арнасының өткізу қабілеттілігіне байланысты сигналды белгілі бір деңгейге дейін сығуға және сапасын айтарлықтай төмендетпей беруге болады. Толқындыққа негізделген классикалық сығу алгоритмдеріне Haar,

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

Huffman және Doubéchy әдістері жатады.

Вейвлет түрлендірулерінің негізгі артықшылықтарының бірі-олар сигналды уақыт және жиілік салаларында бір уақытта талдауға мүмкіндік береді. Бұл әдіс, Фурье түрлендіруінің артықшылықтарын сақтай отырып, қосымша мүмкіндіктер ұсынады. Атап айтқанда, вейвлеттер сигналдың уақыт бойынша орналасуын да, оның жиілік сипаттамаларын да дәл анықтауға мүмкіндік береді (Egorova et al., 2019). Нәтижесінде сигнал құрылымын терең талдау жүзеге асырылады және оның маңызды компоненттері бөлінеді.

MATLAB ортасындағы Wavelet Toolbox пакеті вейвлет талдауын жүргізудің қуатты құралдарының бірі болып табылады. Бұл пакет пайдаланушыға графикалық интерфейсті пайдаланып сигналдарды өңдеуге, талдауға және сығуға мүмкіндік береді. Сонымен қатар, Wavelet Toolbox үздіксіз және дискретті вейвлет түрлендірулерін орындау, сигналдар мен кескіндерді шудан тазарту функцияларын, сондай-ақ деректерді сығу құралдарын қамтиды. Бұл мүмкіндіктер ғылыми зерттеулер мен инженерлік мәселелерді тиімді шешуге жағдай жасайды.

Бұл зерттеу жұмысында сигналды сығуға арналған бағдарламалық жасақтама жасалды және оның нәтижелері қарастырылды. Бағдарлама сигналды компоненттерге бөлуге және олардың әрқайсысын бөлек өңдеуге мүмкіндік береді. Артық немесе маңызды емес компоненттерді жою арқылы сигнал құрылымы сығымдалады, нәтижесінде файл өлшемі айтарлықтай азаяды. Тәжірибелер көрсеткендей, бұл тәсіл ақпараттың негізгі сапалық көрсеткіштерін сақтай отырып, сигнал өлшемін шамамен үш есеге азайтуға мүмкіндік береді.

Ақпараттық-телекоммуникациялық технологиялардың қарқынды дамуы сандық сигналдарды өңдеу саласындағы жетістіктермен тығыз байланысты. Бұл жетістіктер ақпаратты қабылдау, өңдеу және беру процестерін оңтайландыруға мүмкіндік береді. Байланыс арналары арқылы деректерді беру кезіндегі негізгі міндеттердің бірі-ақпаратты мүмкіндігінше аз бұрмалаумен сығу және содан кейін оны дәл қалпына келтіру. Осыған байланысты соңғы жылдары тиімді сығымдау алгоритмдерін әзірлеуге қызығушылық артып келеді.

Жалпы, деректерді сығымдау технологияларын пайдалану көп көлемде ақпаратты сақтау кезінде жад ресурстарын үнемдеуге, сондай-ақ байланыс жүйелерінде берілетін деректер көлемін азайтуға мүмкіндік береді. Бұл өз кезегінде файлдардың өлшемін азайтады және оларды өткізу қабілеті шектеулі арналар арқылы жылдамырақ беруге мүмкіндік береді. Сондықтан ақпаратты сығымдау әдістерін жетілдіру қазіргі заманғы телекоммуникация саласының ең маңызды бағыттарының бірі болып табылады.

Әдістер мен материалдар.

Ақпаратты қысу үшін көптеген әр түрлі тәсілдері мен алгоритмдері бар (Imanbekova et al., 2021). Шығынсыз тәсілдер декомпрессиядан кейінгі бастапқы сигналдың ешқандай да бір бұрмалаусыз толық қалпына келуін қамтамасыз етеді.

Бұл кезде, шығыны бар қысу тәсілдері қалпына келген сигналдардың шамалы ауытқуларын жібереді. Бұндай алгоритмдер неғұрлым маңызды емес болып

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

саналатын мәліметтің жойылған бөлігінің есебінен қысудың жоғарғы дәрежесіне қол жеткізеді. Әдетте, олар шығынсыз тәсілдермен салыстырғанда жоғары тиімділігімен қамтамасыз етеді. Қысу алгоритмінің негізгі сипаттамасы болып қысу коэффициенті табылады, ол бастапқы (қысылмаған) мәліметтердің көлемінің қысылған мәліметтердің көлеміне қатынасымен анықталады:

$$k = \frac{S_0}{S_c} \tag{1}$$

k – қысу коэффициенті;

S_0 – қысылмаған мәліметтердің өлшемі;

S_C – қысылған мәліметтердің өлшемі.

Қысу коэффициентінің мәні неғұрлым үлкен болған сайын, қолданып жатқан алгоритмнің тиімділігі соғұрлым жоғары болып саналады. Соңғы уақыттарда, ақпаратты қысуға байланысты тапсырмалардың шешімім кезінде вейвлет-түрлендіргіші негізіндегі амалдары жиі қолданылып жатыр. Қысудың классикалық алгоритмдерінің арасынан Хаара түрлендіргішін, Хаффман кодтауын, Добеши вейвлеттерін сызып айтуға болады. Қысудың сипаттамаларының айтарлықтай жақсаруы компрессияның жоғары деңгейіне ие арифметикалық кодталудың тәсілін көрсетеді (Daubechies et al., 2001).

Дискретті вейвлет-түрлендіргіштердің заманауи іске асыру анализі оны шартты түрде екі негізгі категорияға мүмкіндік береді: бағдарламалық және бағдарламалық-аппараттық құралдар. Қарастырылып отырған зерттеулердегі дискретті-вейвлет-түрлендіргіштердің бағдарламалық жүзеге асырулары MATLAB ортасында қарастырылады.

Жалпы айтқанда, дискретті вейвлет-түрлендіргішті қолданатын бағдарламалық қамтамасыз етуді өндіруді келесідей құрылым түрінде көрсетуге болады:

1. Қойылған тапсырманың контекстінде сигналдардың сипаттамасының алғашқы берілгендерін жинау мен оны талдау;

2. компьютерлік математика жүйесіндегі вейвлет-түрлендіргішінің кітапханасын қолдану арқылы мәліметтерді алдын ала өңдеу, вейвлеттің сәйкес түрін және бөліну деңгейін таңдау;

3. Бөліну деңгейін анықтау және алынған коэффициенттердің өңдеу алгоритмін таңдау;

4. Таңдалған бағдарламалау тілінің ерекшеліктерін есепке ала отырып дискретті вейвлет-түрлендіргіштің алгоритмін бағдарламалық жүзеге асыру.

Мәліметтерді қысу мен қысу алгоритмін зерттеу кезінде өңдеу алгоритмінің жалпы жұмыс принципі тұжырымдалған. 1-ші суретте көрсетілгендей, вейвлет-түрлендіргіш үшін келесі кезеңдерін қосатын өзгертілген алгоритм ұсынылған:

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



Сур. 1. Вейвлет-түрлендіргіш негізіндегі мәліметтерді қысу алгоритмінің блок-схемасы

Бірінші кезеңде қысылуға кететін файлдың импорты жүзеге асырылады. Ары қарай бөліну деңгейі мен сәйкес келетін вейвлет түріне таңдау жүргізіледі, соның негізінде сигналдың вейвлет-ыдырауы орындалады. Үшінші кезеңде сигналдың компоненттерге бөлінуі орындалады, ол ары қарай тиімді қысуды жүргізуге мүмкіндік береді (Mallat et al., 2008). Одан кейін сигналдың көлемінің кішіреюі мен қысу кезінде пайда болатын мүмкін бұрмалауларды жою жүзеге асырылады. Соңғы кезең қысылған күйіндегі файлды сақтау болып табылады.

Нәтижелер және оларды талқылау.

Әзірленген бағдарламалық қамтамасыз ету әр түрлі вейвлет типтерін пайдаланумен аудиосигналды компоненттерге бөлу функциясына ие. Вейвлет-ыдыратуды орындағаннан кейін ұсақ бөлшектерді басу жүргізіледі, бұл сигналдың сапасын төмендетпей артық ақпаратты жоюға мүмкіндік береді (Tropchenko et al., 2009) Әзірленген алгоритмді пайдалану нәтижесінде қабылданатын сапасын сақтай отырып аудиофайлдарды тиімді қысуға қол жеткізе аламыз. Қысылған файлдар сақталатын мәліметтердің көлемін қысқарта отырып ары қарай қолайлы қолдануға болатындай

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

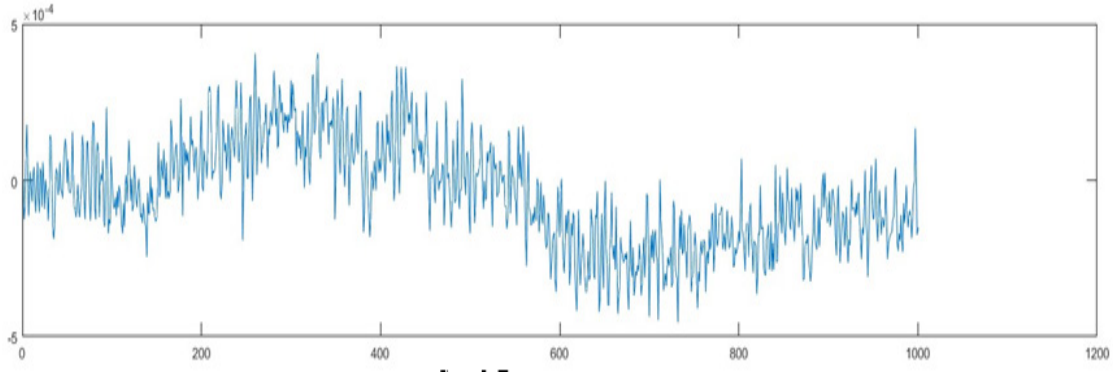
етіп аудиосипаттамаларды сақтайды.

1-ші кестеде өлшемі 427 892 байт болатын аудиофайлды бөлу деңгейінің өзгеруі кезіндегі әр түрлі вейвлеттерді пайдалану арқылы қысу нәтижесі келтірілген. Келтірілген мәліметтер вейвлеттің түрі мен бөлу тереңдігінің қысу тиімділігіне әсерін бағалауға мүмкіндік береді.

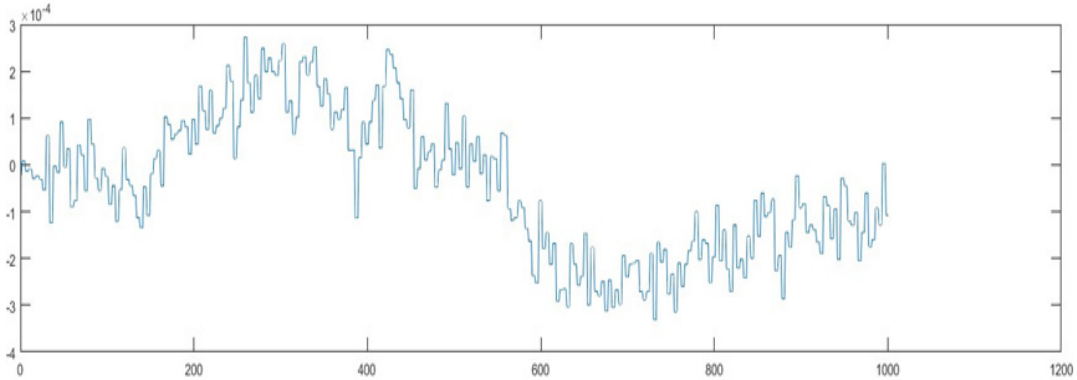
Кесте 1 – Тесттік аудиофайлдың қысылу нәтижесі

Вейвлеттің атауы	n = 2 кезіндегі өлшемі, байт	n = 3 кезіндегі өлшемі, байт	n=2 кезіндегі қысу коэффициенті	n=3 кезіндегі қысу коэффициенті
Haar	262911	263124	1.63	1.62
DB4	210109	176624	2.03	2.43
DB20	144564	87237	2.94	4.92
Sym8	181656	135779	2.35	3.15
Dmey	135778	77028	3.15	5.56
Coif5	169668	116654	2.52	3.66
Fk22	187853	135249	2.28	3.17

Сур. 2–6 бастапқы аудиосигналдарды n бөліну кезінде әр түрлі деңгейде қысу мысалдары көрсетілген.

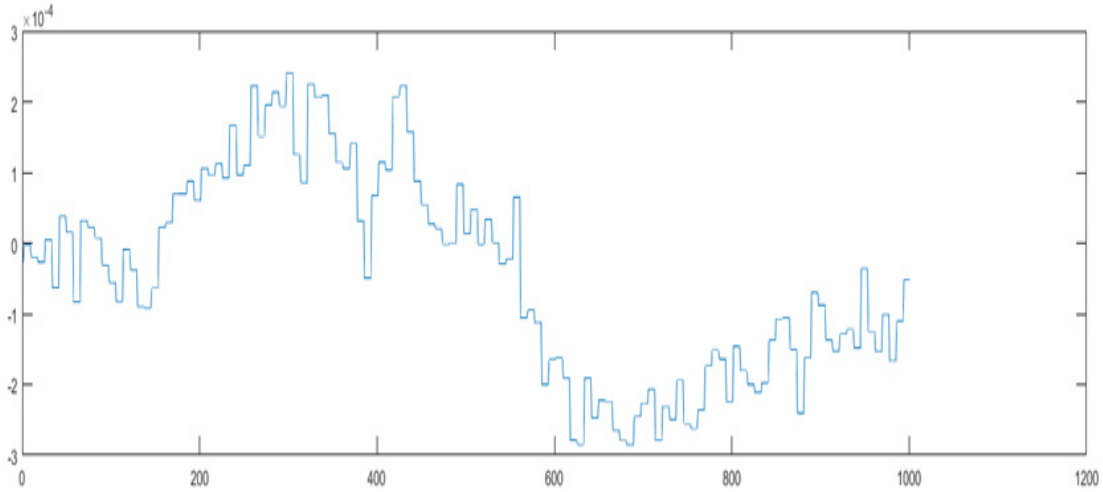


Сур. 2. Бастапқы сигнал

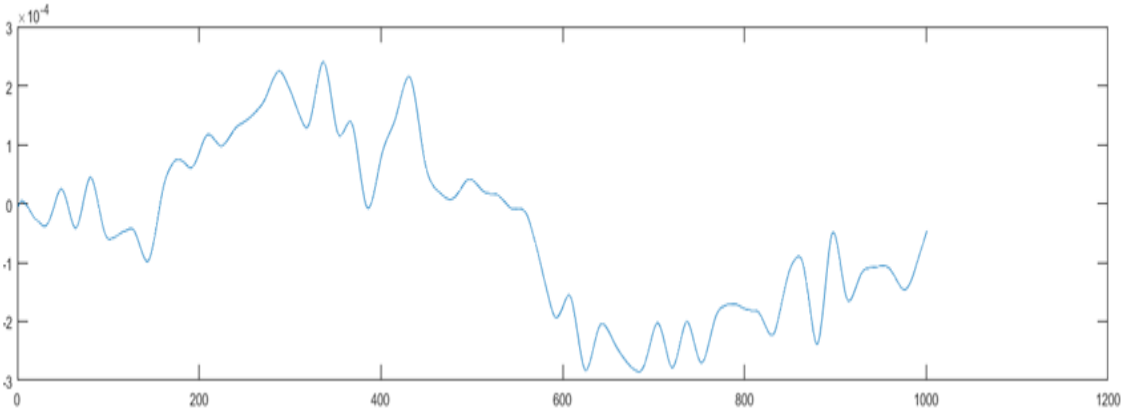


Сур. 3. Хаара вейвлетінің функциясын талдайтын n=2, α=5 кезіндегі қысу

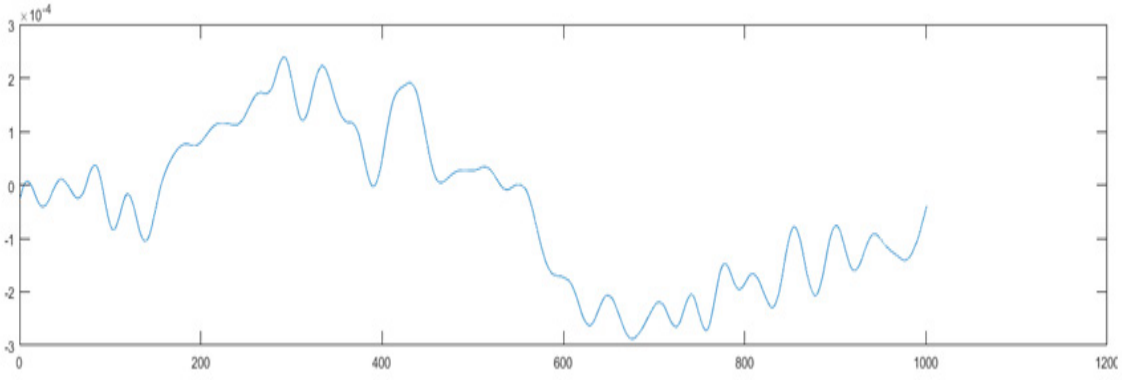
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Сур. 4. Хаара вейвлетінің функциясын талдайтын n=3, α=5 кезіндегі қысу



Сур. 5. Sym 8 вейвлетінің функциясын талдайтын n=4, α=5 кезіндегі қысу.



Сур. 6. fk 22 вейвлетінің функциясын талдайтын n=4, α=5 кезіндегі қысу

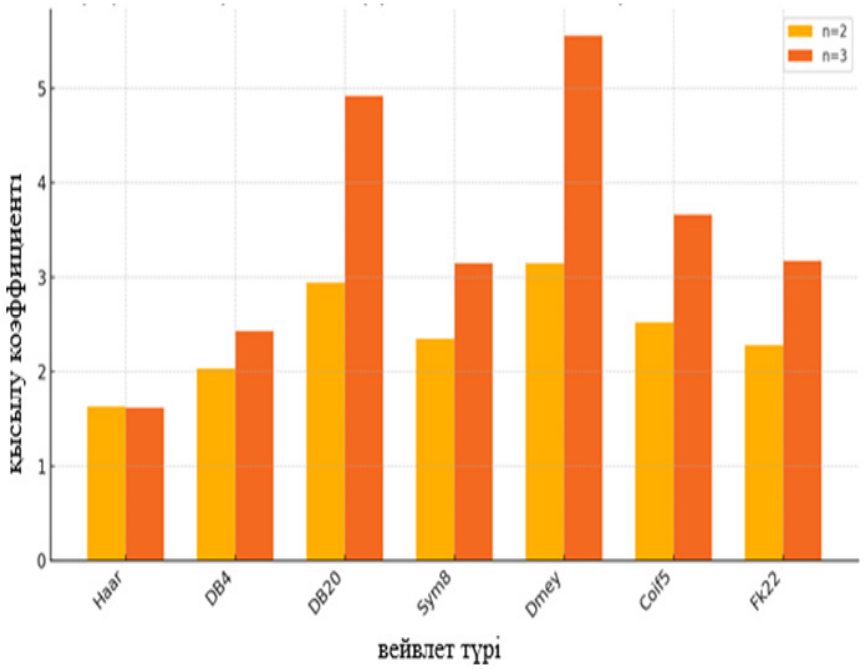
Келтірілген нәтижелерден қысудың тиімділігі пайдаланған вейвлеттің түрі-

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

нен де, таңдалған бөлу деңгейінен де тәуелді екендігі көрініп тұр. n=3n = 3n=3 бөліну деңгейі кезінде қысудың ең тиімді коэффициенті болып Dmey (5.56) и DB20 (4.92) вейвлеті екендігі көрінді, бұл олардың қолайлы сапаны сақтай отырып, аудиосигналын қысу тапсырмаларына жоғары жарамдылығын көрсетеді.

Haar вейвлеті өзінің қарапайымдылығына қарамастан, келтірілген вейвлеттердің арасында ең төмен қысу коэффициентін көрсетіп тұр, бұл қараслырылып отырған тапсырмалардың ішінде ең тиімсізі болып табылды.

Осылайша, қысылу деңгейі мен қайта қалпына келтірілген аудиосигналдың арасындағы компромисіне келу вейвлет мен бөліну деңгейін таңдаудағы маңызды рөлін атқарады.



Сур. 7. Әр түрлі вейвлеттері үшін қысылу коэффициенттерін графикалық салыстыру

Жоғарыда *n=2* и *n=3* бөлу деңгейі кезінде әр түрлі вейвлеттері үшін қысылу коэффициенттерінің графикалық салыстырылуы көрсетілген. Бұл вейвлет түрі мен бөлу деңгейіне байланысты қысу тиімділігі қалай өзгеретіндігін анық көрсетіп тұр.

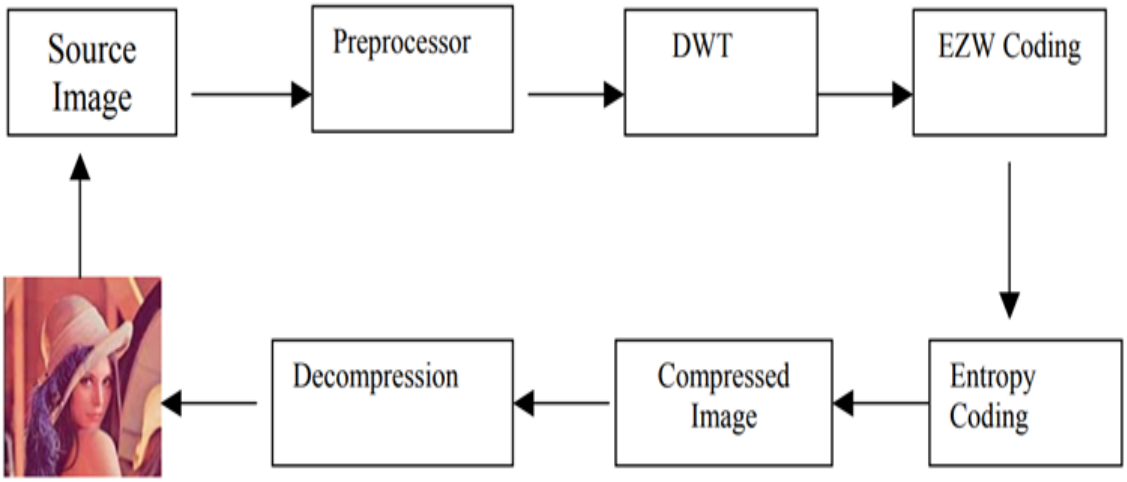
Вейвлет-түрлендіргіш арқылы мәліметтерді қысу

Вейвлет-әдістерін салыстыру үшін C++ тілінде жазылған және MinImage деп аталатын бағдарламалық құрал қолданылды. Бұл бағдарлама бастапқыда тек бір түрдегі вейвлетті тексеру мақсатында жасалған болатын. Кейін қосымша функциялар енгізіліп, басқа вейвлет түрлерін қолдау мүмкіндігі қосылды, сонымен қатар EZW-кодтау алгоритмі іске асырылып, сөндіру тиімділігін арттыруға мүмкіндік берді. Осы тарауда біз MinImage бағдарламасының кейбір жүзеге асырылу ерекшеліктерін қарастырамыз.

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

MinImage вейвлет-кескін компрессоры (Сур. 8 қараңыз) 24 биттік және 8 биттік сұр түсті суреттерді қысуға арналған. Бастапқыда бағдарлама Хаар вейвлетін қолдана отырып, диапазон асты кодтау үшін жасалған болатын. Түрлі вейвлеттерді салыстыру үшін Daubechies және биортогональды сплайн-вейвлеттер сияқты басқа түрлері де енгізілді. Сонымен қатар, бастапқы диапазон асты кодтау EZW-кодтауға ауыстырылып, қысу нәтижелерін жақсарту мүмкіндігі қосылды.

MinImage бағдарламасының маңызды ерекшелігі-сілтеме арқылы қысу параметрлерін өзгерту арқылы әртүрлі қысу дәрежелері мен сурет сапасын алуға мүмкіндік беруі. Пайдаланушы файл көлемін немесе сурет сапасын таңдауға болады. Сонымен қатар, әртүрлі суреттерге әртүрлі вейвлеттерді қолдану арқылы ең тиімді қысу нәтижелеріне жетуге болады.



Сур. 8. Қысудың негізгі сұлбасы

MATLAB ортасында вейвлет-түрлендіргіштің негізінде кескінді қысу

Дискретті вейвлет-түрлендіргішті пайдалану кезінде бөлінудің деңгейін кеңейту сигналдың неғұрлым тиімді қысылуына көмектеседі (Mallat, 1989; Daubechies, 1992). Бұл әр ыдыратудың келесі деңгейінде сигналдың ұсақ бөлшектері одан ары ұсақ болып бөлінетіндігімен түсіндіріледі, олардың айтарлықтай бөлігін визуалдық немесе аудиолық сапасын жоғалтпай алып тастауға болады.

Зерттеуді жалғастыру кезінде MATLAB ортасында екі өлшемді сигнал-кесіндіні қысу үрдісі жүргізілді. Тесттік файл ретінде өлшемі 686 732 байттық кескін пайданалынды.

MATLAB ортасында кескінді қысу үрдісі келесідей негізгі кезеңдерден өтеді:

1. Бастапқы кескінді жүктеу;

2. Кескінді вейвлет-түрлендіргіштің операцияларымен сәйкес келетін форматқа түрлендіру (мысалы, сұр түстеріне түрлендіру);

3. Берілген бөлу деңгейінде дискретті вейвлет-түрлендіруінің орындалуы;

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

4. Кесіндінің қорытынды сапасына әсері төмен коэффициенттерді жою;
5. Өзгертілген коэффициенттердің негізінде кескіндерді қалпына келтіру;
6. Алынған қысылған кесіндіні сақтау мен қысу тиімділігін бағалау.

Берілген іске асыру вейвлет-түрлендіргіштің тек қана аудиосигналдар үшін ғана емес, сонымен қатар, кескін сияқты екі өлшемді мәліметтер үшін де қолданылу мүмкіндігі расталды. Алынған нәтижелер бөлудің оптималды параметрлерін және вейвлеттің сәйкес түрін таңдау кезінде ең тиімді тәсілге көрсетеді. Вейвлет-талдау негізінде ақпаратты қысуға арналған арнайы әзірленген алгоритмге сүйене отырып, статикалық кескіндерді (2D сигналдарды) қысуға мүмкіндік беретін бағдарлама жасалды (Сур. 9 қараңыз).

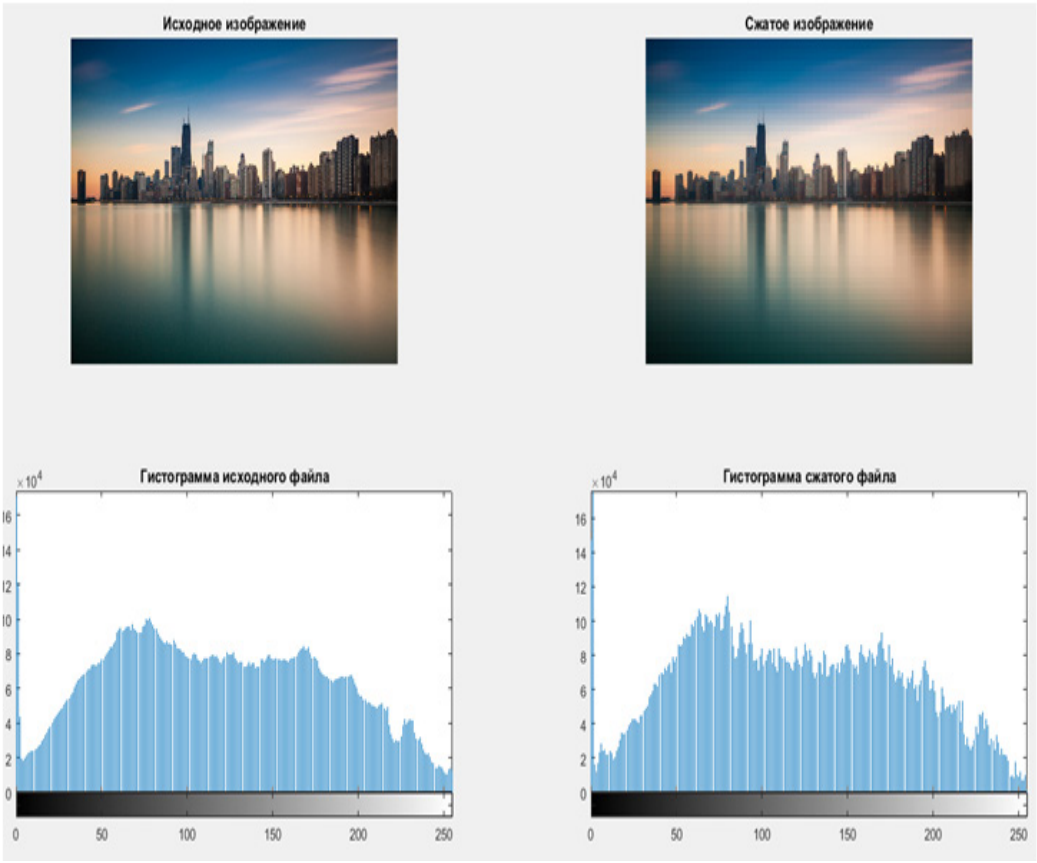
```
clear;
close all;
clc;
% Загрузка файла
file = 'mount.jpg';
X=imread(file);
% n: уровень сжатия, wname: имя вейвлета
n = 6;
wname = 'haar';
x = double(X);
% Вейвлет-разложение изображения
[c,s] = wavedec2(x,n,wname);
% Алгоритм выбора коэфициента шумоподавления
alpha = 1.1; m = 10*prod(s(1,:));
[thr,nkeep] = wdcbm2(c,s,alpha,m);
% Сжатие изображения
[xd,cxd,sxd,perf0,perfl2] = wdencmp('lvd',c,s,wname,n,thr,'s');
disp('Соотношение сжатия');
disp(perf0);
xc = uint8(xd);
% Восстановление сигнала
R = waverec2(c,s,wname);
rc = uint8(R);
x = uint8(x);
% Построение изображений
figure;
subplot(221), imshow(x);
title('Исходное изображение')
subplot(222), imshow(xc);
title('Сжатое изображение')
subplot(223), imhist(x);
title('Гистограмма исходного файла')
subplot(224), imhist(xc);
title('Гистограмма сжатого файла')
```

Сур. 9. 2D кескінін қысуға арналған бағдаоламаның скрипті

Бағдарлама кескінді құрамдас бөліктерге жіктеу мүмкіндігіне ие және көптеген вейвлет түрлерімен жұмыс істейді. Құрамдас бөліктерге ажыратқаннан кейін

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

бағдарлама суреттегі немесе фотодағы артық детальдарды тегістейді (Сур. 9), соның нәтижесінде қажетсіз ақпарат жойылып, шығыста көлемі шамамен 2,5 есе қысқартылған файл алынады.



Сур. 9. Фотосуретті қысу мысалы

Алынған мәліметтер бөліну деңгейінің ұлғаюы қорытынды файлдың өлшемінің айтарлықтай қысқаруына әкелетінін дәлелдейді. Неғұрлым байқалатын әсер Haar, Bior1.1 и Rbio1.1 вейвлеттерінде көрінеді, бұл жерде n=5n = 5n=5 бөліну деңгейінде бастапқы файлдың өлшеміне қарағанда 2,5 есе ден астам азаяды. Db40, Fk22 және Sym8 сияқты күрделі вейвлеттері көлемнің азырақ айқын төмендеуін көрсетеді, бұл терең бөліну кезінде құрылымдық ақпараттың көп бөлігін сақтаумен байланысты болуы мүмкін. Дегенімен, бұндай жағдай болғанның өзінде де кескіннің визуалды қабылдауды жоғалтпай қысуға қол жеткізуге болады.

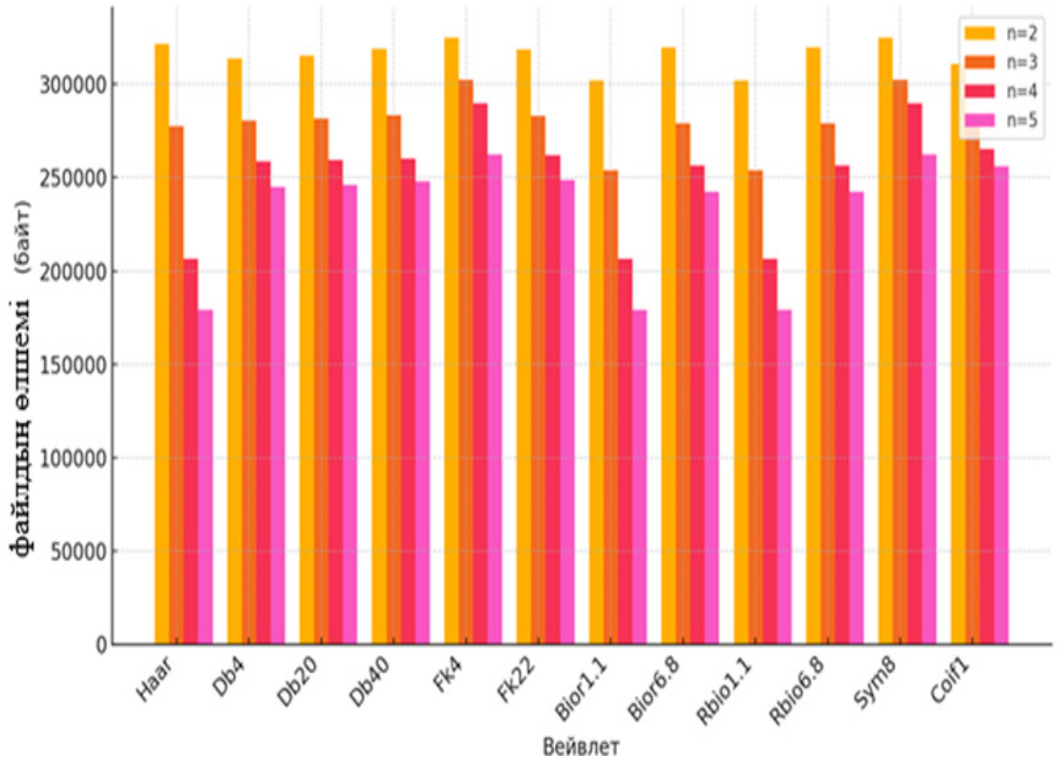
Демек, 2D-сигналдардың вейвлет-қысудың алгоритмінің тиімділігі тікелей вейвлет пен бөлудің тереңдігіне тәуелді. Симметриялық құрылымы бар қарапайым вейвлеттертер (мысалы, Haar, Bior1.1) қысудың жоғары деңгейін қамтамасыз етуі мүмкін. Бұ өз кезегінде оларды мәліметтерді сақтау көлемінің шектеулері маңызды болатын тапсырмалар үшін қолайлы етеді.

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

Кескінді қысуды нәтижелері төмендегі кестеге енгізілген (2-ші кесте).

Кесте 2 – 2D кескінді бөлу деңгейінің әр түрлі параметрлерімен қысу нәтижесі

Вейвлет аталуы	n = 2 кезіндегі өлшемі, байт	n = 3 кезіндегі өлшемі, байт	n = 4 кезіндегі өлшемі, байт	n = 5 кезіндегі өлшемі, байт
Haar	321739	277726	206405	179253
Db4	313929	280484	258686	244978
Db20	315488	281613	259534	246237
Db40	319145	283429	260340	248120
Fk4	324840	302233	289699	262432
Fk22	318657	282986	261870	248844
Bior1.1	301907	253795	206405	179253
Bior6.8	319651	279053	256525	242259
Rbio1.1	301907	253795	206405	179253
Rbio6.8	319651	279053	256525	242259
Sym8	324840	302233	289699	262432
Coif1	310995	280177	265217	256053



Сур. 10. Әр түрлі қысу деңгейіндегі 2D кескінін қысуын салыстыру

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

10-шы суретте әр вейвлет үшін бөлудің әр түрлі деңгейін қолданумен 2D-кескіннің қысылуының графикалық салыстырылуы көрсетілген. Графиктен бөлу деңгейі үлкейген сайын файлдың өлшемі өзгеретіні көрініп тұр, ол өз кезегінде вейвлеттердің әр түрлі типтері үшін қысудың тиімділігін көрсетіп тұр.

Кескіндерді қысу алгоритмдерінің тиімділігін анықтау мақсатында кескін сапасын сипаттайтын бірқатар сандық көрсеткіштер пайдаланылды. Аталған метрикалар қысылған кескін мен бастапқы кескін арасындағы айырмашылық деңгейін бағалауға мүмкіндік береді. Практикада жиі қолданылатын негізгі көрсеткіштер қатарына орташа квадраттық қате (MSE-Mean Square Error), сигнал мен шудың шыңдық қатынасы (PSNR-Peak Signal-to-Noise Ratio) және құрылымдық ұқсастық индексі (SSIM-Structural Similarity Index) жатады.

MSE көрсеткіші бастапқы кескін мен қайта қалпына келтірілген кескін арасындағы пиксельдік айырмашылықтардың орташа квадраттық мәнін сипаттайды (Wang et al., 2004). Бұл көрсеткіштің мәні неғұрлым төмен болған сайын, сығылған кескіннің сапасы соғұрлым жоғары деп бағаланады, яғни бастапқы ақпараттың бұрмалануы аз болады.

PSNR параметрі сигнал мен шудың арақатынасын сипаттай отырып, қысу нәтижесінде алынған кескін сапасының деңгейін анықтайды (Kumar et al., 2022). Бұл көрсеткіштің мәні жоғары болған жағдайда кескін сапасының жақсы сақталғанын білдіреді. Әдетте 30–40 dB диапазонындағы мәндер қанағаттанарлық және жоғары сапаны көрсетеді.

SSIM индексі екі кескіннің құрылымдық ұқсастық дәрежесін бағалауға арналған. Оның мәні-1 мен 1 аралығында өзгереді, мұндағы 1 мәні кескіндердің толық сәйкес келетінін көрсетеді. Бұл көрсеткіш визуалды қабылдауға жақын бағалау жүргізуге мүмкіндік береді.

MSE көрсеткішін есептеу келесі өрнек арқылы жүзеге асырылады:

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(I(i,j) - K(i,j))^2 \quad (2)$$

Ал PSNR мәні төмендегі формула бойынша анықталады:

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \quad (3)$$

Кестеде көрсетілген нәтижелерге сәйкес, вейвлет-түрлендіру негізінде құрылған қысу алгоритмі кескін сапасын айтарлықтай деңгейде сақтай отырып, деректер көлемін тиімді түрде азайтуға мүмкіндік береді. Әсіресе Symlet және Daubechies вейвлеттері жоғары сапа көрсеткіштерімен ерекшеленеді.

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

3-ші кестеде әртүрлі вейвлет түрлерін қолдану арқылы алынған кескіндерді қысу нәтижелерінің сапа көрсеткіштері берілген.

Кесте 3 – Кескінді қысу нәтижелерінің сапа көрсеткіштері

Вейвлет түрі	Қысу коэффициенті	PSNR (dB)	SSIM	MSE
Haar	3.8	34.5	0.92	8.21
Daubechies	4.2	35.7	0.94	7.04
Symlet	4.5	36.3	0.95	6.62

Қорытынды.

1. Вейвлет-түрлендіргішті пайдалану арқылы мәліметтерді қысу тиімділігі: Бұл мақалада қарастырылғанаудиосигналдар мен кекіндерді қысуға арналған вейвлет-түрлендіргіштерді пайдалану жоғары тиімді тәсіл болып табылатындығын растады.

2. Вейвлет түрінің қысу тиімділігіне әсері: Мақалада қысудың тиімділігі сигнлады түрлендіруге арналған вейвлеттің түріне тәуелді екендігі көрсетілген. Dmey және DB20 вейвлеттері (5.56 мен 4.92 сәйкес) аудиосигналдарды қысу кезінде ең жоғары коэффициенттерді көрсетті, бұл оларды сапасын сақтай отырып қысу тапсырмасы үшін қолайлы етеді. Бұл кезде Haar вейвлеті өзінің қарапайымдылығына қарамастан, қысу контекстінде ең тиімсізі болып шықты.

3. Бөлу деңгейінен қысу коэффициентінің тәуелділігі: Бөлу деңгейі қысудың тиімділігіне айтарлықтай әсерін тигізеді. Бөліну деңгейі жоғары болған сайын, сигналдың сапасын төмендетпей көбірек мәліметтерді өшіруге болады. Аудиосигналдар үшін ең үлкен қысу коэффициенті n=3 бөлу деңгейі кезінде, ал кескін үшін – n=5 кезінде болды.

4. MATLAB ортасында іске асыру бағдарламасы: MATLAB ортасында дискретті вейвлет-түрлендіргішін жүзеге асыратын алгортимін қамтамасыз ететін әзірленген бағдарламалық қамтамасыз ету аудиосигналдарды да, кескіндерді де өңдеудің жоғары қабілеті мен функционалдылығын көрсетеді.

5. Әртүрлі мәліметтер үшін вейвлетті таңдау бойынша ұсыныстар: Аудиосигналдар үшін ең тиімді болып Dmey және DB20 вейвлеттері табылады, ал кескіндер үшін Haar мен Bior1.1 вейвлеттерін таңдаған жөн, себебі олар визуалды сапасын сақтай отырып қысудың жоғары коэффициентін қамтамасыз етеді.

6. Зерттеудің тәжірибелік құндылығы: Вейвлет-түрлендіргішті пайдаланумен әзірленген қысу алгоритмі мәліметтетреді шынайы сақтау мен жіберу жүйелерінде қолданылуы мүмкін, мысалы, мультимедианы басқару жүйелері, аудио және кескін файлдарын мұрағаттау және сапаны жоғалтпай деректер көлемін азайту маңызды ресурс шектеулі қолданбаларда.

Жалпы алғанда, вейвлет-түрлендіру негізіндегі қысу алгоритмдері ақпараттың негізгі құрылымдық элементтерін сақтай отырып, деректер көлемін айтарлықтай азайтуға мүмкіндік беретіні анықталды. Бұл әдіс әсіресе үлкен көлемдегі мультимедиялық мәліметтермен жұмыс істейтін жүйелер үшін тиімді болып табылады.

Зерттеудің ғылыми маңыздылығы–вейвлеттік талдаудың мәліметтерді қысу

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

саласындағы теориялық және қолданбалы мүмкіндіктерін кешенді түрде қарастыруында. Ал практикалық маңыздылығы – ұсынылған алгоритмдерді нақты жүйелерде, атап айтқанда телекоммуникациялық желілерде, бұлттық сақтау жүйелерінде, мультимедиялық деректерді өңдеу және архивтеу жүйелерінде қолдануға болатындығында.

Қорытындылай келе, ұсынылған әдістер мен алгоритмдер ақпаратты тиімді қысу, сақтау және тасымалдау мәселелерін шешуде маңызды құрал болып табылады. Болашақта бұл бағыттағы зерттеулерді жетілдіру, әсіресе адаптивті вейвлет әдістерін енгізу және жасанды интеллект негізіндегі қысу алгоритмдерімен біріктіру, зерттеу нәтижелерін одан әрі дамытуға мүмкіндік береді.

REFERENCES

Chen L., Zhao Y. (2024). Towards Robust Text-Guided Image Compression Under Modality Missing. IEEE Transactions on Multimedia. — Vol. 26. — Pp. 4937-4952. DOI: https://doi.org/10.1109/TMM.2024.3364231. [in Eng.].

Daubechies, I. (2001). Desyat' lektsiy po veyvletam [Ten Lectures on Wavelets] (E.V. Mishchenko, Trans.; A.P. Petukhov, Ed.). —M.: RKhD. (In Russ.). Pp. 464. ISBN 5-93972-044-7. [in Russ.].

Egorova E.V., Aksyaitov M.Kh., Rybakov A.N. (2019). Methods for Improving the Efficiency of Wavelet Transformations in Processing, Compression and Restoration of Radio Signals: A Monograph. – Tambov: Ucom, 84 pages. (A research monograph on practical improvements in wavelet applications for radio signals.) [in Russ.].

Graps A. (1995). An Introduction to Wavelets//IEEE Computational Science and Engineering. — Vol. 2. — No. 2. Pp. 50–61. DOI: https://doi.org/10.1109/99.388960 [in Eng.].

Imanbekova T.D., Zhaksylyk A., Kozlov I.A. (2021). Application of wavelet transformation for information compression // QazBSQA Bulletin. Humanities and natural sciences. Economy. — Vol. 1. — No. 79. DOI: https://doi.org/10.51488/1680-080X/2021.1-42 [in Russ.].

Kumar R., Singh S. (2023). Hybrid Approach for Image Denoising using Deep Learning and Traditional Metrics // Multimedia Tools and Applications (Springer). Pp. 145–158. DOI: https://doi.org/10.1007/s11042-023-14456-w. [in Eng.].

Li X., Chen M. (2021). Hybrid image compression algorithm based on SPIHT and DWT, Digital Signal Processing. — Vol. 112. — Article 103001. DOI: https://doi.org/10.1016/j.dsp.2021.103001 . [in Eng.].

Mallat S. (2008). A Wavelet Tour of Signal Processing. – Elsevier, 3rd Edition. Pp. 832. [in Eng.].

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation // IEEE Transactions on Pattern Analysis and Machine Intelligence. — Vol. 11. — No. 7. Pp. 674–693. DOI: https://doi.org/10.1109/34.192463 . [in Eng.].

Tropchenko A.Yu., Tropchenko A.A. Tropchenko, A. Yu., & Tropchenko, A. A. (2009). Metody szhatiya izobrazheniy, audiosignalov i video: uchebnoe posobie [Image, audio and video compression methods: a textbook]. — Saint Petersburg: SPbGU ITMO. Pp. 108. [in Russ.].

Wang Z., Bovik A. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity // IEEE Transactions on Image Processing. — Vol. 13. — No. 4. Pp. 600–612. DOI: https://doi.org/10.1109/TIP.2003.819861 [in Eng.].

Zhang Y., Li H. (2023). An Improved Image Compression Algorithm Based on Integer Wavelet Transform and Context-Based Arithmetic Coding // IEEE Access. — Vol. 11. — Pp. 12485–12497. DOI: https://doi.org/10.1109/ACCESS.2023.3242031 [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.2. Number 26 (2026). Pp. 185–205
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.012
UDC 004.8:004.42:004.65

A MATHEMATICAL AND ALGORITHMIC APPROACH TO THE DEVELOPMENT OF AN INTELLIGENT TEXT-TO-SQL SYSTEM BASED ON LARGE LANGUAGE MODELS

B.Z. Kenzhegulov[1], Zh.T. Bilyalova[1], K.N. Uteuliyeva[1]*, L. Nurgaliyeva[2], Sh.S. Nurzhanova[1]

[1]Atyrau University named after Kh. Dosmukhamedov, Atyrau, Kazakhstan;
[2]Saidot Ltd., Хельсинки, Финляндия.
E-mail: kamka_n@mail.ru

Beket Kenzhegulov — Doctor of Technical Sciences, Professor of the Department of Mathematics and Methods of Teaching Mathematics, Atyrau University named after Kh. Dosmukhamedov, Atyrau, Kazakhstan
https://orcid.org/0000-0001-6230-2926;
Zhupar Bilyalova — Candidate of Pedagogical Sciences, Professor of the Department of Mathematics and Methods of Teaching Mathematics, Atyrau University named after Kh. Dosmukhamedov, Atyrau, Kazakhstan
https://orcid.org/0000-0002-3362-8086;
Kamka Uteuliyeva — Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematics and Methods of Teaching Mathematics, Atyrau University named after Kh. Dosmukhamedov, Atyrau, Kazakhstan
E-mail: kamka_n@mail.ru, https://orcid.org/0009-0004-1195-1642;
Lunara Nurgaliyeva — Master's degree holder in AI Security and Machine Learning, Saidot Ltd., Helsinki, Finland
https://orcid.org/0009-0005-5252-9525;
Sharbat Nurzhanova — Candidate of Pedagogical Sciences, Professor of the Department of Mathematics and Methods of Teaching Mathematics, Atyrau University named after Kh. Dosmukhamedov, Atyrau, Kazakhstan
https://orcid.org/0009-0006-7726-1999.

© B.Z. Kenzhegulov, Zh.T. Bilyalova, K.N. Uteuliyeva, L. Nurgaliyeva, Sh.S. Nurzhanova

Abstract. The article addresses the problem of developing an intelligent Text-to-SQL system designed to automatically transform user queries expressed in natural language into SQL queries for relational databases. The relevance of the topic is associated with the rapid growth of structured data and the need to increase the accessibility of analytical tools

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

for users who do not possess professional-level SQL skills. The aim of the study is to analyze and practically test methods for applying large language models to automate SQL analytics, as well as to compare three approaches: fine-tuning, RAG, and agent-based systems. To achieve this aim, the paper considers the mathematical formulation of the problem, the architecture of the RAG pipeline, the mechanism of vector search, and agent-based error self-correction. The results show that the combination of LLM, RAG, and an agent-based approach improves the adaptability, accuracy, and practical applicability of Text-to-SQL systems in corporate analytics. It is concluded that a combined approach is appropriate for complex analytical queries and dynamically changing databases.

Keywords: Text-to-SQL, large language models, SQL, RAG, agent-based systems, vector search, relational databases

For citation: B.Z. Kenzhegulov, Zh.T. Bilyalova, K.N. Uteuliyeva, L. Nurgaliyeva, Sh.S. Nurzhanova (2026). A mathematical and algorithmic approach to the development of an intelligent text-to-sql system based on large language models // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 185–205. https://doi.org/10.54309/IJICT.2026.26.2.012. (In Eng.).

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

ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР НЕГІЗІНДЕ ИНТЕЛЛЕКТУАЛДЫ TEXT-TO-SQL ЖҮЙЕСІН ӘЗІРЛЕУДІҢ МАТЕМАТИКАЛЫҚ-АЛГОРИТМДІК ТӘСІЛІ

Б.З. Кенжегулов[1], Ж.Т. Билялова[1], К.Н. Утеулиева[1]*, Л. Нургалиева[2], Ш.С. Нуржанова[1]

[1]Х. Досмұхамедов атындағы Атырау университет, Атырау, Қазақстан;
[2]Saidot Ltd., Хельсинки, Финляндия.
E-mail: kamka_n@mail.ru

Кенжегулов Бекет — техника ғылымдарының докторы,математика және математика оқыту әдістемесі кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0001-6230-2926;
Билялова Жупар — педагогика ғылымдарының кандидаты, математика және математика оқыту әдістемесі кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0002-3362-8086;
Утеулиева Камка — физика - математика ғылымдарының кандидаты, Математика және математики 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;
Нургалиева Лунара — магистр, AI Safety ML/LLM Engineering, Saidot Ltd., Хельсинки, Финляндия
https://orcid.org/0009-0005-5252-9525;

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

Нуржанова Шарбат — педагогика ғылымдарының кандидаты, математика және математики оқыту әдістемесі кафедрасының қауымдастырылған профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0009-0006-7726-1999.

© Б.З. Кенжегулов, Ж.Т. Билялова, К.Н. Утеулиева, Л. Нургалиева, Ш.С. Нуржанова

Аннотация. Мақалада пайдаланушылардың табиғи тілде берілген сұраныстарын реляциялық деректер қорларына арналған SQL-сұраныстарға автоматты түрде түрлендіруге бағытталған интеллектуалды Text-to-SQL жүйесін әзірлеу мәселесі қарастырылады. Зерттеу тақырыбының өзектілігі құрылымдалған деректер көлемінің қарқынды өсуімен және SQL тілін кәсіби деңгейде меңгермеген пайдаланушылар үшін аналитикалық құралдардың қолжетімділігін арттыру қажеттілігімен байланысты. Зерттеудің мақсаты — SQL-аналитиканы автоматтандыруда үлкен тілдік модельдерді қолдану әдістерін талдау және практикалық апробациядан өткізу, сондай-ақ Fine-tuning, RAG және агенттік жүйелер сияқты үш тәсілді салыстыру. Осы мақсатқа жету үшін жұмыста мәселенің математикалық қойылымы, RAG-пайплайн архитектурасы, векторлық іздеу механизмі және агенттік қателерді өздігінен түзету тәсілі қарастырылды. Нәтижесінде LLM, RAG және агенттік тәсілді біріктіру корпоративтік аналитикадағы Text-to-SQL жүйелерінің бейімделгіштігін, дәлдігін және практикалық қолданылуын арттыратыны көрсетілді. Қорытындыда күрделі аналитикалық сұраныстар мен динамикалық түрде өзгеретін деректер қорлары үшін біріктірілген тәсілді қолдану тиімді екені анықталды.

Түйінді сөздер: Text-to-SQL, үлкен тілдік модельдер, SQL, RAG, агенттік жүйелер, векторлық іздеу, реляциялық деректер қоры

Дәйексөздер үшін: Кенжегулов Б.З., Билялова Ж.Т., Утеулиева К.Н., Нургалиева Л., Нуржанова Ш.С. (2026). Үлкен тілдік модельдер негізінде интеллектуалды text-to-sql жүйесін әзірлеудің математикалық-алгоритмдік тәсілі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 185–205. https://doi.org/10.54309/IJICT.2026.26.2.012. (Ағыл. тіл.).

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

МАТЕМАТИКО-АЛГОРИТМИЧЕСКИЙ ПОДХОД К РАЗРАБОТКЕ ИНТЕЛЛЕКТУАЛЬНОЙ TEXT-TO-SQL СИСТЕМЫ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Б.З. Кенжегулов[1], Ж.Т. Билялова[1], К.Н. Утеулиева[1]*, Л. Нургалиева[2], Ш.С. Нуржанова[1]

[1]Атырауский университет имени Х.Досмухамедова, Атырау, Қазақстан;
[2]Saidot Ltd., Хельсинки, Финляндия.
E-mail: kamka_n@mail.ru

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

Кенжегулов Бекет — доктор технических наук, профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0001-6230-2926;
Билялова Жупар — кандидат педагогических наук, профессор кафедры математики и методики преподавания математики, Атырауский университет имени Х.Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0002-3362-8086;
Утеулиева Камка — кандидат физико-математических наук, доцент 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;
Нургалиева Лунара — магистр, AI Safety ML/LLM Engineering, Saidot Ltd., Хельсинки, Финляндия
https://orcid.org/0009-0005-5252-9525;
Нуржанова Шарбат — кандидат педагогических наук, профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0009-0006-7726-1999.

© Б.З. Кенжегулов, Ж.Т. Билялова, К.Н. Утеулиева, Л. Нургалиева, Ш.С. Нуржанова

Аннотация. В статье рассматривается проблема разработки интеллектуальной системы Text-to-SQL, предназначенной для автоматического преобразования пользовательских запросов на естественном языке в SQL-запросы к реляционным базам данных. Актуальность темы связана с ростом объемов структурированных данных и необходимостью повышения доступности аналитических инструментов для пользователей, не владеющих SQL на профессиональном уровне. Цель исследования заключается в анализе и практической апробации методов применения больших языковых моделей для автоматизации SQL-аналитики, а также в сравнении трех подходов: Fine-tuning, RAG и агентных систем. Для достижения цели рассмотрены математическая постановка задачи, архитектура RAG-пайплайна, механизм векторного поиска и агентная самокоррекция ошибок. В результате показано, что сочетание LLM, RAG и агентного подхода повышает адаптивность, точность и практическую применимость Text-to-SQL систем в корпоративной аналитике. Сделан вывод о целесообразности комбинированного подхода для сложных аналитических запросов и динамически изменяющихся баз данных.

Ключевые слова: Text-to-SQL, большие языковые модели, SQL, RAG, агентные системы, векторный поиск, реляционные базы данных

Для цитирования: Кенжегулов Б.З., Билялова Ж.Т., Утеулиева К.Н., Нургалиева Л., Нуржанова Ш.С. (2026). Математико-алгоритмический подход к разработке интеллектуальной text-to-sql системы на основе больших языковых моделей

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

// Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 185–205. https://doi.org/10.54309/IJICT.2026.26.2.012. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The modern digital economy generates unprecedented volumes of structured data: according to Statista, the total volume of global data will reach 175 zettabytes by 2025. In this environment, the ability to quickly extract accurate analytical information from databases is a significant competitive advantage for organizations. SQL (Structured Query Language) – the standard query language for relational DBMS – remains the main tool for accessing structured data: as of 2025, relational databases occupy more than 60 % of the DBMS market (Zhong et al., 2017).

At the same time, SQL queries pose a significant barrier for a significant portion of business users. According to Defog.ai analysts, analytics department employees spend 30 to 50% of their work time formulating and debugging data queries, with each non-trivial query potentially requiring the involvement of a qualified developer or data engineer (Yu et al., 2018). This reduces operational efficiency and creates a bottleneck in the decision-making chain.

Large Language Models (LLMs)—neural network architectures based on attention mechanisms (Transformers) trained on trillions of text tokens—open up a fundamentally new opportunity: the automatic transformation of natural language queries into syntactically and semantically correct SQL queries. This task, commonly referred to as Text- to -SQL, is one of the most practical applications of LLMs in the field of data science.

to-SQL problem is not new: the first attempts to solve it were made back in the 1970s using rule-based systems (LUNAR, BASEBALL). However, it was the advent of large pre-trained The introduction of transformer models (BERT, GPT, T5) in 2018–2020, and then super-large LLMs (GPT-4, LLaMA, Claude) in 2022–2024, provided a qualitative leap in accuracy: from 60–70 % on the Spider benchmark to 93 % for the specialized SQLCoder-70B model (Li et al., 2023).

The relevance of this work stems from the gap between the high potential of modern LLMs and the complexity of their practical implementation in corporate analytical systems. A comparative study of three approaches to solving the Text- to -SQL problem—fine-tuning, RAG, and agent-based systems—is particularly important, as each has its own advantages, limitations, and requirements for effective application. Furthermore, the domestic scientific literature lacks comprehensive studies combining theoretical analysis, practical implementation, and economic evaluation of these technologies.

The aim of the work is to research and practically test methods for using large language models to automate data analysis in SQL, to comparatively evaluate three main technologies (paradigms) – Fine- tuning, RAG, agent systems – and to determine the conditions for their effective use in a corporate environment.

To achieve the goal, the following research objectives were defined:

- to -SQL problems using LLM, and to identify their comparative advantages and

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

limitations.

- LLM fine-tuning technologies (Full Fine- tuning, LoRA, QLoRA, RLHF) as applied to the Text- to -SQL task; to explore the practical aspects of deploying the SQLCoder model.
- To review the RAG (Retrieval-Augmented Generation) technology and explore its implementation in the specialized Vanna.AI framework.
- Develop an agent-based text-to-SQL conversion system with an automatic verification and error correction mechanism based on the smolagents framework.

The object of the research is large language models and methods of their application to automate work with relational databases.

The subject of the research is methods, architectures and software frameworks that ensure the transformation of natural language queries into SQL queries (Text-to-SQL task) with the required level of accuracy and practical applicability.

The developed solutions (agent system on smolagents, Vanna.AI configuration with BigQuery / PostgreSQL, templates for local deployment of SQLCoder) can be directly applied in the analytical departments of organizations to reduce dependence on qualified SQL developers and speed up the acquisition of analytical data.

Materials and methods.

Task Text-to-SQL: production, benchmarks And methods solutions

to -SQL problem is formally defined as follows: given a relational database schema $S = \{T_1, T_2, \ldots, T_n\}$ (where each T_i includes a table name, a list of columns and their data types, primary and foreign keys) and a user question in natural language Q, it is necessary to generate an SQL query R, the execution of which on this schema returns the correct answer to the question Q (Lewis et. al., 2020).

The complexity of the task is determined by several factors.

The first is the semantic gap: the same meaning can be expressed by many different natural language formulations, and the model must invariantly map them to the same SQL.

The second is the structural complexity of the schema: real corporate databases contain tens and hundreds of tables, thousands of columns, and their meaning is often encoded in abbreviated or uninformative names.

Third, understanding the business context is essential: the query “show top clients” requires knowledge of how a specific organization defines the concept of “top client” (by revenue, by purchase frequency, or other metrics).

At the time of writing, the BIRD (2023) benchmark is the most relevant, as it includes real databases with «polluted» data and requires the model to incorporate external knowledge (e.g., an understanding of industry terminology). Defog ‘s specialized SQL-Eval is crucial because it uses schemas guaranteed not to have been present in the training set of the models being tested, ensuring a fair assessment of generalization ability.

The main quality metric is the accuracy of execution (Execution Accuracy (EX) is the proportion of queries whose results match the benchmark. Unlike Exact Match Accuracy (EM), EX takes into account semantically equivalent queries with different syntax, making it a more realistic estimate.

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

To objectively evaluate the quality of Text-to-SQL systems, several standard benchmarks have been developed (Table 1).

Table 1 – Main benchmarks for evaluating Text- to -SQL systems

Benchmark	Year		Peculiarities	Metrics
WikiSQL	2017	80 654	Simple queries, one table	Execution Accuracy
Spider	2018	10 181	200 databases, complex JOINs, cross- domain	Exact Match + Execution
BIRD	2023	12,751	Real databases, error annotations, external knowledge	Execution Accuracy + Valid Efficiency
SQL-Eval (Defog)	2023	200 (hold-out)	Schemes NOT from training, corporate patterns	% of correctly executed queries

Currently, there are four main types (paradigms) of LLM application for solving the Text- to -SQL problem

1. Zero -shot and few -shot prompting. Basic LLM (GPT-4, Claude, Llama) takes a schema description and a user question as input and then generates SQL in a single pass without any preliminary model modifications. The method is simple to implement and requires no training costs, but is limited in accuracy: 60–82 % on SQL- Eval. The main drawback is the complete dependence on prompting skills and the model's inability to understand the specifics of a particular database.

2. Fine- tuning. The pre-trained base model undergoes an additional training step on a domain-specific dataset of (question → SQL) pairs specific to the target DBMS and business domain. This knowledge is embedded directly into the model's weights, ensuring high inference accuracy. An example is the SQLCoder family of models with an accuracy of up to 93 % on SQL- Eval (Brown et al., 2020). Limitations: the need for significant computing resources at the training stage and the need for retraining when the database schema changes.

3. RAG (Retrieval - Augmented Generation). Instead of changing the model weights, an external knowledge base is formed (a vector storage of DDL descriptions, SQL examples, and business documentation). Upon receiving a request, a semantic search for relevant context is performed, which is dynamically included in the LLM prompt. The method is flexible, allows for prompt knowledge base updates, and does not require retraining. It is implemented in the Vanna.AI framework.

Agent-based systems. LLM acts as the «brain» of the agent, with access to a set of tools: a database schema inspector, an SQL query executor, and a result verifier. The agent iteratively generates queries, executes them, analyzes errors, and corrects the queries until the correct result is obtained. This allows for automatic correction of errors unachievable with single-pass generation. Implemented using the smolagents framework (Vaswani et al., 2017).

A comparative analysis of the four approaches considered by the main parameters is presented in Table 2.

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

Table 1.2 shows that the choice of approach is determined by the characteristics of a specific project. For a stable database schema and high accuracy requirements, retraining is optimal; for a frequently changing schema and the need for corporate deployment, RAG is optimal; in the presence of complex multi-step analytical queries requiring verification, agent systems are optimal. In practice, the most powerful solutions combine several types: for example, Vanna.AI implements RAG on top of a pre-trained LLM (or retrained SQLCoder), and the agent system can use RAG to build the prompt.

Table 2 – Comparative analysis of approaches to solving the Text- to -SQL problem

Criterion	Fine- tuning	RAG	Agent-based approach	Zero shot (basic LLM)
Accuracy	High (93 % for SQLCoder-70B)	High with a good knowledge base	Very high (self-correction)	Average (65-82%)
Deployment cost	High (GPU, training)	Medium (vector storage)	Average (API calls)	Low
Data requirements	Labeled pairs (question→SQL)	DDL, documentation, SQL examples	Tools + database access	Not required
Knowledge updatable	Requires retraining	Knowledge base update	Dynamically	Not updating
Bug fixes	No (generation in 1 pass)	Partial (via RAG)	Yes (auto-verification)	No
Dependency on the database schema	Requires inclusion in prompt	Automatically from a vector database	On-the-fly circuit inspection	Manual transmission
Examples of systems	SQLCoder, DAIL-SQL	Vanna.AI, LlamaIndex	smolagents + SQL tool	GPT-4 (basic)
Optimal application	Stable circuit, high precision requirements	Corporate analytics, a frequently changing scheme	Complex multi-step queries	Prototypes, one-off requests

In addition to the basic types, a number of universal techniques for improving the quality of Text- to -SQL are used:

– Query decomposition (query decomposition): a complex question is broken down into subquestions , each of which is translated into a subquery or CTE (Common Table) Expression).

– Self -consistency: the model generates several SQL variants, executes them and selects a consistent result.

– Schematic coordination (schema linking): Before SQL is generated, explicit linking of the entities mentioned in the question to the tables and columns of the schema is performed.

– Using ensembles: results from multiple LLMs are aggregated to improve reliability.

Thus, the Text- to -SQL problem is a multi-parameter problem requiring the selection and combination of several technical solutions. In our paper, we sequentially

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

explore three of the most practically relevant paradigms— pretraining, RAG, and agent-based systems—implementing and evaluating each.

Technology Retrieval-Augmented Generation (RAG): architecture And application.

Retrieval-Augmented Generation (RAG) is an architectural paradigm in which a pre-trained language model relies not only on the knowledge encoded in its weights to generate a response, but also on the current context dynamically retrieved from an external store. This type was described in detail in the work of Lewis et al. (2020) and has since become a standard engineering approach for building domain-specific systems based on LLM, including Text-to -SQL systems (Devlin et al., 2019). The primary motivation for using RAG in the Text-to -SQL task is the following: even the most powerful general-purpose LLMs lack a priori knowledge of the schema of a specific corporate database—table names, column values, filtering and aggregation business rules. Passing the entire schema to each query (as DDL) is inefficient for large databases (hundreds of tables), as it bloats the context window. RAG solves this problem: for each specific question, only the relevant schema fragments and SQL examples are extracted from the index and compactly included in the prompt.

The fundamental differences between RAG and Fine- tuning are shown in Table 3.

Table 3 – Comparative analysis of the RAG and Fine -tuning approaches for the Text-to -SQL task

Criterion	Fine- tuning	RAG
Principle	Knowledge is embedded into the model weights during the training phase	Knowledge is stored in an external index and retrieved dynamically
Domain Accuracy	Very high (93% for SQLCoder-70B)	High with a quality index (85-92% for Vanna.AI)
Knowledge update	Requires retraining – expensive and time-consuming	Adding/removing documents from the index is instant
Transparency	Low: The source of knowledge is hidden in the scales	High: The model clearly indicates which context it used
Launch cost	High: GPU, data engineering, annotation	Medium: vector database, embedding calculation
Hallucinations	Possible when going beyond the training	Reduced: The model is based on real-world context
Data privacy	The data is included in the training set	Data is stored in a controlled index and is not transferred to LLM.
Optimal application	Stable database design, high accuracy requirements	Frequently changing scheme, corporate analytics, chatbots

Table 3 shows that RAG is not a universal replacement for fine- tuning: when high accuracy requirements are met with a stable schema, specialized fine-tuning is still preferable. However, RAG offers significant advantages in terms of maintenance costs, transparency, and data privacy, making it the primary choice for enterprise analytics systems with frequently updated schemas.

The RAG system architecture includes two independent pipelines: indexing and query.

pipeline (preparatory stage, performed once):

1. Uploading documents: DDL table scripts, text documentation of business rules,

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

examples of SQL queries with annotations.

2. Chunking: The source documents are cut into fragments, each of which will be independently represented by a vector.

3. Embedding computation: each chunk is transformed into a fixed-dimensional vector using the encoder model.

4. Saving to a vector database: vectors and source texts are saved for later retrieval.

pipeline (executed for each user question):

embedding calculation: the user's question is transformed into a vector of the same dimension.

2. Search for nearest neighbors: the vector DB contains the k most semantically similar pieces.

3. Ranking (optional): The found pieces are sorted by relevance using a reranker.

4. Prompt formation: the user's question is combined with the found pieces of context.

5. SQL generation: the prompt is passed to LLM, the model generates the SQL query.

6. Verification: SQL is executed in the database; if an error occurs, a retry cycle is performed (optional).

For prototyping Text- to -SQL systems, we recommend ChromaDB —an embedded DBMS that requires no dedicated server and is the default in Vanna.AI. For production systems with scalability and hybrid search requirements, Qdrant or pgvector are preferred (the latter is especially advantageous with an existing PostgreSQL infrastructure).

Embedding models. Embedding (vector representation) is the mapping of a text fragment to a point in a multidimensional space such that semantically similar texts are geometrically similar. The quality of embeddings directly determines search relevance. For the Text- to-SQL task, it is important that the model understands both the natural language of questions and the SQL syntax of DDL descriptions. It is recommended to use models from the text-embedding-3 series (OpenAI) for cloud solutions and multilingual-e5-large or BGE-M3 for on-premises deployments with Russian language support.

to -SQL tasks, the recommended slicing parameters are: chunk size – 512–1024 tokens; overlap – 10–15 % (64–128 tokens); strategy – «by DDL» for the schema, « recursive « for the documentation. Too little overlap leads to a loss of context at chunk boundaries; too much overlap leads to duplication and increased storage costs.

Limitations of RAG technology include: search quality depends on the quality of the embedding model; degradation due to an incomplete or poorly structured index; difficulty setting up hybrid search and reranking; and the possibility of «losing» relevant information when the retrieval window is too narrow (top -k). For Text- to -SQL tasks, an additional risk is the situation where similar but logically different DDLs compete in search results, misleading the model regarding relationships between tables.

Chunking documents. The choice of slicing strategy critically impacts search quality: chunks that are too small lose context, while chunks that are too large introduce noise. For the Text-to-SQL task, slicing by logical schema units is optimal: one chunk =

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

one table (CREATE TABLE + column comments). SQL query examples are stored in their entirety, along with the query annotation. The main slicing strategies are presented in Table 4.

Table 4 – Strategies for splitting documents into pieces in RAG systems

Strategy	Description	Pros	Cons
By symbols/tokens	The document is divided into fixed-length chunks with overlapping (overlap)	Fast, easy to implement	Can terminate sentences and SQL statements
By delimiters (Recursive)	First it is divided by paragraphs, then by sentences, then by words.	Maintains the integrity of sentences	Uneven pieces
By Markdown / structure	Separators are headings (#, ##); each section is a separate piece	Semantically clean blocks enriched with metadata	For structured documents only
Semantic slicing	Sentences are combined into pieces based on cosine similarity of embeddings	Maximum semantic integrity of the pieces	High computational costs
Sentence Window	Individual sentences are indexed, but a window of N adjacent sentences is passed to LLM	Precise search + rich context	Excessive context on wide window
By DDL/SQL (for Text-to-SQL)	Each CREATE TABLE is a separate chunk; SQL examples are stored in their entirety	Maximum relevance for the SQL generation task	Requires schema preprocessing

Table 5 – Comparison of vector databases for RAG systems

BD	License	Storage	Hybrid search	Metadata filtering	Peculiarities	Application
ChromaDB	Apache 2.0	Disk / memory	No	Yes	Embeddable, no-server mode; ideal for prototyping	Vanna.AI by default
Qdrant	Apache 2.0	Disk + cache	Yes (dense + sparse)	Yes (payload)	High performance, vector quantization	Production, corporate solutions
Weaviate	BSD-3	Disk	Yes (BM25 + vector)	Yes (GraphQL)	GraphQL API, built-in embedding models	Semantic search of documents
Pinecone	Propriet.	Cloud	Yes	Yes	Managed service, serverless pricing	Cloud production systems
pgvector	PostgreSQL		Partially	Yes (SQL WHERE)	PostgreSQL extension – no additional infrastructure	Integration into an existing PostgreSQL stack
FAISS	MIT	Memory	No	No	Meta AI library; maximum speed, no server	Research, Jupyter laptops

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

Vector Databases. The choice of a vector DBMS determines search performance, infrastructure cost, and filtering flexibility. A comparative analysis of the main solutions is presented in Table 5.

Search Enhancement Techniques. Basic semantic search is often insufficient for complex SQL scenarios. In Table 6, we systematize advanced search techniques and evaluate their applicability to Text- to -SQL.

-to -SQL tasks, the most effective combination is a hybrid search (dense + BM25) with Self- Query filtering by table name. Hybrid search ensures that relevant tables are found even with non-standard query formulations, while Self- Query allows for immediate exclusion of irrelevant tables when the query contains explicit references to specific entities.

Table 6 – Techniques for Improving RAG Search for Text- to -SQL Task

Technique	The essence	When to use	Relevance for Text- to -SQL
Semantic search (Dense)	Searching for nearest neighbors by cosine similarity of vectors	Always - Basic Method	High: Finds relevant DDL and SQL examples
BM25 / Full- text (Sparse)	Classic lexical search using TF-IDF	Queries with exact terms, table names	High: Exact column names, SQL keywords
Hybrid	Linear combination of Dense + Sparse results (RRF, α- fusion)	Production systems – the best balance	Very high: combines semantics and precise terms
HyDE	LLM generates a hypothetical answer; its embedding is used to search	Queries that are far from the document style	Medium: Useful when there is a discrepancy between the styles of the query and the DDL
MMR	Maximum edge-relevant diversity: avoids duplication of found pieces	Broad analytical questions	Medium: Useful when there are many similar SQL examples
Context compression	The found pieces are summed up by LLM before being passed to the generating model	Long documents, narrow context window	Low: DDLs are usually compact
Self- Query	LLM extracts structured filters (date, object type) from the question	Queries with explicit attributes	High: Filtering by table name, SQL dialect

Vanna.AI's Text- to -SQL Capabilities and Practical Applications

Vanna.AI is an open-source Python framework (MIT license), a specialized RAG implementation for the Text- to -SQL task. Unlike general-purpose RAG frameworks (LlamaIndex, LangChain), Vanna.AI was designed exclusively for SQL generation, which determines its architectural decisions and user interface (Hu et al., 2022).

RAG system quality assessment metrics. The quality of a RAG pipeline is assessed using a set of metrics covering both intermediate stages (search, ranking) and the final generation quality (Table 7).

For automated evaluation, the RAGAS (Retrieval-Augmented Generation Assessment) framework is recommended. It allows for the calculation of all the listed

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

metrics without the involvement of a human annotator, using the LLM itself as a «judge.» In the Text- to -SQL task, the Execution metric Accuracy is the final and most informative parameter, as it directly reflects the functional correctness of the generated queries.

Table 7 – RAG system quality assessment metrics for Text -to -SQL

Metrics	Definition	Interpretation for Text- to -SQL
Faithfulness	To what extent is the answer based on the context provided, rather than on the model's «fantasy»	Low fidelity → the model ignores DDL and generates SQL from memory, which leads to errors
Answer Relevance (response relevance)	To what extent does the generated SQL answer the question asked?	You can evaluate it automatically: execute SQL and compare the result with the reference one
Context Recall (completeness of context)	The percentage of relevant chunks (table DDL, SQL examples) that were found and included in the prompt	If the required table is not included in the context, the SQL will be incomplete.
Context Precision	The proportion of truly relevant pieces among all found	Noisy context (irrelevant tables) reduces SQL accuracy
Execution Accuracy (precision of execution)	The percentage of SQL queries whose execution result matches the benchmark	Main end-to-end metrics for Text-to-SQL RAG systems

Vanna.AI's key architectural solution is complete modularity: each of the three components (LLM backend, vector storage, and SQL engine) can be independently replaced without changing the rest of the code. This allows the framework to be adapted to any requirements—from cloud enterprise solutions to fully on-premises deployments without sharing data with third parties.

Supported integrations by category are listed in Table 8.

Table 8 – Supported Vanna.AI Integrations

Category	Supported options	Recommendation
LLM backend	OpenAI (GPT-4o, GPT-4), Anthropic (Claude), Google (Gemini), Ollama (Llama 3, Mistral), HuggingFace, any OpenAI - Interoperable API	GPT-4o – for maximum accuracy; Ollama + Llama 3 – for local private deployment
Vector storage	ChromaDB, Qdrant, Pinecone, Weaviate, pgvector, Vanna Cloud	ChromaDB – quick start; Qdrant / pgvector – production with hybrid search
DBMS / data source	PostgreSQL, MySQL, SQLite, BigQuery, Snowflake, DuckDB, MS SQL Server, ClickHouse and others via SQLAlchemy	Connecting via a standard connection string; multiple databases are supported simultaneously
User interface	Flask web app, Streamlit, Jupyter Notebook, Slack bot, REST API	Flask – ready-to-use chat; Streamlit – for quick prototyping
Training data format	DDL scripts (CREATE TABLE), SQL examples (question + SQL pairs), text documentation	Minimum: 5-10 DDL + 10-20 SQL pairs for basic work; 50+ pairs for production quality

Working with Vanna.AI is based on two consecutive phases: training and query.

The training phase does not involve changing the LLM weights—the term is used in the sense of «filling the vector index.» Three types of training data are loaded into the index:

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

1. DDL descriptions (CREATE TABLE with comments on columns and keys): provide the model with an understanding of the database structure.

2. Text documentation (business rules, glossaries): explains semantics that are not available from the schema (e.g., a category A customer is one who has made more than 10 orders).

3. SQL examples (question→SQL pairs): the most valuable data type; allows the model to learn from successful historical experience.

The corrected query was verified directly in BigQuery and returned correct results, confirming the reliability of the Auto Fix mechanism for correcting common dialect-specific SQL errors.

Feedback and learning mechanism based on successful queries. After each successful SQL query, Vanna.AI prompts you to save the (query → SQL) pair to a vector index. This enables continuous learning: the more successful queries accumulate, the more accurate the system becomes for a specific schema and business domain. Automatic saving can be enabled in the settings (auto_train = True), eliminating the need for manual confirmation.

New Question » to manage the session.

A comparison of Vanna.AI with alternative Python frameworks for Text -to -SQL is given in Table 9.

Table 9 – Comparison of Vanna.AI with alternative Text- to -SQL tools

Criterion	Vanna.AI	LlamaIndex + SQL	LangChain SQL Agent	Defog (cloud)
Open- source	Yes (MIT)	Yes (MIT)	Yes (MIT)	No (SaaS)
SQL Specialization	Full	Partial	Partial	Full
Built-in UI	Yes (Flask, Streamlit)	No	No	Yes (cloud)
Automatic visualization	Yes (Plotly)	No	No	Yes
Feedback mechanism	Yes (auto-save correct pairs)	No	No	Yes
SQL Auto-Correction	Auto Fix button	No	Through an agent	Yes
On-premises deployment	Yes (Ollama + ChromaDB)	Yes	Yes	No
Difficulty of setup	Low (3 lines of code)	Average	High	Very low (SaaS)

Table 9 shows that Vanna.AI occupies the niche of the most ready-to-use open-source solution: with minimal setup complexity, it provides a built-in UI, automatic visualization, and a feedback mechanism—features that are missing from the universal LlamaIndex and LangChain frameworks.

Despite significant practical advantages, a number of limitations of the framework must be taken into account:

– Quality depends on the training data: without a sufficient number of correct SQL examples (less than 10-15 pairs), the accuracy of the system will be comparable to the basic

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

LLM without RAG.

– No guarantees of correctness: Vanna.AI does not formally verify the generated SQL – verification occurs only through actual execution in the database.

– Complex analytical queries with multi-level subqueries, recursive CTEs, or specific window functions may require multiple Auto Fix iterations or manual intervention.

– Cloud Privacy: When using Vanna In cloud or cloud LLM (GPT-4, Claude API), database schema fragments and user questions are transmitted to third-party servers. For sensitive data, local deployment is required (Ollama + ChromaDB).

Conclusion.

RAG technology is an architecturally elegant solution to the Text- to -SQL problem: it combines the power of pre-trained LLMs with a dynamically updated knowledge base of the specific database schema. Comparative analysis shows that RAG outperforms Fine-tuning in flexibility, transparency, and maintenance costs, while achieving slightly lower maximum accuracy.

The Vanna.AI framework is the most mature open-source implementation of this approach: its modular architecture ensures compatibility with any LLM, vector DBMS, and SQL database; a built-in Flask interface, automatic Plotly visualization, and an automatic error correction mechanism make the system ready for industrial use. Practical testing on a dataset Orders (BigQuery) demonstrated correct SQL generation for RFM analysis, YoY comparisons, and ABC-XYZ segmentation tasks. The identified limitations include the dependence of quality on the volume of training data and the need for local deployment under strict privacy requirements, which prompted the study of additional self-correction mechanisms, namely agent-based systems.

Agent-based text-to-SQL conversion system with automatic error correction

An agent system in the context of LLM application is an architecture in which the language model acts not simply as a text generator, but as a "thinking" orchestrator, capable of planning a sequence of actions, calling external tools, interpreting their results, and iteratively adjusting its behavior to achieve a set goal (Dettmers et al., 2023). In the Text- to -SQL task, the agent is able not only to generate SQL, but also to immediately check its correctness through actual execution in the database – and, if necessary, correct errors independently.

To implement the agent system in our work, we use the smolagents framework developed by HuggingFace (Yao et. al., 2023).

The framework features a minimalist design: the agent works through Python code generation rather than through JSON tool call functions, which provides maximum flexibility – the agent can combine the results of several tools in arbitrary expressions.

The fundamental advantage of the agent-based approach over a standard single-pass pipeline is the following: a standard pipeline lacks a verification mechanism—an SQL query may be syntactically correct and execute without errors, but return semantically incorrect data (for example, calculating revenue without accounting for returns). An agent can detect and correct such situations through critical analysis of the results. A comparative analysis of the two approaches is presented in Table 10.

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

Table 10 – Comparison of the standard Text- to -SQL pipeline and the agent system

Criterion	Standard Text- to -SQL pipeline	Agent system
Generation architecture	Single-pass (query → SQL → result)	Iterative (query → SQL → execution → verification → adjustment)
SQL Error Handling	No: The error is returned to the user.	Automatic: The agent analyzes the error message and generates a corrected request
Verification of the correctness of the result	No: SQL may execute without errors but return invalid data.	Partial: the agent analyzes the result for meaningfulness (number of rows, data types)
Working with multiple tables	Requires a full description of all JOINs in the prompt	The agent independently inspects the schema and builds JOINs through tools
Decomposition of complex issues	No	Yes: The agent breaks down a complex question into substeps
Request cost (tokens)	Low (1 LLM challenge)	High (2-5+ LLM calls per iteration)
Delayed response	Minimum (seconds)	Increased (10-60 sec with several iterations)
Optimal application	Simple queries, high load, real time	Complex analytical queries, batch processing, intolerance to errors

An important feature is automatic error handling in scenarios with incorrect SQL. If, in step 2 of the previous version of the agent (with a less powerful model), a query was generated with an SQLite syntax error —ORDER BY after LIMIT—the agent didn't stop, but read the error message and corrected the query in step 3, removing the LIMIT clause before ORDER BY. This is what distinguishes the agent-based approach: error handling is built into the generation cycle, rather than delegated to the user.

When deployed locally, LLM runs through Ollama, a tool that provides an OpenAI -compatible API for local models. Integration with smolagents is accomplished through the LiteLLMModel adapter:

Basic security recommendations for agent systems. An agent with access to the SQL engine is potentially dangerous: LLM can generate a destructive query (DROP TABLE, DELETE without a WHERE clause). The following precautions must be observed:

1. Use a connection with read-only rights (READ ONLY) for the user on whose behalf the agent runs.

2. Restrict the types of SQL statements allowed (SELECT only) at the tool level – add validation before execution.

3. Set a query execution timeout (statement_timeout) to prevent freezing on heavy queries.

4. Use the max_steps parameter to limit the number of agent iterations and prevent infinite loops.

Thus, the agent-based system based on smolagents provides a qualitatively new level of Text- to -SQL reliability: automatic result verification, iterative self-correction, and the ability to decompose complex queries—features unachievable in single-pass pipelines. The tradeoff—increased cost and iteration latency—is acceptable for analytical tasks that do not require real-time responses.

To clearly display the logic of the agent's operation – including the self-correction cycle – we will construct a block diagram (Figure 1).

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

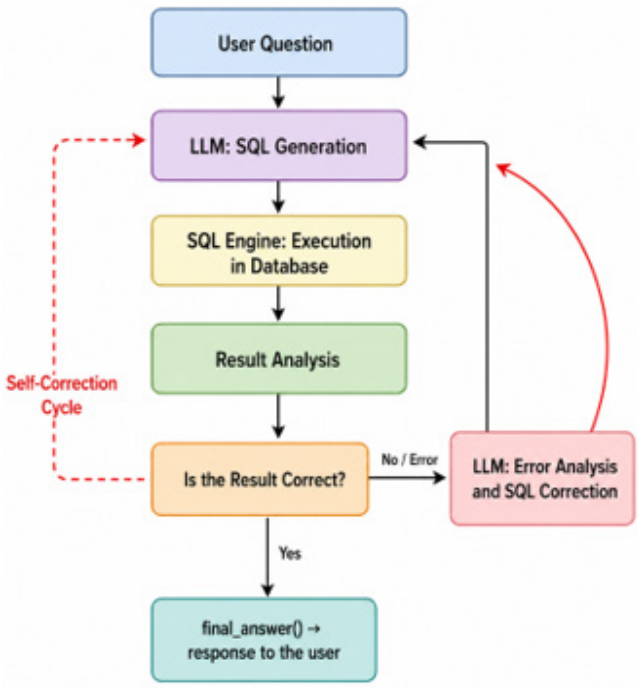


Fig. 1. Workflow Cycle of the Text-to-SQL Agent System

Local agent system deployment configurations. In scenarios where data privacy is important, the agent system can be deployed entirely locally, without forwarding any requests to external servers. Table 11 summarizes the available configurations, from minimal to production.

Table 11 – Configurations for local deployment of the Text- to -SQL agent system

Configuration	LLM	Vector. Storage	Min. requirements	Application
Minimum	Ollama + Llama 3.2 3B	ChromaDB (in-memory)	8GB RAM, no GPU	Demos, prototypes, testing
Basic	Ollama + Llama 3.1 8B	ChromaDB (persistent)	16 GB RAM / 8 GB VRAM	Small teams, private data
Recommended	Ollama + Llama 3.3 70B (4-bit)	Qdrant (persistent)	64 GB RAM / 40 GB VRAM	Corporate analytics service
Maximum	SQLCoder-70B (float16)	Qdrant + pgvector	80+ GB of VRAM (2× A100)	High-load production

Conclusion. The smolagents -based agent system implements a fundamentally different approach to the Text -to -SQL task compared to a single- pass pipeline: iterative SQL generation and verification automatically correct both syntax errors and semantically incorrect queries. A practical implementation with two tables (receipts, waiters) confirmed the functionality of the self-correction mechanism: when an SQLite error occurred, the agent automatically corrected the query without user intervention.

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

Discussion.

The results demonstrate that the Text- to -SQL problem cannot be effectively solved using a single, universal approach. The analysis confirmed that each of the technologies considered—fine- tuning, RAG, and agent-based systems—has its own area of rational application. The article has already demonstrated that fine- tuning ensures high accuracy with a stable database schema, RAG increases flexibility and knowledge refreshability, and agent-based systems enable iterative validation and automatic error correction of SQL queries.

The most significant result of the study is the justification for a combined approach that uses large language models in conjunction with search-augmented generation and agent-based self-correction. This approach partially overcomes the main limitations of basic LLMs: dependence on prompt quality, lack of knowledge of the specific database schema, and the risk of generating syntactically correct but semantically incorrect SQL queries. Unlike single-pass generation, agent-based architecture includes query execution, result analysis, and error correction, increasing the system's reliability when working with complex analytical queries.

A comparison of approaches shows that fine- tuning is feasible in a stable subject area and with labeled query-SQL pairs. However, in a corporate environment, database schemas frequently change, with new tables, attributes, and business rules added. Under such conditions, retraining the model becomes costly. Therefore, RAG is a more flexible solution: knowledge of the database structure, DDL descriptions, sample SQL queries, and business documentation are stored in an external index and can be quickly updated without changing the model's weights. This is especially important for organizations that require constant updating of analytical data.

However, RAG also has limitations. The quality of the response directly depends on the completeness and quality of the vector index, the document partitioning strategy, the choice of embedding model, and the top -k extraction parameter. If the relevant table or the required SQL example is not included in the context, the model may generate an incomplete or erroneous query. Furthermore, if similar tables or ambiguous column names are present, there is a risk of selecting irrelevant context. Therefore, the effectiveness of a RAG system is determined not only by the quality of the LLM but also by the quality of the knowledge base preparation.

The agent-based approach discussed in this article complements RAG with a self-checking mechanism. The agent-based system not only generates an SQL query but also executes it, analyzes errors, and generates a corrected version of the query. The article notes that the smolagents -based agent system differs from a standard pipeline in that it uses an iterative process: query → SQL → execution → verification → correction. This is especially important for multi-step analytical tasks that require working with multiple tables, JOIN operations, grouping, filtering, and qualifying conditions.

However, the use of agent-based systems comes with increased computational costs and response latency. While standard generation requires a single LLM call, an agent-based cycle can involve multiple iterations, increasing token consumption and processing

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

time. Therefore, the agent-based approach is most appropriate not for simple one-off queries, but for complex analytical scenarios where result reliability is more important than minimal response latency.

Security requires special attention. Since the agent accesses the SQL engine, there is a risk of generating unwanted or potentially dangerous queries. Therefore, the article rightly suggests security measures: using read-only permissions, limiting valid SQL statements to the SELECT command, setting execution timeouts, and limiting the number of agent iterations. These measures are mandatory when implementing Text- to -SQL systems in a corporate environment.

Thus, the study's results confirm that the most promising approach lies not in the isolated use of fine- tuning, RAG, or agent-based systems, but in their integration into a single hybrid architecture. In such an architecture, fine- tuning can provide the model's basic linguistic and SQL competence, RAG provides the relevant context of a specific database, and the agent-based mechanism verifies and corrects the result. This approach improves the adaptability, reliability, and practical applicability of intelligent Text- to -SQL systems in corporate analytics.

Conclusion.

This paper explores and tests methods for using large language models to automate SQL data analysis. All stated objectives were achieved and all tasks outlined in the introduction were solved.

A systemic analysis of the Text- to -SQL problem is conducted: a formal formulation is provided, the evolution of benchmarks (WikiSQL, Spider, BIRD, SQL-Eval) is considered, and the comparative advantages of four solution types are identified. The main conclusion: there is no universally best approach – the choice is determined by the specific project requirements for accuracy, flexibility, cost, and data privacy. The LLM fine-tuning technology is studied in detail as applied to Text -to -SQL; it is shown that the specialized SQLCoder-70B model, with an accuracy of 93 % on SQL- Eval, outperforms GPT-4 (82 %) and Claude 2 (65 %) when working with unfamiliar schemas, demonstrating the fundamental advantage of domain-specific fine- tuning over general LLMs.

RAG technology was explored as an alternative to fine-tuning for dynamic corporate environments. All components of the RAG pipeline were analyzed: document slicing strategies, vector databases (ChromaDB, Qdrant, pgvector), embedding models, and search enhancement techniques (hybrid search, Self-Query, MMR). Practical testing of the Vanna.AI framework on a real dataset was conducted. Orders in BigQuery confirmed the approach's effectiveness for typical business analytics tasks: RFM analysis, YoY comparisons, and ABC-XYZ segmentation. The built-in Auto Fix mechanism demonstrated the system's ability to correct SQL errors without user intervention. It was established that the performance of the RAG system is critically dependent on the volume and quality of the training data in the vector index.

An agent-based text-to-SQL conversion system based on the smolagents framework was developed. A comparative analysis of the agent-based approach and a standard pipeline showed that the agent outperforms single-pass generation in terms of

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

reliability (iterative verification), the ability to handle complex multi-table queries, and automatic error correction, albeit at the expense of higher token costs and response latency.

Based on the conducted research, the following practical recommendations for choosing a paradigm were formulated:

- Fine- tuning (SQLCoder): if you have a stable database schema, high accuracy requirements, and are ready to invest in GPU infrastructure and data labeling.
- RAG (Vanna.AI): with a frequently changing scheme, the need for a corporate chatbot for analysts, a lack of resources for additional training, and a priority for a quick start.
- Agent system (smolagents): for complex multi-step analytical queries requiring decomposition and verification, in batch mode or when response latency is acceptable.
- Combined approach: for maximum quality – RAG on top of pre-trained SQLCoder as LLM backend, with agent-based verification loop.

Overall, the practical significance is determined by the fact that the developed solutions—Vanna.AI configurations, the agent system on smolagents , and local deployment templates—can be directly applied in the analytical departments of organizations without significant adaptation.

The following are promising areas for further research:

- integration of RAG and agent-based approaches into a single pipeline with automatic strategy selection depending on the complexity of the request;
- development of specialized Russian-language training datasets for the Text- to -SQL task;
- study of the application of multi-agent systems, where different agents specialize in different types of SQL queries;
- Evaluation of the impact of DDL documentation quality (completeness of column comments) on the accuracy of RAG systems in real corporate schemes.

REFERENCES

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems. - Curran Associates. Vol. 33. Pp. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165 [In Eng.].

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT // Association for Computational Linguistics. Vol. 1. Pp. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805 [In Eng.].

Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems. - Curran Associates. Vol. 36. https://doi.org/10.48550/arXiv.2305.14314 [In Eng.].

Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations. — OpenReview. https://doi.org/10.48550/arXiv.2106.09685 [In Eng.].

Li J., Hui B., Qu G., Yang J., Li B., Wang B., Qin B., Cao R., Geng R., Huo N., Zhou X., Ma C., Li G., Chang K.C.C., Huang F., Cheng R., Li Y. (2023). Can LLM Already Serve as A Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-SQLs. arXiv preprint // Cornell University. arXiv:2305.03111. https://doi.org/10.48550/arXiv.2305.03111 [In Eng.].

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.T., Rocktäschel T., Riedel S., Kiela D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural

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

Information Processing Systems // Curran Associates. Vol. 33. Pp. 9459–9474. https://doi.org/10.48550/arXiv.2005.11401 [In Eng.].

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.T., Rocktäschel T., Riedel S., Kiela D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. - Curran Associates. Vol. 33. Pp. 9459–9474. - https://doi.org/10.48550/arXiv.2005.11401 [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. Curran Associates. Vol. 30. Pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762 [In Eng.].

Yu T., Zhang R., Yang K., Yasunaga M., Wang D., Li Z., Ma J., Li I., Yao Q., Roman S., Zhang Z., Radev D.R. (2018). Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing // Association for Computational Linguistics. Pp. 3911–3921. https://doi.org/10.18653/v1/D18-1425 [In Eng.].

Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models // International Conference on Learning Representations. — OpenReview. https://doi.org/10.48550/arXiv.2210.03629 [In Eng.].

Zhong V., Xiong, C., Socher R. (2017). Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. arXiv preprint // Cornell University. arXiv:1709.00103. https://doi.org/10.48550/arXiv.1709.00103 [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.2. Number 26 (2026). Pp. 206–218
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.013

DESIGN AND IMPLEMENTATION OF A WEB-BASED PROTOTYPE FOR CARDIOLOGICAL IMAGE MANAGEMENT

M. Kabdullin[1]*, A. Kairbekov[1], A. Kabdullin[1], L. Naizabayeva[2], A. Mukhit[3]
[1]Satbayev University, Almaty, Kazakhstan;
[2]International Information Technology University, Almaty, Kazakhstan;
[3]Kazakhstan's Medical University "KSPH", Almaty, Kazakhstan.
E-mail: m.kabdullin@satbayev.university

Maxat Kabdullin — Master of Technical Sciences, Senior Lecturer, Department of Information Systems, Institute of Automation and Information Technologies, Satbayev University, Principal Investigator of research project No. AP25794129, Almaty, Kazakhstan
E-mail: m.kabdullin@satbayev.university, https://orcid.org/0000-0002-7320-6259;
Abylai Kairbekov — Master of Technical Sciences, Senior Lecturer, Department of Information Systems, Institute of Automation and Information Technologies, Satbayev University, Almaty, Kazakhstan
https://orcid.org/0000-0001-7123-0901;
Azat Kabdullin — Master of Technical Sciences, Lecturer, Department of Information Systems, Institute of Automation and Information Technologies, Satbayev University, Almaty, Kazakhstan
https://orcid.org/0000-0001-5589-2542;
Lyazat Naizabayeva — Doctor of Technical Sciences, Research Professor, Department of Information Systems, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-4860-7376;
Anar Mukhit — Master's Student, Kazakhstan's Medical University "KSPH", Almaty, Kazakhstan
https://orcid.org/0009-0003-2471-0365.

© M. Kabdullin, A. Kairbekov, A. Kabdullin, L. Naizabayeva, A. Mukhit

Abstract. This paper presents the development of a web-based prototype for secure storage and preprocessing of cardiological medical images. The system includes modules for uploading DICOM files, performing basic preprocessing, and securely storing patient data with role-based access control. Built with the Django framework, the application follows the Model–View–Template architecture to ensure modularity and scalability. The prototype was tested using open medical datasets and demonstrated stable performance in scenarios typical for diagnostic workflows. The system addresses both technical and

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

ethical requirements for medical data handling, aiming to support digital transformation in healthcare.

Keywords: medical imaging, DICOM, web application, data security, cardiology

For citation: Kabdullin M., Kairbekov A., Kabdullin A., Naizabayeva L., Mukhit A. (2026). Design and implementation of a web-based prototype for cardiological image management // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 206–218. https://doi. org/10.54309/IJICT.2026.26.2.013. (In Eng.).

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

Acknowledgements. *This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant: №AP25794129 "Development of an Algorithm for Filtering and Preprocessing Biomedical Images for Cardiology Diagnostics".*

КАРДИОЛОГИЯЛЫҚ КЕСКІНДЕРДІ БАСҚАРУҒА АРНАЛҒАН ВЕБ-НЕГІЗДЕЛГЕН ПРОТОТИПТІ ӘЗІРЛЕУ ЖӘНЕ ІСКЕ АСЫРУ

М. Кабдуллин[1]*, А. Каирбеков[1], А. Кабдуллин[1], Л. Найзабаева[2], А. Мухит[3]

[1]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан;

[2]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;

[3]Қазақстанның «ҚШПХ» медициналық университеті, Алматы, Қазақстан.

E-mail: m.kabdullin@satbayev.university

Кабдуллин Максат — техникалық ғылымдар магистрі, Ақпараттық жүйелер кафедрасының, Автоматтандыру және ақпараттық технологиялар институтының аға оқытушысы, № AP25794129 ғылыми жобасының жетекші зерттеушісі, Қ. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
E-mail: m.kabdullin@satbayev.university, https://orcid.org/0000-0002-7320-6259;

Каирбеков Абылай — техникалық ғылымдар магистрі, Ақпараттық жүйелер кафедрасының, Автоматтандыру және ақпараттық технологиялар институтының аға оқытушысы, Қ. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-7123-0901;

Кабдуллин Азат — техникалық ғылымдар магистрі, Ақпараттық жүйелер кафедрасының, Автоматтандыру және ақпараттық технологиялар институтының оқытушысы, Қ. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
https://orcid.org/0000-0001-5589-2542;

Найзабаева Лязат — техникалық ғылымдар докторы, Ақпараттық жүйелер кафедрасының зерттеуші-профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-4860-7376;

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

Мухит Анар — магистрант, Қазақстанның «ҚШПХ» медициналық университеті, Алматы, Қазақстан
https://orcid.org/0009-0003-2471-0365.

© М. Кабдуллин, А. Каирбеков, А. Кабдуллин, Л. Найзабаева, А. Мухит

Аннотация. Мақалада кардиологиялық медициналық кескіндерді қауіпсіз сақтау және алдын ала өңдеуге арналған веб-негізделген прототипті әзірлеу мәселесі қарастырылады. Жүйе DICOM файлдарын жүктеу, бастапқы алдын ала өңдеуді орындау және рөлдік қолжетімділікті басқару арқылы пациенттердің деректерін қауіпсіз сақтау модульдерін қамтиды. Қосымша Django фреймворкі негізінде әзірленіп, жүйенің модульділігі мен масштабталуын қамтамасыз ету мақсатында Model–View–Template архитектурасын қолданады. Прототип ашық медициналық деректер жиынтықтарын пайдалана отырып тестіленіп, диагностикалық жұмыс үдерістеріне тән сценарийлерде тұрақты жұмыс өнімділігін көрсетті. Жүйе медициналық деректерді өңдеуге қойылатын техникалық және этикалық талаптарды ескеріп, денсаулық сақтау саласының цифрлық трансформациясын қолдауға бағытталған.

Түйінді сөздер: медициналық кескіндерді визуализациялау, DICOM, веб-қосымша, деректер қауіпсіздігі, кардиология

Дәйексөздер үшін: Кабдуллин М., Каирбеков А., Кабдуллин А., Найзабаева Л., Мухит А. (2026). Кардиологиялық кескіндерді басқаруға арналған веб-негізделген прототипті әзірлеу және іске асыру // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 206-218. https://doi.org/10.54309/ IJICT.2026.2.013. (Ағыл. тіл.).

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

Алғыс. *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (грант № AP25794129 «Кардиологиялық диагностикаға арналған биомедициналық кескіндерді сүзгілеу және алдын ала өңдеу алгоритмін әзірлеу»).*

РАЗРАБОТКА И РЕАЛИЗАЦИЯ ВЕБ-ПРОТОТИПА ДЛЯ УПРАВЛЕНИЯ КАРДИОЛОГИЧЕСКИМИ ИЗОБРАЖЕНИЯМИ

М. Кабдуллин[1]*, А. Каирбеков[1], А. Кабдуллин[1], Л. Найзабаева[2], А. Мухит[3]
[1]Казахский национальный технический университет им. К. Сатпаева, Алматы, Казахстан;
[2]Международный университет информационных технологий, Алматы, Казахстан;
[3]Казахстанский медицинский университет «КШПХ», Алматы, Казахстан.
E-mail: m.kabdullin@satbayev.university

Кабдуллин Максат — магистр технических наук, старший преподаватель кафедры

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

информационных систем Института автоматизации и информационных технологий Казахский национальный технический университет им. К. Сатпаева, главный исследователь научного проекта № AP25794129, Алматы, Казахстан
E-mail: m.kabdullin@satbayev.university, https://orcid.org/0000-0002-7320-6259;
Каирбеков Абылай — магистр технических наук, старший преподаватель кафедры информационных систем Института автоматизации и информационных технологий Казахский национальный технический университет им. К. Сатпаева, Алматы, Казахстан
https://orcid.org/0000-0001-7123-0901;
Кабдуллин Азат — магистр технических наук, преподаватель кафедры информационных систем Института автоматизации и информационных технологий Казахский национальный технический университет им. К. Сатпаева, Алматы, Казахстан
https://orcid.org/0000-0001-5589-2542;
Найзабаева Лязат — доктор технических наук, профессор-исследователь кафедры информационных систем Международного университета информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-4860-7376;
Мухит Анар — магистрант, Казахстанского медицинского университета «КШПХ», Алматы, Казахстан
https://orcid.org/0009-0003-2471-0365.

© М. Кабдуллин, А. Каирбеков, А. Кабдуллин, Л. Найзабаева, А. Мухит

Аннотация. В статье представлена разработка веб-ориентированного прототипа для безопасного хранения и предварительной обработки медицинских изображений в кардиологии. Система включает модули загрузки DICOM-файлов, выполнения базовой предварительной обработки и безопасного хранения данных пациентов с использованием ролевого управления доступом. Приложение разработано на основе фреймворка Django с применением архитектуры Model–View–Template, обеспечивающей модульность и масштабируемость системы. Прототип был протестирован с использованием открытых медицинских наборов данных и продемонстрировал стабильную производительность в сценариях, характерных для диагностических рабочих процессов. Разработанная система учитывает как технические, так и этические требования к обработке медицинских данных и направлена на поддержку цифровой трансформации в сфере здравоохранения.

Ключевые слова: медицинская визуализация, DICOM, веб-приложение, безопасность данных, кардиология

Для цитирования: Кабдуллин М., Каирбеков А., Кабдуллин А., Найзабаева Л., Мухит А. (2026). Разработка и реализация веб-прототипа для управления кардиологическими изображениями // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 206-218. https://doi.org/10.54309/IJICT.2026.2.013. (На англ.).

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

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность. *Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № AP25794129 «Разработка алгоритма фильтрации и предварительной обработки биомедицинских изображений для кардиологической диагностики»).*

Introduction.

In the age of digitalization, healthcare requires the processing and storage of large volumes of medical data, including biomedical images. Images obtained during cardiac examinations, such as echocardiography, cardiac computed tomography, and magnetic resonance imaging, are key sources of data for accurate diagnosis. Modern diagnostics of cardiovascular disease, one of the leading causes of death worldwide (WHO, 2023), largely depends on the quality of recognition and accessibility of image data.

Before recognizing and storing large volumes of data, it is necessary to develop a specialized web application that combines secure storage and automatic format conversion, ensuring compatibility with analysis and visualization software, and pre-processing of biomedical images for deep recognition and subsequent diagnostics. The use of web technologies in this context enables cross-platform support, remote data access, and integration with external expert systems. Information security and compliance with biomedical data privacy requirements must be considered during development.

The main goal of this paper is to design and implement of a web application for a research project founded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant: №AP25794129 “Development of an algorithm for filtering and pre-processing biomedical images for cardiac diagnostics”, which provides secure storage, automatic format conversion and pre-processing of biomedical images.

This work addresses the following tasks: formulating requirements for the functionality, security, and architecture of a web application; designing its structure in accordance with medical imaging standards; implementing modules for loading, storing, viewing, and pre-processing images; integrating basic information security tools; testing and assessing the compliance of the developed prototype with the stated requirements.

The relevance of this study stems not only from the pre-processing of large biomedical data flows and automated format conversion to a unified standard, but also from the growing threats to information security in the medical field. According to the 2023 HIMSS report, approximately 80% of healthcare institutions have experienced incidents involving data compromise, including vulnerabilities in image storage and transmission systems (HIMSS, 2023). Some studies highlight that many PACS systems still lack modern cryptographic protection, making them vulnerable to unauthorized access (Eichelberg et al., 2004). In the context of the digital transformation of healthcare, there is a growing need for distributed, secure, and easy-to-use solutions that integrate the functions of storing, pre-processing, and visualizing medical images—which forms the subject area of this study.

Cardiology diagnostics rely on various types of cardiac and vascular imaging. All

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

these modalities generate digital images, often serially (dynamic «cine» series showing cardiac motion). The generally accepted standard for storing and exchanging such images is the DICOM format (Kumar et al., 2024). The DICOM format describes both image pixel data and extensive metadata (patient information, examination parameters, image geometry, etc.), ensuring reproducibility and unambiguous interpretation of medical images across different systems. As a result, all leading medical equipment manufacturers support DICOM, which enables the storage, transmission, and viewing of cardiac images (CT, MRI, angiography, etc.) in a single, standardized format (Eichelberg et al., 2004; Kumar et al., 2024).

Medical images contain personal (and often highly sensitive) patient data, so their processing is subject to numerous standards and regulations regarding personal data protection. Development must consider international and local standards in the Republic of Kazakhstan.

The development of a web application for our study must encompass a set of standards (international and national). To this end, we will formulate a list of specific information security requirements for a web application for medical images:

- Strict user authentication and authorization based on least privilege must be implemented. Patient data must be stored in encrypted form, and internet transmission must be carried out only through secure channels;
- Logging and monitoring: the system must maintain a log of access to each image and be able to detect anomalies;
- Backup and the ability to restore data without loss are essential for availability (Raza et al., 2024);
- Strict confidentiality.

Compliance with local laws and ethical standards is an important aspect of development. According to the Law of the Republic of Kazakhstan «On Personal Data and Their Protection,» medical images associated with identifiable individuals are considered personal data and require special protection measures and consent for their processing (Republic of Kazakhstan, 2013). In clinical and research contexts, informed patient consent is often used, especially if the images are used not only for treatment but also for scientific purposes (Republic of Kazakhstan, 2013; DeMartini et al., 2023; WHO, 2021). International practice also emphasizes that the use of medical images requires clear consensus forms, especially if they are published or transferred to third-party systems (Mazurowski et al., 2020).

Taking these requirements into account during development will protect the system from typical threats and create the preconditions for successful implementation in medical institutions.

Based on an analysis of the subject area and identified trends in medical imaging, the key objectives addressed in this study were formulated. First, an analysis of cardiac imaging characteristics, their use cases, and user requirements was conducted, including regulatory and ethical considerations. A critical review of existing PACS systems and DICOM viewers (both commercial and open-source) was conducted, focusing on

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

their functionality, security level, and limitations, including the lack of web access and integration difficulties.

Based on the data obtained, a web application architecture was developed, including secure storage, a pre-processing module, and a web interface. A prototype was implemented with functions for loading, storing, and viewing cardiac images, and basic information security measures were implemented. Testing was conducted using an open dataset to assess the system's functional suitability. Technical documentation was also developed, including a description of the architecture, database model, and implemented security mechanisms.

Completing these tasks will enable us to achieve our research goal: creating a prototype web application capable of securely storing and processing cardiac biomedical images. This application should meet both the practical needs of physicians (convenient access to images, basic processing tools) and strict ethical requirements. Information security will also be improved due to the storage of user data. Implementation of the project will contribute to the solution of the problem of securely storing growing volumes of medical images and will enable data preprocessing directly in a secure environment, which is essential for healthcare systems in the era of digital transformation.

Materials and methods.

The web application is built on the Django framework, which utilizes the Model-View-Template (MVT) design pattern – a variation of the MVC architecture. In accordance with MVT, Django implements a separation of three key components: data models, views, and user interface templates. This approach helps maintain a clear structure and scalability of the project by separating data processing logic, application business logic, and the presentation layer.

The project's architecture is organized according to a modular principle: functionality is divided into several Django applications. Each application is responsible for its own domain: HOME_AUTH (user management and homepage), PATIENTS (patient data management), and MEDICAL_IMAGES (medical image uploading and processing). This approach aligns with the recommended Django project structure, where models are grouped into applications based on their logical role (Django Software Foundation, 2024).

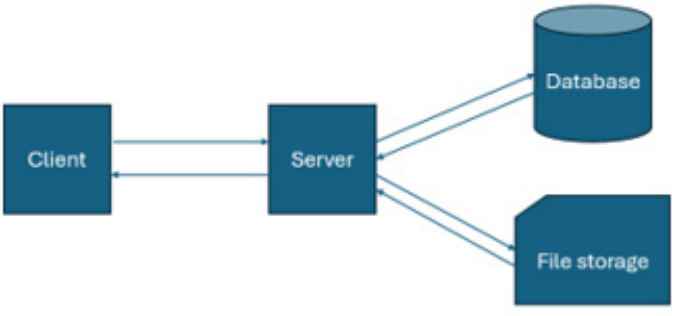


Fig. 1. High-level architecture of the application.

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

Django's built-in mechanism is used for URL routing: each path is associated with a function or view class that handles the request. Access to key sections of the application is restricted to unregistered users using the @login_required decorator, which redirects all unauthenticated users to the login page. This ensures that functionality for working with sensitive medical data is available only to authorized users. Additionally, access rights are differentiated using Django mechanisms: administrator rights (superuser) for full access, staff rights for access to the admin panel, and regular medical professional accounts with limited rights. Django provides a built-in permissions system that allows you to set permissions for operations on specific models (view, add, edit, delete) and verify these permissions through the authorization system. Thus, the application architecture provides role-based access control: access to certain actions (such as deleting or editing records) is granted only to users with the appropriate permissions or roles.

Django's built-in authentication system is used for authentication. When a user logs in, a session is created, and Django associates it with the user object, storing the user ID in session data. Logout is performed by calling the appropriate framework method, which completely clears the session data. Using a ready-made mechanism ensures secure account management without the need for manual implementation of secure password storage or session management (Django Software Foundation, 2024).

The web application's user interface is designed to be intuitive and functional, providing convenient access to medical images and patient data. The interface is divided into several logical modules: an authentication screen, the main work panel after login, and navigation and data filtering elements. All interactions are handled via secure HTTPS requests to the server application, which stores and processes images and patient information.

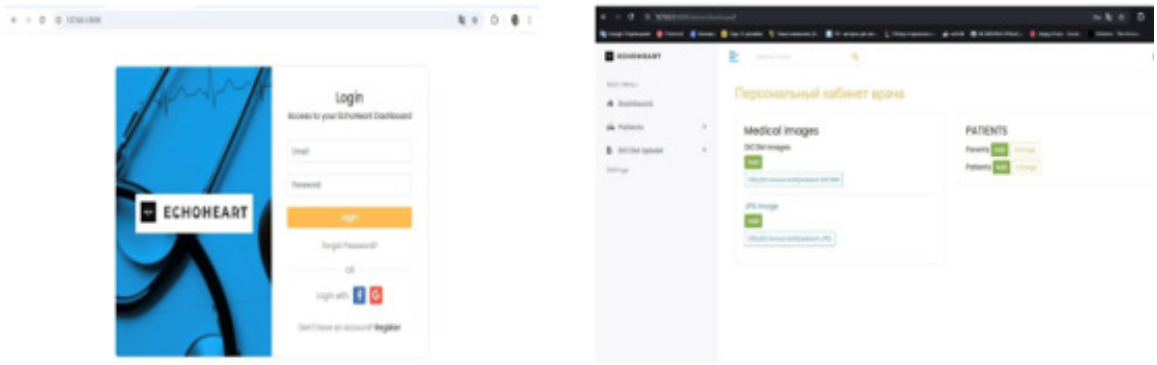


Fig. 2. Login Page and profile page.

The start page begins with authorization. For security, all forms are protected with CSRF tokens and data validation, as shown in Figure 2: Login Page

The «DICOM Upload» section is responsible for uploading medical images to the database. On the server, preliminary verification occurs, a SHA-256 checksum is calculated to ensure integrity and prevent tampering, and then the images are saved to

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

the data storage. The list of supported formats depends on the DICOM standard—an open standard for storing and transmitting medical images. The interface can display upload progress and statistics on processed files. Once uploaded, the images become available in the corresponding patient's profile: the user can switch to a specific patient's section and see previews or full-size image views, as shown in Figure 3.

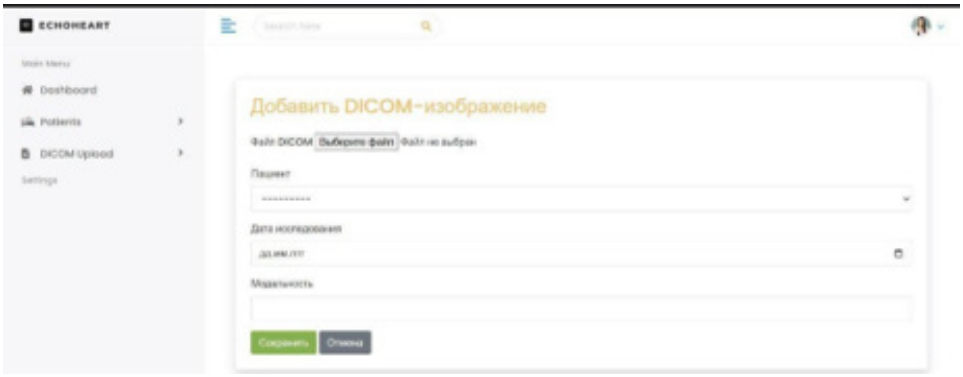


Fig. 3. Image Loading.

The viewing section displays a list of the patient's uploaded images, for example, as cards with the examination date and file type. Files can be sorted by patient name as shown in Figure 4. This functionality provides ease of working with large numbers of images, as described in studies on medical image storage systems.

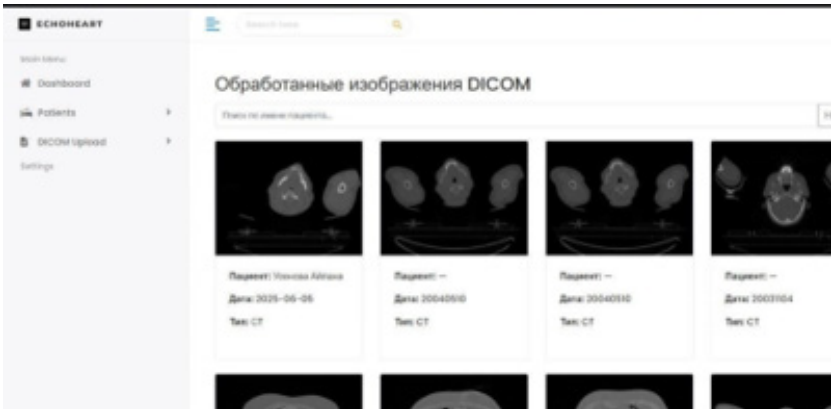


Fig. 4. DICOM format processed images tab.

Since the application handles sensitive medical information, interface security is critical. Authentication and authorization are implemented using Django: access to the doctor panel is restricted by the @login_required decorator, and the user role is checked for each request. Only authorized doctors have access to patient data. Furthermore, all forms are protected with a CSRF token, preventing cross-site request forgery.

Privilege checks are implemented at the interface level: for example, if a doctor is not assigned to a given patient, the controls for their data will be inactive or hidden.

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

A confirmation dialog box is displayed when attempting to delete or replace an image. Maximum file size restrictions and type checking are implemented.

Patient privacy principles are regulated by law: for example, the EU GDPR and the US HIPAA consider medical images to be sensitive data and require explicit patient consent for their processing. The interface includes consent and access logging. Data transfer between the client and server is performed over a secure HTTPS protocol, and stored files are encrypted or accessible only on the server behind a firewall. Furthermore, the interface does not display redundant information: for example, image metadata (passport data) is accessible only to authorized physicians and is hidden from the public. This ensures that security recommendations for medical information systems are implemented: access rights are verified at every step, forms are protected, authentication is implemented, and compliance with personal data protection regulations is ensured.

During the development of a web application for secure storage and preprocessing of medical images, one of the key steps was working with a dataset containing real DICOM files. To demonstrate and test the capabilities of the system being developed, the open source «Cardiac MRI Images Dataset,» published on the Mendeley Data platform, was used (Sharifrazi, 2024).

The dataset is a collection of cardiac images. The DICOM format is an international standard for storing, processing, and transmitting medical images and associated metadata. The images in this dataset cover various phases of the cardiac cycle and contain anonymized patient data, making it suitable for research purposes without violating medical privacy rights. This is especially important in the context of compliance with personal data protection regulations (specifically, GDPR and HIPAA).

The use of an open dataset in this study allowed us to implement and debug the full lifecycle of medical image processing in a web application, from uploading to secure storage and visualization, confirming the viability and practical value of the proposed solution.

Results and discussion.

The testing was conducted in several stages, including a functional check of the system, an analysis of the ethical storage of medical data, and an assessment of the usability of the application for end users.

Functional testing. The correctness of the main functions was verified: loading biomedical images (in DICOM format), saving them, and displaying them via the web interface. Testing was conducted on a set of cardiac images with different characteristics. After loading, all image-related data (DICOM metadata: patient information, examination date, image parameters, etc.) was preserved without corruption. Attention was paid to verifying the integrity of DICOM files during transmission and storage – the application uses the DICOM standard, which guarantees compatibility and information security when exchanging medical images. Image display was tested in various browsers to ensure correct scaling and basic navigation (zooming, scrolling through image series, etc.). To implement DICOM file viewing, specialized Python libraries were used, such as pydicom (for reading/writing DICOM) and Pillow (for displaying images). For example, using

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

the pydicom library, a web application "unpacks" a DICOM file, extracts pixel data and metadata, which made it possible to verify the correct visualization of the medical image and associated information. All functional requirements were met: the system successfully attaches images uploaded by the physician to the corresponding patient records, allows the physician to view them through a browser, perform pre-processing (e.g., adjusting brightness/contrast for improved perception), and save the results (DICOM Standards Committee).

The web application is built on the Django framework, which provides default protection against many common web security threats, including SQL injection, cross-site scripting, and cross-site request forgery. All forms in the application contain CSRF tokens, which were confirmed by tests simulating repeated requests with a fake origin – the framework successfully rejected such attempts. The application's codebase was analyzed for vulnerabilities using a static analyzer; minor warnings were identified and fixed.

The evaluation of the implemented functions was based on five criteria: functional correctness, load resilience, system integration, usability, and reliability. Each function was assessed using a 5-point scale, where a score of 5 indicates full compliance with the requirements and stable operation, while a score of 1 reflects the presence of critical deficiencies. During testing, 50 DICOM files (average size 4.2 MB) were uploaded to the system; the average processing time for brightness and contrast adjustment was 1.1 seconds per image. A summary of the evaluation results is provided in Table 1.

Table 1. Evaluation of Implemented System Functions Based on Key Quality Criteria

Tested Function	Testing Result	Score (1–5)
DICOM file upload	Files uploaded successfully; supports various sizes and modalities	5
Image storage	Images are saved in the database, linked to patients, with no distortion	5
Image viewing via browser	Viewing is available; navigation and zooming function correctly	5
Authentication and authorization	Access granted only to authorized users according to roles	4
Data integrity check (SHA-256)	Checksum is correctly calculated and verified	5
Image preprocessing	Brightness and contrast adjustment works reliably	4
Django security mechanisms (CSRF, XSS, SQLi)	Protection is active; simulated attacks are blocked	5
Cross-browser display	Interface renders correctly in Chrome, Firefox, and Safari	5
Static code analysis	Minor warnings identified and resolved	4
Interface usability	Interface is intuitive and requires no training	5

Overall, testing demonstrated that the developed web application meets the requirements. Functionality is implemented correctly medical images are loaded and displayed without distortion, and preprocessing is performed correctly. The system provides the required level of data protection, preventing unauthorized access and complying with

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

regulatory requirements for medical information security. The application's performance and scalability are sufficient for real-world use: even with increasing data volume or the number of users, response times remain short. The results of usability testing indicate that the interface is sufficiently intuitive, allowing doctors and administrators to operate the system with minimal onboarding. Therefore, based on the chapter's conclusion, it can be concluded that the testing objectives have been achieved—the web application has demonstrated correct operation and is ready for deployment.

Conclusion.

This research involved developing a web application for secure storage, automatic format conversion, and pre-processing of biomedical images (including DICOM format). This project was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant no. AP25794129, «Development of an Algorithm for Filtering and Pre-Processing Biomedical Images for Cardiology Diagnostics»). The application was developed using the Django framework, with the architecture built according to the Model–View–Template pattern. This ensured a clear separation of logic, views, and user interface templates, improving the system's maintainability and scalability. Security and scalability were emphasized during the design process: Django authentication and authorization mechanisms were utilized, along with protection against CSRF attacks and encryption measures for data transfer. The project is designed for future expansion: it includes the ability to scale horizontally (for example, by separating media storage and load balancing), as well as code modularity for easily incorporating new components. The development of the web application was accompanied by testing of key modules—functional and integration tests were conducted, which ensured the correct operation of the system and the reliability of its main components.

The practical significance of the developed web application lies in creating a secure and effective tool for working with medical imaging information. With the growing volume of diagnostic data and the rapid development of telemedicine, solutions that ensure reliable storage and initial processing of images are particularly relevant. The proposed system automates routine processes for medical professionals, including centralized image storage, quick access to examination history, and notifications, thereby improving the efficiency and quality of medical care. Compliance with information security requirements ensures compliance with current legislation and ethical standards for the handling of medical data. Thus, the developed solution represents a relevant and practically relevant tool for digital medicine, contributing to improved diagnostic processes and clinical decision-making.

The project's future development focuses on expanding the system's functionality and scalability. Priority areas include the implementation of a DICOM web viewer, allowing medical images to be viewed directly in a browser; the integration of cloud storage to improve data availability and reliability; the use of machine learning methods for automated image analysis (including segmentation and anomaly detection); the integration of external medical information systems; and the development of analytical modules. These improvements are aimed at further optimizing electronic document management processes and increasing the efficiency of medical personnel.

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

In accordance with the objectives of the funded project, the developed web application will be launched as a functional online system. The release will include public access to the application interface, source code, documentation, and a sample dataset to support reproducibility and further research. This step aligns with the principles of open science and aims to enhance transparency and collaboration in biomedical imaging research.

REFERENCES

Django Software Foundation (2024). Authentication using sessions," Django 5.0 Documentation. [Online]. Available: https://docs.djangoproject.com/en/5.0/topics/auth/.[In Eng.].

DICOM Standards Committee. Digital Imaging and Communications in Medicine (DICOM). [Online]. Available: https://dicom.nema.org/.[In Eng.].

DeMartini J., Sheehan A., Shah N. (2023). Best practices for sharing clinical images: Considerations for patient privacy, informed consent, and transparency. JAMA Dermatol. — Vol. 159. — No. 4. Pp. 341–343. DOI:10.1001/jamadermatol.2023.0154. [In Eng.].

European Society of Cardiology. MRI and Echocardiography Statistics. [Online]. Available: https://www.escardio.org/. [In Eng.].

European Society of Cardiology (2022). Cardiovascular Disease Statistics, ESC Press. [In Eng.].

Eichelberg M., Aden T., Riesmeier J., Dogac A., Laleci G.B. (2004). A survey and analysis of Electronic Healthcare Record standards. Comput. Methods Programs Biomed. — Vol. 73. — No. 3. Pp. 223–239. [In Eng.].

Feigenbaum H., Armstrong W.F., Ryan T. (2018). Feigenbaum's Echocardiography, 8th ed., Lippincott Williams & Wilkins. [In Eng.].

He K., Zhang X., Ren S., Sun J. (2016). Deep Residual Learning for Image Recognition, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). Pp. 770–778. [In Eng.].

HIMSS. (2023). 2023 HIMSS Healthcare Cybersecurity Survey. Healthcare Information and Management Systems Society. [Online]. Available: https://www.himss.org/resources/2023-himss-healthcare-cybersecurity-survey. [In Eng.].

Kremkau F.W. (2019). Diagnostic Ultrasound: Principles and Instruments, 9th ed., Elsevier Health Sciences. [In Eng.].

Kumar K.P., Chinna K.P., Shivaprasad S. (2024). Secure approach to sharing digitized medical data. Biomed. Signal Process. Control. — Vol. 87. — Article 105346. DOI:10.1016/j.bspc.2023.105346. [In Eng.].

Mazurowski M.A., Buda M., Saha A. (2020). Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. Radiology. — Vol. 297. — No. 3. Pp. 663–670. DOI:10.1148/radiol.2020192536. [In Eng.].

Marwick T.H. (2020). Cardiac Imaging: The Requisites, 5th ed., Elsevier. [In Eng.].

Republic of Kazakhstan (2013). Law on Personal Data and Their Protection. No. 94-V. [Online]. Available: https://www.dlapiperdataprotection.com/index.html?t=law&c=KZ. [In Eng.].

Raza M.H., Lawton T., Hunt D., Baxter D. (2024). Retrospective radiology research: Do we need informed consent? J. Bioeth. Inq. — Vol. 21. — No. 2. DOI:10.1007/s11673-024-10368-6. [In Eng.].

Sharifrazi D. (2024). CAD Cardiac MRI Dataset. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/danialsharifrazi/cad-cardiac-mri-dataset[In Eng.].

U.S. Department of Health & Human Services, Health Insurance Portability and Accountability Act (HIPAA). [Online]. Available: https://www.hhs.gov/hipaa/.[In Eng.].

World Health Organization (2021). Ethical considerations for health data privacy, confidentiality and security. WHO, Geneva. [Online]. Available: https://www.who.int/publications/i/item/9789240030839. [In Eng.].

World Health Organization (2023). Cardiovascular diseases (CVDs). Fact Sheets, WHO. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [In Eng.].

Zipes D.P., Libby P., Bonow R.O., Mann D.L. (2021). Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine, 11th ed., Elsevier. [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.2. Number 26 (2026). Pp. 219–235
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.014
УДК 004.8:004.9

APPROACHES TO BUILDING A HYBRID INTELLIGENT DECISION SUPPORT SYSTEM BASED ON FUZZY LOGIC, XAI, AND LARGE LANGUAGE MODELS

L. Kurmangaziyeva[1], Sh. Yelezhanova[1], S. Iskakova[2], Sh. Kodanova[2]*, H.İ. Bülbül[3]
[1]Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[2]Safi Utebayev Atyrau Oil and Gas University;
[3]Gazi University, Ankara, Turkiye.
E-mail: kodanova_s@mail.ru

Lyailya Kurmangaziyeva — Candidate of Technical Sciences, Professor of the Department of "Software Engineering" Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000000306407306;
Shynar Yelezhanova — Candidate of Physico-Mathematical Sciences, Professor of the Department of "Software Engineering", Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0001-9815-9594;
Shynar Kodanova — Candidate of Technical Sciences, Associate Professor, Faculty 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;
Sandugash Iskakova — Candidate of Technical Sciences, Associate Professor, Faculty of Information Technology, Safi Utebayev Atyrau Oil and Gas University, Atyrau, Kazakhstan
https://orcid.org/0000-0002-6589-854X;
Halil İbrahim Bülbül — PhD, Professor, Department of Computer and Instructional Technologies, Gazi Eğitim Fakültesi, Gazi University, Ankara
https://orcid.org/0000-0002-6525-7232.

© Kurmangaziyeva, Sh. Yelezhanova, S.Iskakova, Sh.Kodanova H.İ. Bülbül

Abstract. The paper presents the software architecture and practical implementation of an intelligent web-based decision support system that integrates fuzzy logic, explainable artificial intelligence (XAI), and a large language model (LLM) within a unified hybrid framework. A key architectural principle is the separation of the computational decision-making layer from the generative layer. The fuzzy controller processes the input indicators, evaluates them using predefined membership functions and expert rules, and produces a

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

deterministic recommendation. The XAI module then improves transparency by displaying the activated fuzzy rules, membership degrees, and the factors that contributed to the final decision. In contrast, the LLM does not determine the decision itself; instead, it provides context-aware natural-language interpretation, explanatory feedback, and adaptive content based on the computed result. The server side of the system is implemented in Python using the Flask framework, while persistent user and decision data are stored in an SQLite database. Communication with the generative component is organized through an application programming interface. Pilot evaluation demonstrated a high level of agreement with expert judgments, stable system operation, and an end-to-end response cycle suitable for interactive decision support and adaptive web-based applications.

Keywords: intelligent web system, decision support, fuzzy logic, explainable AI, large language models, Flask, SQLite, API, adaptive system

For citation: Kurmangaziyeva L., Yelezhanova Sh., Iskakova S., Kodanova Sh., Bülbül H.İ. (2026). Approaches to building a hybrid intelligent decision support system based on fuzzy logic, XAI, and large language models // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 219–235. https://doi. org/10.54309/IJICT.2026.26.2.014. (In Eng.).

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

АЙҚЫН ЕМЕС ЛОГИКА, ХАІ ЖӘНЕ ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР НЕГІЗІНДЕ ГИБРИДТІ ЗИЯТКЕРЛІК ШЕШІМ ҚАБЫЛДАУДЫ ҚОЛДАУ ЖҮЙЕСІН ҚҰРУ ТӘСІЛДЕРІ

Л. Курмангазиева[1], Ш. Ележанова[1], С. Искакова[2], Ш. Коданова[2]*, Х.И. Бұлбұл[3]

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

[2]Сафи Өтебаев атындағы Атырау мұнай және газ университеті, Атырау, Қазақстан;

[3]Гази университеті, Анкара, Түркия.

E-mail: kodanova_s@mail.ru

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

https://orcid.org/0000000306407306;

Ележанова Шынар — физика-математика ғылымдарының кандидаты, «Бағдарламалық инженерия» кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан

https://orcid.org/0000-0001-9815-9594;

Коданова Шынар — техника ғылымдарының кандидаты, Сафи Өтебаев атындағы Атырау мұнай және газ университетінің Ақпараттық технологиялар факультетінің қауымдастырылған профессоры, Атырау, Қазақстан

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

Искакова Сандуғаш — техника ғылымдарының кандидаты, Сафи Өтебаев атын-

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

дағы Атырау мұнай және газ университетінің Ақпараттық технологиялар факультетінің қауымдастырылған профессоры, Атырау, Қазақстан
https://orcid.org/0000-0002-6589-854X;
Бұлбұл Халил Ибрагим — PhD, Гази білім беру факультеті, Компьютерлік және оқыту технологиялары кафедрасының профессоры, Гази университеті, Анкара
https://orcid.org/0000-0002-6525-7232.

© Л. Курмангазиева, Ш. Ележанова, С. Искакова, Ш. Коданова, Х.И. Бұлбұл

Аннотация. Мақалада айқын емес логиканы, түсіндірілетін жасанды интеллектті (XAI) және үлкен тілдік модельді (LLM) бірыңғай гибридті архитектура аясында біріктіретін интеллектуалды веб-негізделген шешім қабылдауды қолдау жүйесінің бағдарламалық архитектурасы мен практикалық іске асырылуы ұсынылған. Жүйені құрудың негізгі қағидасы — шешім қабылдаудың есептеу деңгейін генеративті деңгейден бөлу. Айқын емес контроллер алдын ала анықталған тиесілілік функциялары мен сарапшылық ережелер негізінде кіріс көрсеткіштерін өңдеп, детерминирленген ұсыныс қалыптастырады. XAI модулі белсендірілген айқын емес ережелерді, тиесілілік дәрежелерін және қорытынды шешімге әсер еткен факторларды көрсету арқылы жүйенің ашықтығын арттырады. Бұл ретте LLM шешім қабылдау процесіне тікелей қатыспайды, керісінше есептелген нәтиже негізінде контекстке тәуелді мәтіндік түсіндірмелерді, дербестендірілген кері байланысты және бейімделген мазмұнды қалыптастыру үшін пайдаланылады. Жүйенің серверлік бөлігі Python бағдарламалау тілінде Flask фреймворкі арқылы іске асырылған, ал пайдаланушы деректері мен қабылданған шешімдердің нәтижелері SQLite дерекқорында сақталады. Генеративті компонентпен өзара әрекеттесу бағдарламалық интерфейс (API) арқылы ұйымдастырылған. Пилоттық тестілеу жүйе нәтижелерінің сарапшылық бағалаулармен жоғары сәйкестігін, жүйенің тұрақты жұмысын және интерактивті шешім қабылдауды қолдау жүйелері мен бейімделген веб-қосымшалар үшін қолайлы толық өңдеу циклінің уақытын көрсетті.

Түйінді сөздер: зияткерлік веб-жүйе, шешім қабылдауды қолдау, айқын емес логика, XAI, LLM, Flask, SQLite, API, бейімделген жүйе

Дәйексөздер үшін: Курмангазиева Л., Ележанова Ш., Искакова С., Коданова Ш., Бұлбұл Х.И. (2026). Айқын емес логика, XAI және үлкен тілдік модельдер негізінде гибридті зияткерлік шешім қабылдауды қолдау жүйесін құру тәсілдері // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 219-235. https://doi.org/10.54309/IJICT.2026.26.2.014. (Ағыл. тіл.).

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

ПОДХОДЫ К ПОСТРОЕНИЮ ГИБРИДНОЙ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ, XAI И БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

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

Л. Курмангазиева[1], Ш. Ележанова[1], С. Искакова[2], Ш. Коданова[2]*, Х.И. Бюльбюль[3]
[1]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан;
[2]Атырауский университет нефти и газа имени Сафи Утебаева, Атырау, Казахстан;
[3]Университет Гази, Анкара, Турция.
E-mail: kodanova_s@mail.ru

Курмангазиева Ляйля — кандидат технических наук, профессор кафедры «Программная инженерия» Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000000306407306;
Ележанова Шынар — кандидат физико-математических наук, профессор кафедры «Программная инженерия» Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0001-9815-9594;
Коданова Шынар — кандидат технических наук, ассоциированный профессор факультета информационных технологий Атырауского университета нефти и газа имени Сафи Утебаева, Атырау, Казахстан
E-mail: kodanova_s@mail.ru, https://orcid.org/0000-0002-1589-4268;
Искакова Сандугаш — кандидат технических наук, ассоциированный профессор факультета информационных технологий Атырауского университета нефти и газа имени Сафи Утебаева, Атырау, Казахстан
https://orcid.org/0000-0002-6589-854X;
Бюльбюль Халил Ибрагим — PhD, профессор кафедры компьютерных и образовательных технологий педагогического факультета Гази, Университет Гази, Анкара
https://orcid.org/0000-0002-6525-7232.

© Л. Курмангазиева, Ш. Ележанова, С. Искакова, Ш. Коданова, Х.И. Бюльбюль

Аннотация. В статье представлена программная архитектура и практическая реализация интеллектуальной веб-системы поддержки принятия решений, объединяющей нечеткую логику, объяснимый искусственный интеллект (XAI) и большую языковую модель (LLM) в рамках единой гибридной архитектуры. Ключевым принципом построения системы является разделение вычислительного уровня принятия решений и генеративного уровня. Нечеткий контроллер обрабатывает входные показатели на основе заданных функций принадлежности и экспертных правил и формирует детерминированную рекомендацию. Модуль XAI повышает прозрачность системы за счет отображения активированных нечетких правил, степеней принадлежности и факторов, повлиявших на итоговое решение. При этом LLM не участвует непосредственно в принятии решения, а используется для формирования контекстно-зависимых текстовых объяснений, персонализированной обратной связи и адаптивного контента на основе рассчитанного результата. Серверная часть системы реализована на языке Python с использованием фреймворка Flask, а постоянное хранение пользовательских данных

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

и результатов принятия решений осуществляется в базе данных SQLite. Взаимодействие с генеративным компонентом организовано посредством программного интерфейса API. Пилотное тестирование продемонстрировало высокий уровень соответствия результатов системы экспертным оценкам, стабильность функционирования и приемлемую продолжительность полного цикла обработки запроса для интерактивных систем поддержки принятия решений и адаптивных веб-приложений.

Ключевые слова: интеллектуальная веб-система, информационная система, поддержка принятия решений, нечеткая логика, XAI, LLM, Flask, SQLite, REST API, адаптивная система

Для цитирования: Курмангазиева Л., Ележанова Ш., Искакова С., Коданова Ш., Бюльбюль Х.И. (2026). Подходы к построению гибридной интеллектуальной системы поддержки принятия решений на основе нечеткой логики, XAI и больших языковых моделей // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 219-235. https://doi. org/10.54309/IJICT.2026.2.014. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The development of intelligent information systems is accompanied by a shift from hard-coded scenarios to adaptive software environments capable of accounting for input data uncertainty, explaining decisions, and interacting with the user in natural language. This trend is particularly noticeable in educational systems, HR analytics, and digital decision support services, where initial indicators rarely have clear boundaries, and the user needs to understand not only the outcome but also the reasons for its achievement (Chrysafiadi et al., 2023; Susnjak, 2024).

Classical intelligent tutoring systems provide personalization through a user model, adaptation rules, and accumulated results, but their development requires a combination of several technological approaches. Fuzzy systems are convenient for representing expert rules in a form close to natural language and work well with blurred boundaries between levels (Zadeh, 1965; Mamdani et al., 1975). Modern XAI approaches complement the computational model with interpretation mechanisms, which increases trust in algorithmic decisions and allows for the identification of contradictions between rules, input data, and the result (Arrieta et al., 2020; Schwalbe et al., 2024).

The emergence of large language models has significantly expanded the capabilities of the user interface of intelligent systems. LLMs are capable of creating customized explanations, generating new tasks, and adapting text to the user's context. However, transferring the decision-making function itself to a generative model creates risks of non-determinism and insufficient traceability. Therefore, for an applied information system, it is advisable to separate the computational loop from the generative loop: the decision should be generated by a supervised algorithm, and the LLM should be used for the semantic representation and adaptation of the result (Vaswani et al., 2017; Susnjak, 2024).

The aim of the study is to develop and experimentally evaluate an intelligent

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

web-based decision support system in which fuzzy inference, XAI visualization, user data storage, and a generative LLM module are combined in a single modular software architecture.

To achieve this goal, the following tasks were completed: a component-based web application architecture was developed; a server layer was implemented using Python / Flask ; a model for storing user sessions and results in SQLite was designed; a fuzzy inference module was integrated; an XAI interface was developed ; programmatic exchange with LLM via API was organized ; the information flow from the user request to the delivery of the adapted result was defined; and a pilot test of the performance and consistency of the solutions was conducted.

Materials and methods.

The system uses two basic metrics as input: user performance and task completion time. These parameters were chosen as a minimum set to evaluate not only the correctness of the answer but also the nature of the task completion. In practical scenarios, the set of features can be expanded with additional parameters: number of attempts, error type, sequence of actions, history of previous sessions, or expert characteristics.

Table 1. Functional requirements for an intelligent web system

Component	Main function	Result
Web interface	Entering data and displaying the result	Custom request / response
Fuzzy module	Processing undefined input indicators	Numerical index and level
XAI module	Explanation result	Rules, degrees accessories, explanation
LLM module	Semantic adaptation and generation	Personalized text / task
Base data	Storage profiles and sessions	History, results, parameters
API layer	Connection of internal and external services	Structured exchange data

Architecture intellectual web systems.

The software architecture is designed around the principle of separation of concerns. The user interface is responsible for receiving and visually presenting data; the Flask controller organizes request routing and processing; the fuzzy logic module performs computational decision making; the XAI module generates explainable decision features; the storage layer ensures session persistence; and the external LLM service is used only after a controlled result is obtained. This approach allows for independent modification of individual components and reduces code coupling.

Flask was chosen as the server platform. Its routing and minimal core coupling allow for separate services for the compute module, database, and external API. The official Flask documentation recommends an application factory and modular component registration as a means of scalable project organization (Pallets). Projects (2026). In the system under study, this idea is used to logically separate the view, controller, and data model.

Unlike fully generative systems, the LLM does not have the power to independently determine the user category. The input to the generative module is the already calculated

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

level, error information, and acceptable content constraints. Thus, the architecture implements the principle of deterministic decision – generative interpretation.

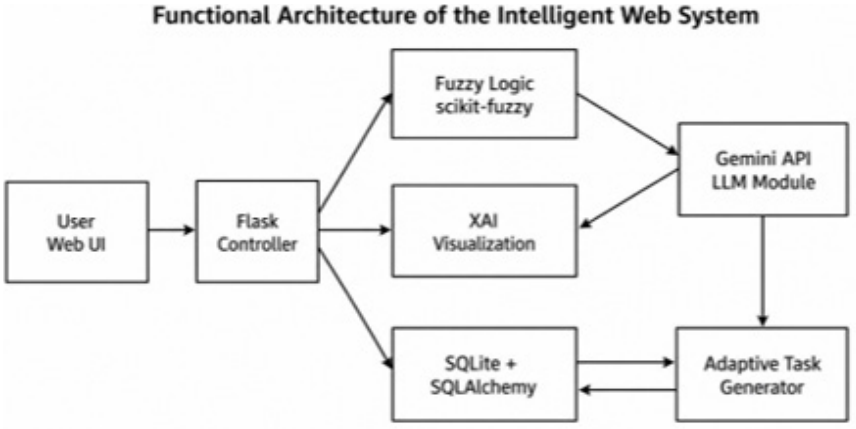


Fig. 1. Functional architecture of an intelligent web system

Table 2. Software technologies and their role

Technology	Purpose
Python	Server logic, computing modules, library integration
Flask	HTTP Routing and Backend Web Applications
scikit-fuzzy	Mamdani's membership and inference functions
SQLite	Local transactional storage data
SQLAlchemy ORM	Table Access Abstraction and Object Data Model
JavaScript	Dynamic XAI visualization in the client area
Gemini API	Generating explanations and adaptive content
Ngrok	External HTTPS access at the testing stage

The server side is implemented as a set of routes corresponding to user scenarios. A GET request returns the interface and the initial session state; a POST request receives user data; a service handler performs validation and type conversion; then the data is passed to the fuzzy inference module. After the calculation, the result is saved, an XAI structure is generated, and only then is the request initiated to the LLM. This sequence eliminates the possibility that an error in an external generative service could affect the underlying solution.

$$\text{Request} \rightarrow \text{Validate} \rightarrow \text{Fuzzy}(X) \rightarrow \text{Explain}(y^*) \rightarrow \text{Persist} \rightarrow \text{Generate}(\text{Context}) \quad (1)$$

At the programmatic level, exchange between components is built through structured objects. Dictionaries/ JSON representations are used for internal interaction, including the session identifier, input parameters, calculated level, activated rules, text interpretation, and generation metadata. When integrating an external model, an API call with explicit instructions and context is used. Modern Gemini The API supports the transmission of structured content and system instructions, which makes it possible to limit the role of the generative module to the scope of the application task (Google AI for

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

Developers (2026).

Table 3. Logical endpoints of the web application

Route	Method	Function
/	GET	Loading interface
/ evaluate	POST	Receiving indicators and running fuzzy estimation
/ explain	POST	Formation of XAI explanation
/ generate	POST	Adaptive Content Generation through LLM
/ history	GET	Receipt stories custom sessions
/ health	GET	Control availability services

Exception handling is an important requirement. If the external API is unavailable, the user must receive the calculated level and XAI explanation without generative augmentation. If a validation error occurs, the request must not be passed to the computation module. If a database failure occurs, the operation must be completed by rolling back the transaction. This ensures that the failure of a single component does not lead to an incorrect change in the decision-making result.

For testing purposes, a local Flask server was published via a secure HTTPS tunnel. This solution is convenient for exploratory testing of the interface on various devices, but is not intended for production deployment. For production use, the architecture suggests migrating the application to a permanent server with a WSGI / ASGI gateway, centralized logging, and secrets management.

A relational model is used to store user profiles, session parameters, and results. SQLite was chosen due to its ease of deployment and suitability for a pilot test. The ORM layer separates application logic from specific SQL syntax and allows system entities to be represented as program classes. As the load increases, the schema can be migrated to a server DBMS without any additional work. changes logic computing modules.

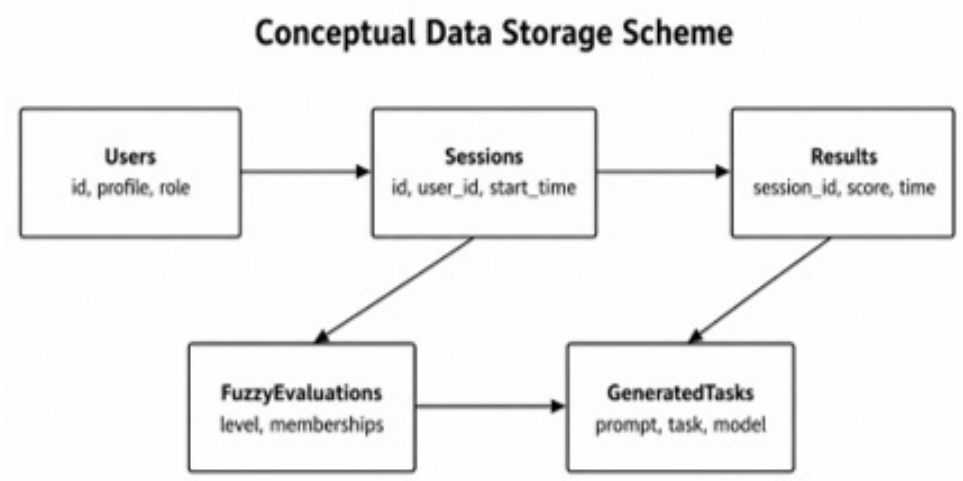


Fig. 2. Conceptual diagram of user and analytical data storage

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

Table 4. Main entities of the information model

Essence	Key fields	Purpose
Users	ID, profile, role	User identification and profile settings
Sessions	id, user_id, start_time	Connection of actions within one attempt
Results	session_id, score, duration	Factual indicators execution
FuzzyEvaluations	session_id, level, memberships	Fuzzy inference result and XAI data
GeneratedTasks	session_id, prompt, response, model	Story generative interactions

The user-session relationship is implemented as a one-to-many relationship, as a single user can complete multiple attempts. The result and fuzzy evaluation are linked to a specific session. The history of generative tasks is also stored by session ID, ensuring reproducibility of the user scenario and allowing analysis of the data transferred to LLM.

Separating the original results, the calculated estimate, and the generative response is crucial for auditing. If necessary, the solution can be recalculated based on the original values independently of the text generated by the external model. This reduces the risk of logical confusion between actual data and probabilistically generated content.

The fuzzy module functions as the computational core. Unlike the first mathematically oriented formulation, in this paper it is considered a software service that accepts normalized input parameters and returns a structured result. The minimal configuration uses two variables: performance x_1 and time x_2. Each variable is described by three linguistic terms, and the rule base contains nine combinations.

$$X = (x_1, x_2),\ x_1 \in [0,100],\ x_2 \in [0, T\max] \tag{2}$$

$$y^* = D\left(Agg\left(\{ Rk\left(\Phi\left(X\right)\right)\}\right)\right),\ k = 1,\ldots,9 \tag{3}$$

Membership functions, rules, and defuzzification are implemented using scikit - fuzzy. Programmatically, the module has an interface of the form evaluate (score, time) and returns an object containing the numerical value y *, the text level, the membership degrees of the inputs, the activated rules, and their trigger strengths. This contract allows the same computational service to be used in both the educational and HR scenarios, varying the configuration of the membership functions and the interpretation dictionary.

Table 5. Example of Fuzzy service result interface

Field	Type	Content
crisp_value	float	Defuzzified index
level	string	Category: Low/Medium/High or Easy/Medium/Hard
memberships	object	Degrees accessories input parameters
rules	array	List activated production rules
rule_strength	array	Forces triggering rules
timestamp	datetime	Time calculation

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

The computational service contains no interface code and does not depend on an external LLM. This simplifies unit testing: for a given set of inputs, the expected value y, the category, and the set of active rules can be tested. The determinism of such a core is an important property of a decision support system and distinguishes the proposed architecture from solutions where the final recommendation is generated solely by a language model.

Fuzzy rules are also convenient for system maintenance: an expert can change thresholds and rules without retraining the neural network. This also allows for storing configuration separately from the application code and, in the future, implementing an administrative interface for managing the knowledge base.

Explainability in the developed system is implemented at two levels. The first level is mathematically straightforward: the user or expert is shown the membership degrees of the input values and the production rules that influenced the outcome. The second level is semantic: a short text explanation is generated based on the same structured data. This division corresponds to the modern XAI approach, which emphasizes not only the formal attributes of interpretability, but also the suitability of the explanation for a specific user (Barredo Arrieta et al., 2020; Afzaal et al., 2024).

The client-side XAI component receives data from the backend in JSON format. For each input variable, the position of the actual value relative to the membership functions is displayed, followed by a list of activated rules and their contribution. For example, the system might show that the result has high membership in the term «High,» time has partial membership in the terms «Normal» and «Slow,» and the final level is formed by a combination of the two most active rules.

Table 6. Structure of XAI explanation

Element	What is displayed	Purpose
Entrance signs	Points, time, normalized values	Examination original data
Accessories	μLow, μMedium, μHigh, etc.	Understanding fuzzy classifications
Active rules	IF–THEN rules	Tracking logic conclusion
Strength rules	Numerical degree activation	Grade influences
Result	y* and text category	Fixation solutions
Text explanation	Brief interpretation	Availability result to the user

A key principle is that the textual explanation should not replace the underlying mathematical tracing. Even if LLM is used to reformulate the explanation in natural language, the interface preserves the numerical membership degrees and the original rules. This ensures verifiability and allows an expert to detect instances where the generative formulation does not fully match the calculation.

It's recommended to use compact diagrams and an expandable «Why did the system make this decision?» block in the application interface. This design reduces information overload: the average user sees a clear explanation, while the expert receives a detailed traceback if needed.

The generative module is called after fuzzy inference is complete. Its task is not to select a level, but to generate content corresponding to the user's computed state. In the

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

educational scenario, the LLM generates a new task or explanation; in the HR scenario, it generates a structured interpretation of the result and recommendations for the next evaluation stage. This approach reduces the risk that free generation will change the underlying decision.

The request context is generated automatically and includes: the calculated level; a brief XAI explanation; the type of error detected; domain constraints; and response format requirements. The request specifically specifies that the model should not change the calculated level and should operate only within the provided context. This complies with the prompt driven principle engineering, when a generative model is used as a specialized component of an information system rather than as an autonomous source of truth.

$$Prompt = f(level, error\_type, XAI, domain\_constraints, output\_schema) \quad (4)$$

The API layer handles network timeouts, service unavailability, and incorrect response formats. If a generation error occurs, the system returns the basic deterministic result to the user and prompts them to repeat the generative part separately. This functional degradation increases the reliability of the web application.

Table 7. Control of the generative module

Mechanism	Implementation	Effect
Fixation solutions	The level is passed as an immutable context	LLM does not define a base category
Limitation subject areas	System instructions and prompt template	Decrease irrelevant answers
Structure answer	JSON requirement / given sections	Simplification processing on the backend
Timeout/fallback	Interception network exceptions	System works without LLM
Logging	Saving prompt and response	Audit and reproducibility

Gemini documentation The API provides for the generation of content with system instructions and structured query parameters (Google AI for Developers, 2026). In the research prototype, this is used to convey level context and errors. In production implementation, the external service could be replaced with a local LLM while maintaining the same internal generative service interface.

The system operates as a sequential pipeline. The user enters or generates initial indicators; the backend checks their ranges; a session is created or updated; the fuzzy module performs the evaluation; the XAI module generates interpretable features; the results are stored in the database; then the context of the generative query is built and the LLM is called. The final answer unites deterministic assessment, explanation and adaptive content.

The algorithm can be represented as the following sequence:

1) get score and duration;
2) perform range checking;
3) save the raw data with the session ID;

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

4) call the Fuzzy service;
5) save y *, memberships and rules;
6) generate an XAI explanation;
7) prepare a prompt;
8) call LLM;
9) check the response;
10) save the generative result and return the aggregated object to the client.

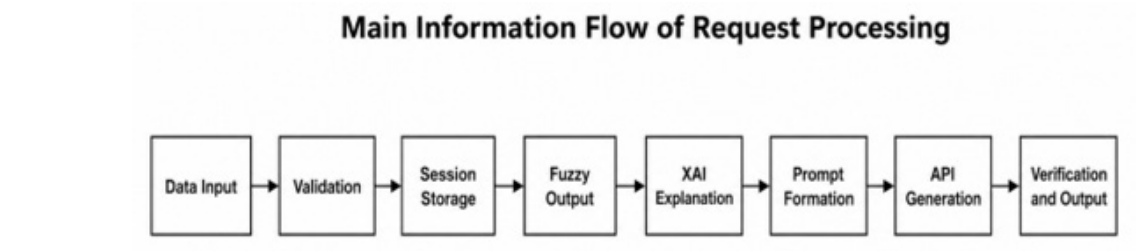


Fig. 3. Information flow for processing a user request

From a software engineering perspective, such a pipeline allows for the separation of synchronous critical operations from potentially slower external generation. When scaling, the LLM call can be moved to a task queue, and the interface can receive the result asynchronously. For the pilot version, the entire cycle was executed sequentially, as the measured latency remained acceptable for interactive mode.

The information flow also supports auditing: each stage leaves data that allows one to reconstruct the decision-making process. This is especially important for systems in which the results can be analyzed by a teacher, HR specialist, or administrator. The presence of initial parameters, rules, and a saved prompt makes it possible to compare versions of the algorithm and the external model.

The pilot testing was carried out as an interactive pilot test of the prototype. The main criteria were the correctness of the entire software pipeline, the consistency of the final recommendations with expert assessments, the suitability of the XAI representation, and the delay from sending input data to receiving a complete response. The system was tested in an educational scenario and using an HR assessment as an example, which allowed us to assess the architecture's transferability across subject areas.

Table 8. Summary indicators pilot testing

Indicator	Result	Interpretation
General agreement with expert assessments	95.4 %	High proximity of the hybrid solution to the expert one
HR scenario	94.0 %	Stable correspondence expert assessment
Educational scenario	91.3 %	Acceptable quality adaptive assessments
Full cycle answer	1.5–2.0 s	Suitable for interactive web work

The obtained values are considered the results of a pilot, not an industrial, verification.

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

They confirm the functionality of the selected software architecture, but do not replace testing on large samples or load testing. Most importantly for this article, the addition of XAI and the generative layer did not compromise the manageability of the underlying computing solution.

During testing, the web interface was used from various client devices. HTTPS tunnel access allowed for testing display and interaction on smartphones, tablets, and personal computers. This mode is particularly useful during the prototyping stage, as it allows for quick identification of responsive layout issues, network latency, and browser incompatibility.

The functional separation of the modules was also experimentally confirmed. When the generative API was disabled, the ability to calculate the level and display the XAI explanation remained. Therefore, the LLM is an extension component, but not critical to the correctness of the core. This property improves the system's robustness and facilitates future replacement of the external model.

Table 9. Comparison options architecture

Criterion	Fuzzy only	LLM only	Fuzzy + XAI + LLM
Determinism solutions	High	Low / depends from generation	High
Explainability	Average	Unstable	High
Adaptive text	Limited	High	High
Audit solutions	High	Limited	High
Addiction from external API	No	High	Partial

Results.

Table 10. Experimental performance indicators of the hybrid system

Indicator	The resulting value
Consistency of the hybrid model with expert assessment	95.4%
System consistency with HR experts in a test scenario	94%
Matching adaptive tasks to the learner's profile	91.3%
Average time for a full request processing cycle	1.5–2.0 s
Reducing the volume of manual feedback	78%

To test the performance of the developed hybrid intelligent system, a series of computational and functional experiments were conducted. The main goals of the testing were to evaluate the correct operation of the fuzzy block, the consistency of the hybrid model's results with expert decisions, the effectiveness of adaptive task generation, and the processing time of user requests.

The experimental implementation of the system was implemented as a web application. Flask served as the server side, *the scikit-fuzzy* library was used for fuzzy inference, SQLite and SQLAlchemy were used for data storage, and the generative component was implemented via calls to the Gemini API. An XAI module was created

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

as a separate functional block, designed to visually and meaningfully explain the system's generated solutions.

The developed system ensures sequential processing of information according to the following scheme: input of initial data → validation → saving session parameters → fuzzy inference → XAI interpretation → formation of a prompt → generation of a response via API → verification and issuance of the result.

During the computational experiment, the agreement between the hybrid model's results and expert assessments was assessed. The comparison revealed a 95.4 % agreement rate, demonstrating a high degree of consistency between the system's decisions and expert assessments.

Table 11. Results of the system accuracy and consistency assessment

Criterion assessments	Result
Consistency of the hybrid model with expert assessment	95.4 %
Consistency with experts in the test scenario	94.0 %
Correspondence of generated tasks to the user profile	91.3 %

Particular attention was paid to testing the adaptive mechanism. After determining the user's current state using fuzzy logic, the system generated a set of parameters, which were used to create a structured prompt for the LLM. As a result, the generated tasks varied depending on the user's level, the results obtained, and their previous activity.

In pilot testing, the proportion of tasks matching the user profile defined by the system was 91.3 %. This result demonstrates that the use of intermediate fuzzy evaluation allows for fairly precise control of the content and difficulty level of automatically generated tasks.

Table 12. Evaluation of the adaptive mechanism

Parameter	Meaning
Matching the task to the current user profile	91.3 %
Correctness choice level system	94 %
Decrease necessity manual adjustments	78 %

Using the generative module reduced the amount of manual feedback generation and editing. Experimental use showed a reduction in labor costs for preparing personalized feedback of approximately 78% compared to a completely manual approach.

The execution time of the full request processing cycle was also analyzed. This cycle included input data validation, fuzzy inference, explanation generation, prompt preparation, API access, result retrieval, and subsequent processing.

The average time of a full cycle was 1.5–2.0 s, which allows the developed system to be used in interactive mode.

Additionally, the system's resilience to external API errors was tested. If there is no response or an exception occurs, the request does not terminate the user session. The system returns a controlled *fallback* response, and the current session parameters and intermediate results are saved in the database.

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

Table 13. Request processing time characteristics

Stage	Characteristic
Validation input data	is being carried out locally
Fuzzy conclusion	is being carried out locally
Formation of XAI interpretation	is being carried out locally
Preparation prompt	is being carried out automatically
Accessing the Gemini API	basic external temporal factor
Full cycle processing	1.5–2.0 s

The data storage structure includes the Users, Sessions, Results, FuzzyEvaluations and GeneratedTasks tables, which ensures the storage of the history of user interaction with the system, the results of fuzzy evaluation, generated tasks and the parameters of individual sessions.

Thus, the experimental results confirm the viability of the proposed architecture. The system ensures a high level of consistency with expert decisions, adaptation of generated content to the user profile, acceptable response times, and a reduction in manual effort.

Discussion.

The obtained results demonstrate that combining fuzzy logic, explainable artificial intelligence, and a large language model enables the creation of a flexible, adaptive system capable of taking into account individual user characteristics and operating under uncertainty. Unlike rigid, deterministic rules, fuzzy inference allows for a more accurate interpretation of intermediate states and the formation of final assessments without sharply dividing users into fixed categories.

The high consistency of the system's results with expert assessments confirms the validity of the chosen approach. A 95.4 % score indicates that the hybrid model's decisions in most test scenarios are consistent with expert opinion. However, the system does not completely replace human experts, but rather serves as a decision support tool, particularly in situations where the initial data is ambiguous.

The resulting 91.3 % compliance rate between adaptive tasks and the user profile demonstrates that fuzzy assessment results can be effectively used to guide the generation of educational content. This is a significant advantage over traditional systems, which select tasks only from a pre-defined pool. In the developed system, task content is generated dynamically based on the results of the current session.

A separate advantage is the use of the XAI component. It allows for explanation of which parameters and membership function values influenced the system's final solution. This increases the transparency of the intelligent module's operation and makes the results more understandable for both the user and the instructor.

The practical effectiveness of the system is confirmed by a reduction in the volume of manual feedback by approximately 78 %. Automated generation of assignments and recommendations reduces the instructor's workload and allows them to focus on complex or non-standard cases. However, completely eliminating expert review is impractical, as

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

the quality of the generated LLM content may depend on the context and structure of the prompt.

The average request processing time of 1.5–2.0 seconds allows for interactive use of the developed system. Latency is primarily driven by calls to the external API, while validation, fuzzy inference, and XAI interpretation generation are performed locally.

One of the limitations of the proposed architecture is its dependence on an external language model and API availability. To improve resilience, a fallback response mechanism is provided, ensuring that an external service failure does not lead to a complete system shutdown. Furthermore, the results of intermediate calculations and session parameters are stored in the database.

In the future, it would be advisable to expand the experimental sample, conduct testing with a larger number of users, and compare the results of different language models. Also promising is improving the membership functions, the fuzzy rule base, and XAI interpretation methods.

Thus, the proposed approach demonstrates the practical feasibility of combining fuzzy logic, XAI, and generative models to build adaptive intelligent web systems. The results confirm the architecture's potential in terms of accuracy, interpretability, performance, and the ability to generate personalized content.

Conclusion.

This paper develops an intelligent web-based decision support system that integrates fuzzy logic, XAI, and a large language model into a single software architecture. Unlike purely generative solutions, the proposed approach separates computation and generation: the base solution is generated by a deterministic Fuzzy module, XAI provides mathematical and visual traceability, and LLM performs semantic adaptation and generates subsequent content.

The software implementation is built in Python and Flask, and the storage layer is built on SQLite with object-relational mapping. Component separation allows for independent maintenance of the user interface, computational core, XAI visualization, database, and external generative service. The information flow is organized so that LLM failure does not affect the correctness of the main assessment: the user retains access to the calculated level and its explanation.

The pilot validation confirmed the system's functionality in two subject-matter scenarios. Overall agreement between the hybrid system and expert assessments was 95.4 %; in the HR scenario, it was 94 %; and in the education scenario, it was 91.3 %. The full response cycle was in the range of 1.5–2.0 seconds, allowing the prototype to be used in interactive web mode. These results should be considered a validation of the architecture, not a definitive assessment of the industrial solution's performance.

The scientific and practical significance of this work lies in the proposed framework for integrating symbolic and generative artificial intelligence at the information system level. Fuzzy logic ensures controlled decision making under uncertain inputs; XAI makes the decision mechanism auditable; and LLM extends the interface through natural language generation. Thus, the generative model is used as an interpretation and personalization

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

service, while maintaining the mathematical traceability of the underlying solution.

Future research opportunities include load testing, automatic calibration of membership functions, rule base versioning, migration to a server-side DBMS, containerization, asynchronous LLM invocation, the use of local language models, and advanced quality assurance of generated content. A separate area of focus is the development of an administrative interface for rule management and decision explainability monitoring.

REFERENCES

Arrieta B. 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. — Vol. 58. — Pp. 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. [in Eng.].

Afzaal M., Zia A., Nouri J. & Fors U. (2024). Informative feedback and explainable AI-based recommendations to support students' self-regulation // Technology, Knowledge and Learning. — Vol. 29. — Pp. 331–354. DOI: https://doi.org/10.1007/s10758-023-09650-0. [in Eng.].

Chrysafiadi, K., Virvou, M., Tsihrintzis, G.A. & Alexandris, G. (2023). Evaluating the user's experience, adaptivity and learning outcomes of a fuzzy-based intelligent tutoring system for computer programming for academic students in Greece. Education and Information Technologies. — Vol. 28. — Pp. 6453–6483. DOI: https://doi.org/10.1007/s10639-022-11444-3. [in Eng.].

Google AI for Developers. (2026). Gemini API: Generating content and API documentation. https://ai.google.dev/api/generate-content. [in Eng.]

Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python. 2nd ed. O'Reilly Media. [in Eng.].

Hobert S. & Berens F. (2024). Developing a digital tutor as an intermediary between students, teaching assistants, and lecturers // Educational Technology Research and Development. — Vol. 72. — Pp. 797–818. DOI: https://doi.org/10.1007/s11423-023-10293-2. [in Eng.].

Kouam F., A.W. (2024). The effectiveness of intelligent tutoring systems in supporting students with varying levels of programming experience. Discover Education. — Vol. 3. — Pp. 278. DOI: https://doi.org/10.1007/s44217-024-00385-3. [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.].

Pallets Projects. (2026). Flask Documentation, version 3.1.x. https://flask.palletsprojects.com/. [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.].

Susnjak T. (2024). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT // International Journal of Artificial Intelligence in Education. — Vol. 34. — Pp. 452–482. DOI: https://doi.org/10.1007/s40593-023-00336-3. [in Eng.].

Schwalbe G. & Finzel B. (2024). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery. — Vol. 38. — Pp. 3043–3101. DOI: https://doi.org/10.1007/s10618-022-00867-8. [in Eng.].

SQLite Consortium. (2026). SQLite Documentation. https://www.sqlite.org/docs.html. [in Eng.].

SciPy Community. (2026). scikit-fuzzy Documentation. https://pythonhosted.org/scikit-fuzzy/. [in Eng.].

SQLAlchemy Authors. (2026). SQLAlchemy Documentation. https://docs.sqlalchemy.org/. [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. [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.].

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. 236–247
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.015

DDPG FOR TRAJECTORY GENERATION

Z. Momynkulov[1]*, B. Omarov[2]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: mrjohnnyunited@gmail.com

Zeinel Momynkulov — master International Information Technology University, Almaty, Kazakhstan
E-mail: z.momynkulov@iitu.edu.kz, https://orcid.org/0000-0002-1918-065X;
Bauyrzhan Omarov — master, Al-Farabi Kazakh National University, Almaty, Kazakhstan https://orcid.org/0000-0002-9312-4429.

© Z. Momynkulov, B. Omarov

Abstract. Deep Deterministic Policy Gradient (DDPG) is an effective reinforcement learning method designed for continuous control tasks. In this work, we investigate the use of DDPG to generate optimal motion trajectories for robotic systems, focusing on smooth, efficient, and collision-free movements. The approach centers on designing a reward function that balances trajectory smoothness, successful task execution, and obstacle avoidance. A simulated training environment is constructed to develop and assess the learned policies. Results from the experiments show that DDPG is capable of producing high-quality trajectories. These outcomes emphasize the promise of deep reinforcement learning in enabling real-time trajectory optimization for robotic applications.

Keywords: deep deterministic policy gradient, reinforcement learning, optimal trajectory generation, continuous action space

For citation: Momynkulov Z., Omarov B. (2026). DDPG for trajectory generation // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 236–247. https://doi. org/10.54309/IJICT.2026.26.2.015. (In Eng.).

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

Acknowledgements. *This study was funded by the research project "BR24992947— Development of Robots, Scientific, Technical, and Software for Flexible Robotization and Industrial Automation (RPA) in Automotive Industrial Enterprises in Kazakhstan Using Artificial Intelligence".*

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

ТРАЕКТОРИЯЛАРДЫ ГЕНЕРАЦИЯЛАУ ҮШІН DDPG

З. Момынкулов[1]*, Б. Омаров[2]
[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.
E-mail: mrjohnnyunited@gmail.com

Момынкулов Зейнел — магистр, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: z.momynkulov@iitu.edu.kz, https://orcid.org/0000-0002-1918-065X;
Омаров Бауыржан — магистр, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-9312-4429.

© З. Момынкулов, Б. Омаров

Аннотация. Терең детерминирленген саясат градиенті (Deep Deterministic Policy Gradient, DDPG) — үздіксіз басқару міндеттерін шешуге арналған тиімді күшейте оқыту әдісі. Бұл жұмыста роботтандырылған жүйелер үшін оңтайлы қозғалыс траекторияларын генерациялау мақсатында DDPG әдісін қолдану зерттеледі. Зерттеу барысында қозғалыстардың бірқалыптылығын, тиімділігін және соқтығысусыз орындалуын қамтамасыз етуге баса назар аударылады. Тәсілдің негізгі бөлігі траекторияның бірқалыптылығын, тапсырманың сәтті орындалуын және кедергілерді айналып өтуді теңгеретін сыйақы функциясын әзірлеуге негізделеді. Үйретілген басқару саясаттарын әзірлеу және бағалау үшін модельденген оқыту ортасы құрылады. Эксперимент нәтижелері DDPG әдісінің жоғары сапалы траекторияларды генерациялай алатынын көрсетті. Алынған нәтижелер роботтандырылған қолданбаларда траекторияларды нақты уақыт режимінде оңтайландыру үшін терең күшейте оқыту әдістерін қолданудың перспективалылығын көрсетеді.

Түйінді сөздер: терең детерминирленген саясат градиенті, күшейте оқыту, траекторияларды оңтайлы генерациялау, үздіксіз әрекеттер кеңістігі

Дәйексөздер үшін: Момынкулов З., Омаров Б. (2026). Траекторияларды генерациялау үшін DDPG // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 236–247. https://doi.org/10.54309/IJICT.2026.26.2.015. (Ағыл. тіл.).

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

Алғыс: *Бұл зерттеу «BR24992947 — Қазақстандағы автомобиль өнеркәсібі кәсіпорындарында жасанды интеллектті пайдалана отырып, икемді роботтандыру және өнеркәсіптік автоматтандыру (RPA) үшін роботтарды, ғылыми-техникалық шешімдер мен бағдарламалық қамтамасыз етуді әзірлеу» ғылыми-зерттеу жобасы аясында қаржыландырылды.*

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

DDPG ДЛЯ ГЕНЕРАЦИИ ТРАЕКТОРИЙ

З. Момынкулов[1]*, Б. Омаров[2]
[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
E-mail: mrjohnnyunited@gmail.com

Момынкулов Зейнел — магистр Международного университета информационных технологий, Алматы, Казахстан
E-mail: z.momynkulov@iitu.edu.kz, https://orcid.org/0000-0002-1918-065X;
Омаров Бауыржан — магистр Казахского национального университета имени аль-Фараби, Алматы, Казахстан
https://orcid.org/0000-0002-9312-4429.

© З. Момынкулов, Б. Омаров

Аннотация. Метод глубокого детерминированного градиента политики (Deep Deterministic Policy Gradient, DDPG) представляет собой эффективный метод обучения с подкреплением, предназначенный для решения задач непрерывного управления. В настоящей работе исследуется применение метода DDPG для генерации оптимальных траекторий движения роботизированных систем с целью обеспечения плавности, эффективности и отсутствия столкновений при перемещении. Основное внимание уделяется разработке функции вознаграждения, обеспечивающей баланс между плавностью траектории, успешным выполнением поставленной задачи и избеганием препятствий. Для обучения и оценки получаемых политик управления формируется моделируемая среда обучения. Результаты экспериментов показывают, что метод DDPG способен генерировать траектории высокого качества. Полученные результаты свидетельствуют о перспективности применения методов глубокого обучения с подкреплением для оптимизации траекторий движения роботизированных систем в режиме реального времени.

Ключевые слова: глубокий детерминированный градиент политики, обучение с подкреплением, оптимальная генерация траекторий, непрерывное пространство действий

Для цитирования: Момынкулов З., Омаров Б. (2026). DDPG для генерации траекторий // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 236–247. https://doi.org/10.54309/IJICT.2026.26.2.015. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарности: *Настоящее исследование выполнено в рамках научно-исследовательского проекта «BR24992947 — Разработка роботов, научно-технических решений и программного обеспечения для гибкой роботизации и промышленной автоматизации (RPA) на предприятиях автомобильной промышленности Казахстана*

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

с использованием искусственного интеллекта».

Introduction.

Reinforcement Learning (RL) is a fundamental area of machine learning that allows agents to acquire behaviors by interacting with their environment through trial and error. Among RL methods, Deep Deterministic Policy Gradient (DDPG) is a highly effective algorithm for problems involving continuous action spaces. This work investigates the use of DDPG for discovering and generating optimal trajectories, as well as evaluating its overall performance. DDPG proves especially useful in scenarios where an agent must plan trajectories within a multidimensional continuous domain, such as robotic navigation or motion planning. The study discusses the core elements of the DDPG algorithm and its role in trajectory generation, where environmental feedback is shaped by the reward function. The range of tasks where DDPG can be applied is quite broad, especially in areas that require decision making in continuous and high-dimensional action spaces. One of the most common applications is robotic control, where the algorithm can be used to generate precise joint movements or end-effector trajectories. For example, robotic arms in industrial settings can be trained with DDPG to pick, place, or manipulate objects in a smooth and efficient way.

Mobile robots, drones, and self-driving vehicles all operate in environments where actions such as steering angle, velocity, or acceleration are continuous rather than discrete. DDPG provides a framework to learn optimal navigation policies that balance safety, efficiency, and goal achievement. In addition, DDPG finds applications in energy systems and resource optimization problems. For instance, in smart grid management or continuous production control, the algorithm can learn strategies that minimize energy cost while satisfying system constraints. Similarly, in finance or portfolio optimization, where decisions are continuous in nature, DDPG can support adaptive strategies for trading or allocation of assets.

Materials and methods.

Several studies highlight the application of reinforcement learning methods, including DDPG, in industrial environments.

For instance, in (Kober et al., 2013), a methodology is proposed for improving the precision of assembly robots through reinforcement learning by dynamically adjusting control stiffness, leading to enhanced contact precision and reduced errors. Similarly, the work in (Xue et al., 2018) explores the integration of AI-driven reinforcement learning models for managing fleets of automated guided vehicles (AGV) in manufacturing, optimizing task scheduling and improving system efficiency by preventing deadlocks. Furthermore, (Gong et al., 2022) introduces an optimized DDPG-based path planning method for mobile robots, enhancing their ability to navigate complex environments efficiently while ensuring collision-free movement. Several studies have highlighted the application of reinforcement learning methods, particularly DDPG, in industrial environments. The work in (Zhao et al., 2023) presents an enhanced DDPG algorithm incorporating a hierarchical memory structure, improving trajectory planning accuracy

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

and adaptability in robotic arm motion control. Path-following control of autonomous ground vehicles has been extensively explored using classical controllers (e.g., Pure Pursuit, Stanley, PID), nonlinear methods (sliding mode, backstepping), and, more recently, reinforcement learning (RL). Traditional approaches offer stability and real-time feasibility but struggle with nonlinear dynamics and parameter tuning. Nonlinear adaptive controllers improve robustness yet still rely heavily on accurate models. To overcome these limitations, deep RL methods have emerged as model-free alternatives. In work (Cao et al., 2023) applied the Deep Deterministic Policy Gradient (DDPG) algorithm for vehicle path following, achieving smoother steering and reduced cross-track error compared to conventional methods. However, the given approach focuses only on control and simulation scenarios and leaving open challenges in joint speed-steering coordination, generalization to diverse environments, and integration with model-oriented control. The present work addresses these issues by combining data-driven learning with model-based constraints for improved adaptability and performance.

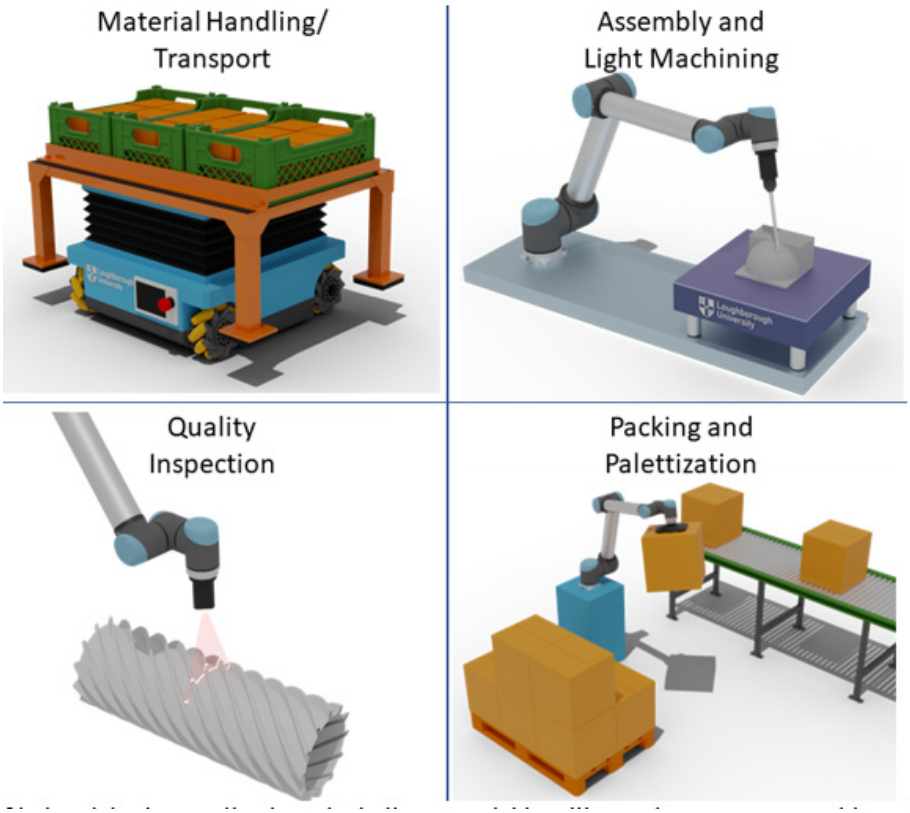


Fig. 1. Examples of industrial robot applications, including material handling and transport, assembly and light machining, automated quality inspection, and packing with palletization (Buerkle et al. 2023).

In Figure 2 the initial positions are the origin. The graph on the right side demonstrates the errors observed in the left given subplot (Cao et al., 2023).

Another work (Chen et al., 2022) addresses trajectory planning for a 7-DOF

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

manipulator in uncertain, dynamic environments by applying model-free deep reinforcement learning (DRL). They compare the stochastic policy-based Soft Actor-Critic (SAC) and the deterministic Deep Deterministic Policy Gradient (DDPG) to achieve collision-free manipulation under uncertainty. Their key contributions include: devising a dense reward function combining distance to goal and distance from obstacles; using relative rather than absolute state representations for better learning convergence; demonstrating that SAC achieves higher stability, efficiency and zero-collision performance after around 5000 episodes, whereas DDPG shows higher variance and slower convergence. While this work advances DRL for high-dimensional manipulators in dynamic settings, it still remains in simulation, does not integrate explicit model-based planning/constraints, and separates action types (end-effector motion only) rather than holistic vehicle-manipulator control.

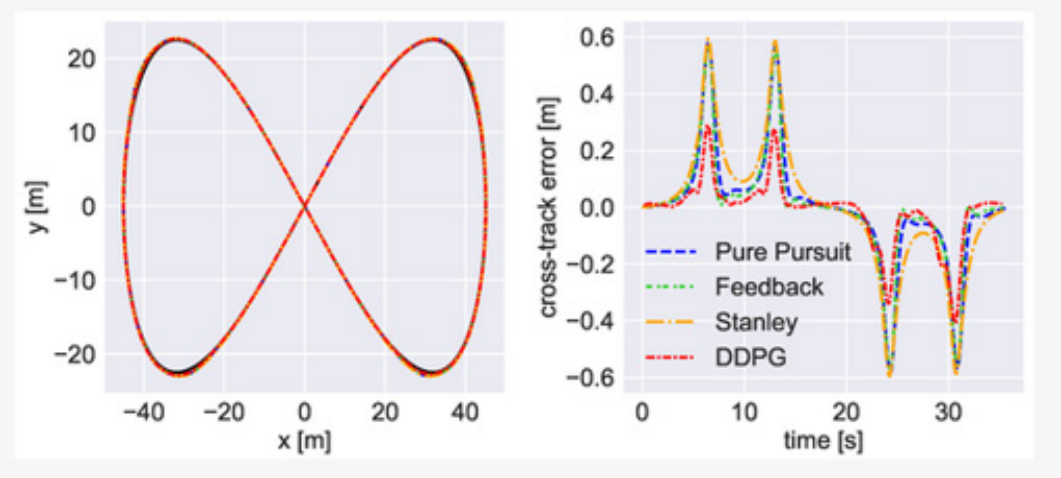


Fig. 2. The trajectories for calculating the cross-track error (Cao et al., 2023).

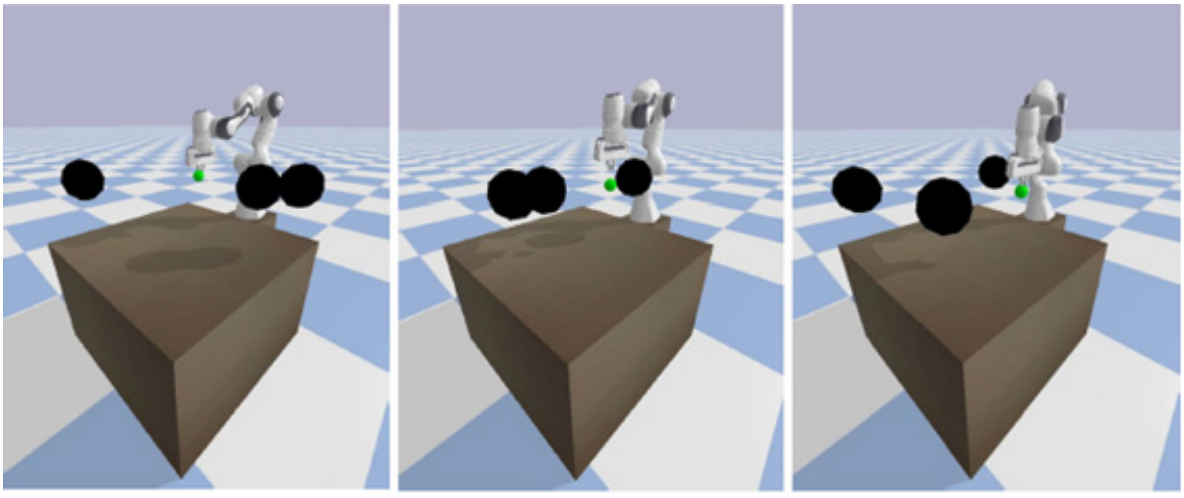


Fig. 3. First, the robot learns to reach the goal. Then avoid collisions with dynamic obstacles. Third, keep reaching the target position (Chen et al., 2022).

The development of deterministic policy gradient methods was preceded by the

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

emergence of stochastic policy gradient approaches, which formed the basis of early policy optimization techniques. Traditional reinforcement learning algorithms such as Q-learning (Watkins et al., 1992) and REINFORCE (Williams, 1992) were mainly suited for discrete action spaces, limiting their effectiveness in domains that require precise continuous control. To address this challenge, Silver et al. proposed the Deterministic Policy Gradient (DPG) algorithm, which optimizes a deterministic policy directly instead of relying on action sampling from probability distributions (Williams, 1992). Through the deterministic policy gradient theorem, DPG provided a more stable and efficient framework for learning in continuous control environments.

Building on the foundation of DPG, Lillicrap et al. introduced its deep learning extension, known as the Deep Deterministic Policy Gradient (DDPG) algorithm (Silver et al., 2014). This method combines the power of deep neural networks with the actor–critic framework, where the actor is responsible for learning a deterministic policy and the critic evaluates the value function. The design borrows elements from the Deep Q-Network architecture, enabling effective learning in high-dimensional continuous action spaces.

To ensure training stability, DDPG employs a replay buffer, which stores past interactions and allows the algorithm to learn from a diverse set of previous experiences. This mechanism reduces the problem of correlated updates that often destabilize reinforcement learning. In addition, DDPG makes use of target networks-slower-updating copies of the actor and critic-that introduce delayed updates and therefore mitigate variance, further stabilizing the learning process.

Unlike classical Q-learning methods restricted to discrete actions, DDPG is tailored for continuous domains. This makes it highly suitable for advanced control problems, such as robotic motion and autonomous manipulation, where agents must choose from an infinite range of possible actions to achieve fine-grained control over their environment (Lillicrap et al., 2015). At the core of DDPG is the actor–critic architecture. This design uses two neural networks that operate together. The Actor network is responsible for deciding which action to take in a given state, effectively representing the agent's strategy or policy. The Critic network, on the other hand, evaluates the quality of the chosen action by estimating how good it is in terms of expected future rewards. The Actor therefore learns how to act, while the Critic learns how to judge the Actor's choices. The Actor is improved by following feedback from the Critic, and the Critic itself is trained using gradient-based optimization methods to minimize the difference between its predictions and the actual outcomes. Another key element is the replay buffer. In reinforcement learning, experiences collected by the agent are often highly correlated because they come from sequential interactions with the environment. If the algorithm learns directly from this sequential data, it risks overfitting to recent experiences and failing to generalize. To address this, DDPG stores past experiences in a memory buffer. During training, random samples of experiences are drawn from this buffer to train the networks. This randomization breaks the correlation between consecutive experiences, smooths the learning process, and improves the efficiency of training.

Together, these components-value estimations through the Q-function (1), the actor–

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

critic framework, the use of target networks, and the replay buffer-form the theoretical foundation of DDPG. Each element plays a crucial role in enabling stable and efficient learning in environments with continuous action spaces, making the algorithm especially effective for tasks such as robotic control and autonomous navigation.

$$Q^{\pi}(s_t, a_t) = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a, \pi)] \quad (1)$$

Application of DDPG to the problem of trajectory generation

In continuous action spaces, the task begins with two defined points: a starting point and a target point. The goal is to generate a trajectory from the start to the target in the most efficient and shortest possible way. In reinforcement learning, this requires the design of a reward function that provides positive feedback to encourage maximization of performance. To increase the complexity of the task, additional points are introduced that act as obstacles, preventing a direct path to the goal. As a result, the problem can be divided into main challenge:

- A task that contains obstacles;
- A task with multiple goals.

Current approach uses three components of the reward function. The first component is the agent's distance to the target. The agent is penalized by the negative distance to the current target (2).

$$R_{dist} = -\|position - obstacle\| \quad (2)$$

The *position* refers to the current location of the trajectory, while the obstacle represents the location of a barrier in the environment. In the second case, it is necessary to consider how the agent behaves when encountering an obstacle. For this experiment, two variables were introduced: *dist_to_obstacle* and *penalty_dist_to_obstacle.* The first indicates the distance between the agent and the obstacle, and the second specifies the penalty applied to the agent at each step of the trajectory (3).

$$R_{obst} = \begin{cases} -10, if \|position - obstacle\| < 0.2 \\ 0, else \end{cases} \quad (3)$$

Next, it is necessary to assign a reward for reaching the goal. For this purpose, two variables are introduced: *goal_distance* and *goal_reward,* with values set to 0.1 and 10, respectively. Since multiple goals may exist, defining these values in advance is practical. This setup means that when the agent approaches a goal within a distance of less than 0.1, it will receive a reward of 10. Such a reward acts as an incentive for the agent to successfully complete its tasks (4).

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

$$R_{goal} = \begin{cases} +10, \wedge \, if \, \|position - goal\| < 0.1 \\ 0, \wedge \, else \end{cases} \quad (4)$$

The final reward function will be as follows (5).

$$R_{total} = -\|position - obstacle\| + \sum_{obstacle} R_{obst} + R_{goal} \quad (5)$$

Results and discussion.

The training process was divided into two stages. To ensure representativeness, the agent was trained for only 200 episodes. In the first experiment, the agent was assigned a single goal, with the task of reaching it while avoiding obstacles. The training was performed on an NVIDIA RTX 4070 GPU with CUDA support, and the total duration was approximately two hours. For visualization, the start position is indicated by a green dot, obstacles are marked in red, and the target points are shown in blue. Figures 2 illustrate the agent's learning process.

It is important to note how the agent progressively learns to reach all designated points. By episode 74, the trajectories identified by the agent already show attempts to pass through the target locations. By episode 141, the paths become more consistent and reliable, and this trend continues. Nevertheless, it should be emphasized that improved outcomes could be achieved by extending the number of training episodes and refining the reward function.

When compared with traditional trajectory planning approaches such as A* or linear interpolation, the deep deterministic policy gradient algorithm demonstrates clear advantages, especially in dynamic and continuous domains. A* relies on discrete state spaces and requires a predefined map of the environment, while linear interpolation simply connects waypoints with straight lines, often ignoring obstacles. In contrast, DDPG learns a control policy directly within continuous state and action spaces, enabling the generation of smooth and adaptive trajectories that respect physical constraints while avoiding obstacles, even in partially observable or changing environments. Furthermore, DDPG optimizes agent behavior based on cumulative rewards, which makes it possible to incorporate complex objectives such as energy efficiency, trajectory smoothness, or safety margins. Unlike classical methods that must be recalculated whenever the environment changes, a trained DDPG agent is capable of generalizing to new, unseen conditions. This adaptability makes DDPG particularly effective for real-time robotic control, where both perception and decision-making must function continuously and flexibly. Figure 3 illustrates all trajectories for comparison.

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

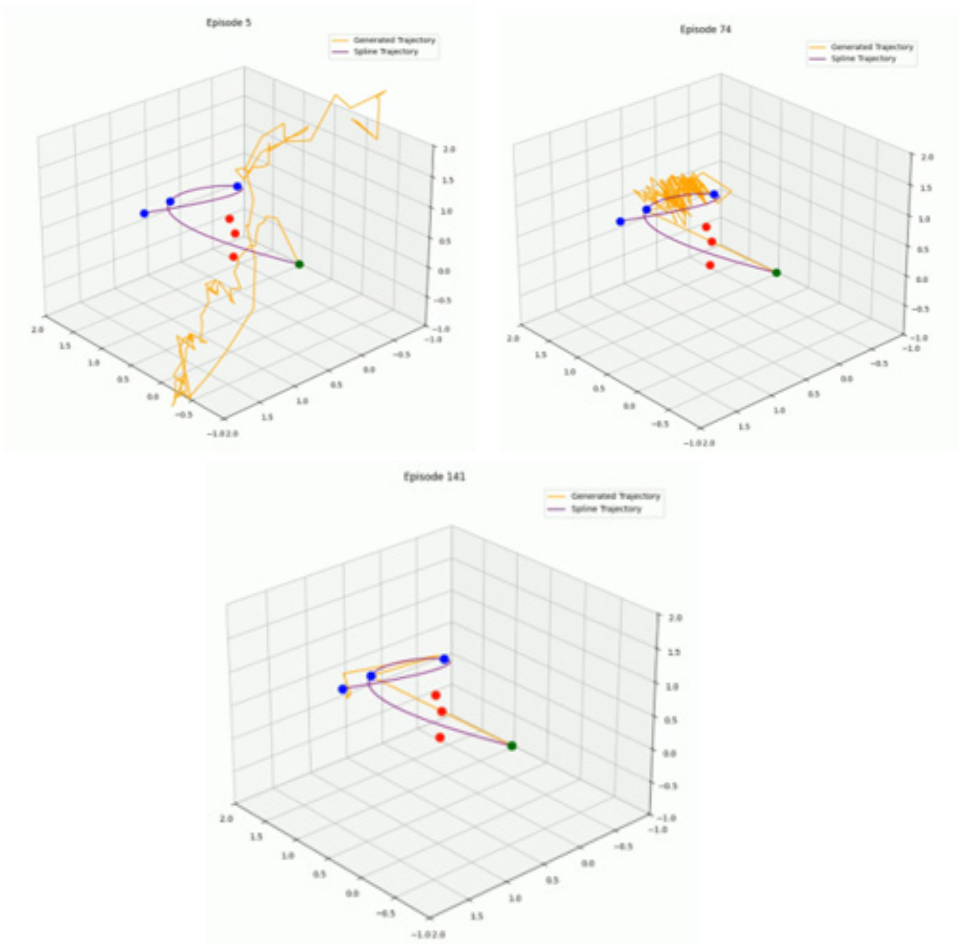


Fig. 4. Agent learning process demonstration from different episodes.

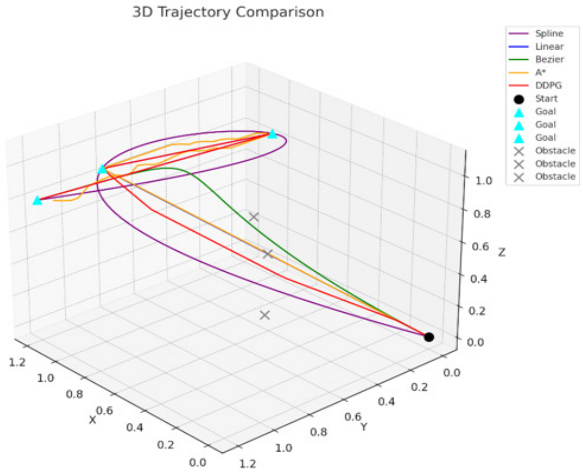


Fig. 5. Comparison of different algorithms.

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

Table 1. Comparison of algorithms for task

Method	Frequency	MAE
Linear	0.386	0.808
Bezier	0.487	0.771
A*	0.370	0.777
DDPG	0.389	0.796

Table 1 provides a quantitative comparison of five trajectory generation methods: Spline (used as the reference), Linear Interpolation, Bezier Curve, A* Search, and DDPG. The evaluation relies on two main metrics: Mean Absolute Error (MAE) and Maximum Error, both measured relative to the spline trajectory. Linear interpolation demonstrates a slightly higher error (MAE = 0.39), highlighting its limitations in managing curvature and ensuring path continuity. The Bezier curve, although producing the smoothest visual path, results in the largest deviation (MAE = 0.49) because it lacks precise control over intermediate goal positions. The DDPG-based trajectory performs at a level similar to A*, with an MAE of 0.39, showing improvement over linear interpolation but still not reaching the accuracy of the reference spline. This gap may be due to restricted training episodes or incomplete policy optimization.

Conclusion.

This paper demonstrated the use of the DDPG algorithm for trajectory generation by targets. The results show potential for future applications in training robots and manipulators, with improvements possible through reward function adjustments. Future work will focus on refining the reward design, testing alternative reinforcement learning methods, and enhancing stability and sample efficiency. Another important direction is the inclusion of safety constraints and collision avoidance to increase real-world applicability. Finally, integration with a physical robotic manipulator will allow real-time testing in dynamic environments, ensuring smoother, more precise, and faster execution, ultimately strengthening the model's robustness for industrial use.

REFERENCES

Buerkle A., Eaton W., Al-Yacoub A., Zimmer M., Kinnell P., Henshaw M., Coombes M., Chen W. H., and Lohse N. (2023). Towards industrial robots as a service (IRaaS): Flexibility, usability, safety and business models // Robotics and Computer-Integrated Manufacturing. — Vol. 81. — Article 102484. DOI: 10.1016/j.rcim.2022.102484. [in Eng.].

Cao, Z., Zheng, S., Liu, J., Li, L., Xiao, J. (2023). Path Following for Autonomous Ground Vehicle Using DDPG Algorithm: A Reinforcement Learning Approach. Applied Sciences. —Vol. 13. — No. 11. Pp. 6847. [in Eng.].

Chen, C., Liu, Z., Liu, X., Wang, X., Wu, Y., Zhang, W. (2022). Deep Reinforcement Learning-Based Trajectory Planning Under Uncertain Constraints. Frontiers in Neurorobotics, 16:883562. DOI: https://doi.org/10.3389/fnbot.2022.883562[in Eng.].

Gong H., Wang P., Ni C., and Cheng N. (2022). Efficient path planning for mobile robot based on deep deterministic policy gradient. Sensors. — Vol. 22. — No. 9. Pp. 3579. [in Eng.].

Kober J., Bagnell J.A., and Peters J. (2013). Reinforcement learning in robotics: A survey. Int. J. Robot. Res. — Vol. 32. — No. 11. Pp. 1238–1274. [in Eng.].

Lillicrap T. P. et al. (2015). Continuous control with deep reinforcement learning. arXiv (Cornell University). [in Eng.].

Liu Z., Liu Q., Xu W., Wang L., and Zhou Z. (2022). Robot learning towards smart robotic manufacturing: A review. Robotics and Computer-Integrated Manufacturing. — Vol. 77. — Pp. 102360. DOI: 10.1016/j.rcim.2022.102360.

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

[in Eng.].

Silver D., Lever G., Heess N., Degris T., Wierstra D., and Riedmiller M. (2014). Deterministic Policy gradient algorithms. PMLR. [in Eng.].

Watkins C. J. C. H. and Dayan P. (1992). Q-learning," Machine Learning. — Vol. 8. —No. 3–4. Pp. 279–292. [in Eng.].

Williams R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning. — Vol. 8. — No. 3–4. Pp. 229–256. [in Eng.].

Xue T., Zeng P., and Yu H. (2018). A reinforcement learning method for multi-AGV scheduling in manufacturing. 2022 IEEE International Conference on Industrial Technology (ICIT), Pp. 1557–1561, Feb. 2018. [in Eng.].

Zhao D., Ding Z., Li W., Zhao S., and Du Y. (2023). Robotic Arm Trajectory Planning Method using Deep Deterministic Policy Gradient with Hierarchical Memory Structure. IEEE Access. — Vol. 11. — Pp. 140801–140814. [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.2. Number 26 (2026). Pp. 248–259
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.016

«WHY IS THIS NEWS NEGATIVE?»: AN ATTENTION-BASED VISUALIZATION TECHNIQUE FOR EXPLAINABLE SENTIMENT ANALYSIS. AN ANALYTICAL SURVEY AND CONCEPTUAL FRAMEWORK

F. Myrzabekov*, N. Mustafin, M. Mukushev
Astana IT University, Astana, Kazakhstan.
E-mail: fmyrzabekov@gmail.com

Farkhat Myrzabekov — PhD, Senior Lecturer, Astana IT University, Astana, Kazakhstan https://orcid.org/0009-0004-9866-1495;
Nurzhan Mustafin — Astana IT University, Astana, Kazakhstan https://orcid.org/0009-0002-0454-0322;
Medet Mukushev — PhD, Assistant professor, Astana IT University, Astana, Kazakhstan https://orcid.org/0000-0002-3655-992.

© F. Myrzabekov, N. Mustafin, M. Mukushev

Abstract. The proliferation of Transformer-based architectures has marked a new epoch in Natural Language Processing, with sentiment analysis systems achieving unprecedented accuracy. This performance, however, is coupled with a profound lack of transparency, rendering these models as opaque «black boxes.» This interpretability deficit is a significant barrier to their deployment in high-stakes domains like financial analysis and public policy, where the justification for a decision is as critical as the decision itself. This paper conducts a comprehensive survey of this problem and proffers a conceptual framework for an explainable sentiment analysis system. We propose leveraging the Transformer's intrinsic self-attention mechanism to generate salient visualizations that elucidate the textual evidence driving a model's classification. The paper examines the theoretical underpinnings of this method, engages with the pivotal «Attention is not Explanation» debate, and charts a course for future research. We present the architecture of a system engineered not merely to classify sentiment but to furnish a transparent rationale for its predictions, thereby fostering greater trust and accountability in advanced AI systems.

Keywords: Explainable AI (XAI), Sentiment Analysis, Natural Language Processing (NLP), Attention Mechanism, Transformers, Data Visualization, Model Interpretability

For citation: Myrzabekov F., Mustafin N., Mukushev M. (2026). Why is this news negative?: an attention-based visualization technique for explainable sentiment analysis. An analytical survey and conceptual framework // International journal of information

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

and communication technologies. Vol. 7. No. 26. Pp. 248–259. https://doi. org/10.54309/IJICT.2026.26.2.016. (In Eng.).

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

«БҰЛ ЖАҢАЛЫҚ НЕЛІКТЕН ТЕРІС?»: ТҮСІНДІРІЛЕТІН СЕНТИМЕНТТІК ТАЛДАУҒА АРНАЛҒАН НАЗАР АУДАРУ МЕХАНИЗМІНЕ НЕГІЗДЕЛГЕН ВИЗУАЛИЗАЦИЯ ӘДІСІ. ТАЛДАУЛЫҚ ШОЛУ ЖӘНЕ ТҰЖЫРЫМДАМАЛЫҚ НЕГІЗ

Ф. Мырзабеков, Н. Мустафин, М. Мукушев*
Astana IT University, Астана, Қазақстан.
E-mail: fmyrzabekov@gmail.com

Мырзабеков Фархат — PhD докторы, сеньор-лектор, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0004-9866-1495;
Мустафин Нуржан — Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0002-0454-0322;
Мукушев Медет — PhD, ассистент-профессор, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-3655-992.

© Ф. Мырзабеков, Н. Мустафин, М. Мукушев

Аннотация. Трансформерлерге негізделген архитектуралардың кеңінен таралуы табиғи тілді өңдеу саласында жаңа дәуірдің басталуына ықпал етті, бұл ретте сентименттік талдау жүйелері бұрын-соңды болмаған жоғары дәлдікке қол жеткізді. Алайда мұндай жоғары өнімділік модельдердің ашықтығының елеулі деңгейде болмауымен қатар жүреді, соның нәтижесінде олар іс жүзінде түсіндірмесі қиын «қара жәшік» модельдеріне айналады. Интерпретацияланудың жеткіліксіздігі бұл модельдерді қаржылық талдау және мемлекеттік саясат сияқты шешімдердің негіздемесі шешімнің өзімен бірдей маңызды болып табылатын жоғары жауапкершілікті салаларда қолдануға елеулі кедергі келтіреді. Бұл мақалада аталған мәселенің жан-жақты талдауы жүргізіліп, түсіндірілетін сентименттік талдау жүйесінің тұжырымдамалық негізі ұсынылады. Біз модельдің жіктеу шешіміне әсер ететін мәтіндік белгілерді айқындайтын ақпараттық визуализацияларды қалыптастыру үшін трансформердің ішкі өзіндік назар аудару (self-attention) механизмін пайдалануды ұсынамыз. Мақалада осы әдістің теориялық негіздері қарастырылып, «Attention is not Explanation» («Назар аудару механизмі түсіндірме болып табылмайды») атты маңызды ғылыми пікірталас талданады, сондай-ақ болашақ зерттеулердің перспективалы бағыттары айқындалады. Мәтіннің сентименттік сипатын жіктеп қана қоймай, сонымен қатар өз болжамдарына қатысты ашық әрі түсінікті негіздеме ұсынуға бағытталған жүй-

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

енің архитектурасы берілген. Бұл тәсіл заманауи жасанды интеллект жүйелеріне деген сенімділікті және олардың есептілігі деңгейін арттыруға ықпал етеді.

Түйінді сөздер: түсіндірілетін жасанды интеллект (XAI), сентименттік талдау, табиғи тілді өңдеу (NLP), назар аудару механизмі, трансформерлер, деректерді визуализациялау, модельдің интерпретациялануы

Дәйексөздер үшін: Мырзабеков Ф., Мустафин Н., Мукушев М. (2026). Бұл жаңалық неліктен теріс?: түсіндірілетін сентименттік талдауға арналған назар аудару механизміне негізделген визуализация әдісі. Талдаулық шолу және тұжырымдамалық негіз // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 248–259. https://doi.org/10.54309/IJICT.2026.26.2.016. (Ағыл. тіл.).

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

ПОЧЕМУ ЭТА НОВОСТЬ НЕГАТИВНА?: МЕТОД ВИЗУАЛИЗАЦИИ НА ОСНОВЕ МЕХАНИЗМА ВНИМАНИЯ ДЛЯ ОБЪЯСНИМОГО АНАЛИЗА ТОНАЛЬНОСТИ. АНАЛИТИЧЕСКИЙ ОБЗОР И КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ

Ф. Мырзабеков, Н. Мустафин, М. Мукушев*
Astana IT University, Астана, Казахстан.
E-mail: fmyrzabekov@gmail.com

Мырзабеков Фархат — доктор PhD, сеньор-лектор, Astana IT University, Астана, Казахстан
https://orcid.org/0009-0004-9866-1495;
Мустафин Нуржан — Astana IT University, Астана, Казахстан
https://orcid.org/0009-0002-0454-0322;
Мукушев Медет — PhD, ассистент-профессор, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-3655-992.

© Ф. Мырзабеков, Н. Мустафин, М. Мукушев

Аннотация. Распространение архитектур на основе трансформеров ознаменовало новую эпоху в области обработки естественного языка, в которой системы анализа тональности достигли беспрецедентно высокой точности. Однако такая производительность сопровождается существенным недостатком прозрачности, вследствие чего данные модели фактически представляют собой непрозрачные «черные ящики». Этот дефицит интерпретируемости является существенным препятствием для применения подобных моделей в критически важных областях, таких как финансовый анализ и государственная политика, где обоснование принятого решения имеет такое же важное значение, как и само решение. В настоящей статье проводится всесторонний обзор данной проблемы и предлагается концептуальная модель системы

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

объяснимого анализа тональности. Мы предлагаем использовать внутренний механизм самообслуживания внимания (self-attention) трансформера для формирования информативных визуализаций, позволяющих выявить текстовые признаки, определяющие классификацию, выполненную моделью. В статье рассматриваются теоретические основы данного подхода, анализируется ключевая дискуссия «Attention is not Explanation» («Механизм внимания не является объяснением»), а также определяются перспективные направления дальнейших исследований. Представлена архитектура системы, разработанной не только для классификации тональности текста, но и для предоставления прозрачного обоснования своих прогнозов, что способствует повышению доверия к современным системам искусственного интеллекта и уровня их подотчетности.

Ключевые слова: объяснимый искусственный интеллект (XAI), анализ тональности, обработка естественного языка (NLP), механизм внимания, трансформеры, визуализация данных, интерпретируемость модели

Для цитирования: Мырзабеков Ф., Мустафин Н., Мукушев М. (2026). Почему эта новость негативна?: метод визуализации на основе механизма внимания для объяснимого анализа тональности. Аналитический обзор и концептуальная модель // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 248–259. https://doi.org/10.54309/IJICT.2026.26.2.016. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The contemporary digital landscape is characterized by an unprecedented deluge of unstructured text from news media, social platforms, and other digital fora. Automated sentiment analysis has evolved into a cornerstone technology within Natural Language Processing (NLP) (Pang et al., 2008), with applications influencing financial trading, public health responses, and corporate strategy.

The evolution of sentiment analysis models has been swift, progressing from lexicon-based systems to the current state-of-the-art: large-scale, pre-trained Transformer architectures such as BERT (Devlin et al., 2019). These models have consistently shattered performance benchmarks by learning deep, contextual representations of language. Yet, this success has engendered a critical challenge: the profound opacity of their decision-making processes. The millions of parameters within these networks form a complex system whose internal logic is largely inscrutable to human observers. This «black box» problem creates a crisis of trust and accountability, as a model's prediction is often useless without understanding the reasoning behind it.

This demand for transparency has catalyzed the emergence of Explainable AI (XAI) as a vital sub-discipline (Gunning, 2017). The core mandate of XAI is to develop systems that can provide clear and understandable justifications for their outputs. This paper addresses this imperative by proposing a conceptual framework for an explainable sentiment analysis system. Our central thesis is that the self-attention mechanism, an integral component of the Transformer architecture (Vaswani et al., 2017), can be

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

repurposed as an effective tool for explanation. We argue that attention weights can be systematically extracted, aggregated, and visualized to illuminate the textual fragments that most influenced a sentiment classification.

The primary contribution of this work is to provide a holistic and critical treatment of this attentionbased explanation approach. We conduct a synthesis of the literature, present a detailed conceptual architecture, and perform a rigorous analysis of the method's limitations and ethical considerations. Our goal is to provide a robust conceptual foundation for developing NLP systems that are not only accurate but also transparent and accountable.

Materials and methods.

Literature review: Foundations and related work

Our framework is situated at the confluence of sentiment analysis, Transformer architectures, and XAI for NLP.

The methodological trajectory of sentiment analysis

The progression of sentiment analysis reflects a search for more context-aware language representations. Early systems were rule-based, using lexicons with pre-assigned polarity scores (Turney, 2002). These were superseded by classical machine learning (e.g., SVMs) on Bag-of-Words (BoW) features (Pang et al., 2002), which lacked syntactic understanding.

A major leap came with deep learning, specifically word embeddings (e.g., Word2Vec (Mikolov et a., 2013)) and Recurrent Neural Networks (RNNs) like LSTMs (Hochreiter et al., 1997), which could capture sequential context. However, their sequential nature posed computational bottlenecks. The Transformer architecture (Vaswani et al., 2017) resolved these issues with its parallel processing capability and the self-attention mechanism, enabling the paradigm of large-scale pre-training and task-specific fine-tuning seen in models like BERT (Devlin et al., 2019).

The self-attention mechanism allows a model to weigh the importance of all other tokens in a sequence when encoding a particular token. For an input sequence of token embeddings X, it computes Query (Q), Key (K), and Value (V) matrices by multiplying X with learned weight matrices. The attention output is then calculated as:

$$Attention\left(Q,K,V\right)= softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is the dimensionality of the key vectors. This operation, illustrated in Figure 1, is performed in parallel across multiple «attention heads,» allowing the model to capture diverse syntactic and semantic relationships. The complexity arising from these interactions motivates the search for model-specific explanation techniques.

XAI methodologies can be broadly categorized as model-agnostic or model-specific. Model-agnostic methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg et al., 2017) offer flexibility but often incur significant computational costs. Attention-based explanation, being model-specific, is computationally efficient.

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

However, its validity rests on the assumption that attention weights are a faithful proxy for feature importance. This was challenged by Jain and Wallace (Jain et al., 2019), who argued that «Attention is not Explanation.» A counterargument was presented by Wiegreffe and Pinter (Wiegreffe et al., 2019), who contended that with proper controls, «Attention is not *not* Explanation.» More recent work continues to explore this debate, investigating the faithfulness of saliency-based methods for Large Language Models (Sood et al., 2023) and surveying the broader challenges in NLP interpretability (Alqaraawi et al., 2024). Our work is positioned within this dialogue, advocating for the pragmatic utility of attention as an explanatory tool while acknowledging its contested status.

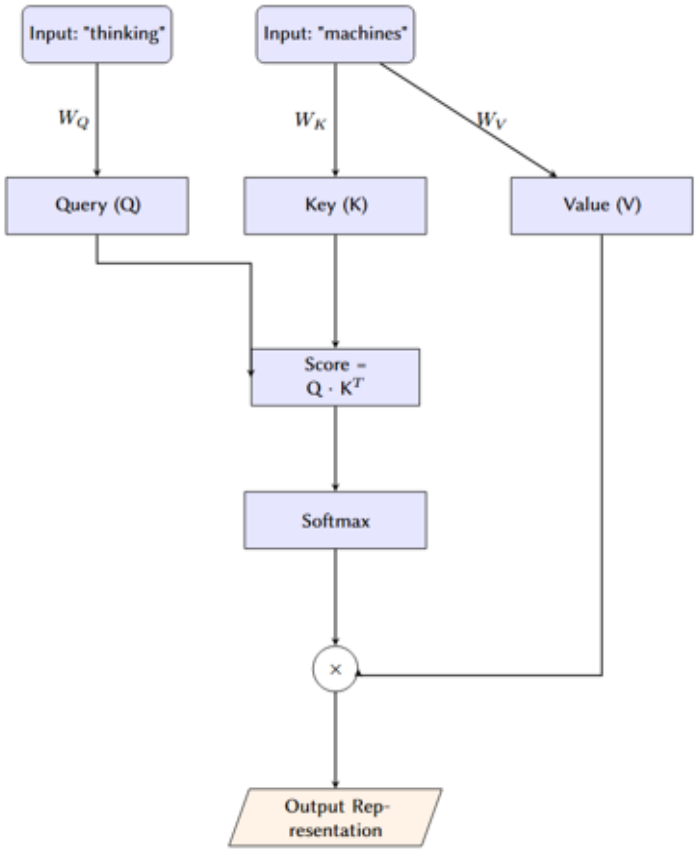


Fig. 1. Simplified visualization of the self-attention mechanism.

When processing the word «thinking,» the model computes a Query vector. This is compared against Key vectors from all other words (here, «machines») to generate a score. The scores are normalized via softmax to create attention weights, which then scale the Value vectors to produce a new, context-aware representation.

We delineate a modular framework for an end-to-end explainable sentiment analysis system. The proposed architecture follows a logical pipeline from data ingestion to explanation delivery, as illustrated in Figure 2.

At the system's core is a pre-trained Transformer like BERT (Devlin et al., 2019)

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

or RoBERTa (Liu et al., 2019). The model is finetuned on a domain-specific dataset for sentiment classification, preserving its linguistic knowledge while adapting to the task.

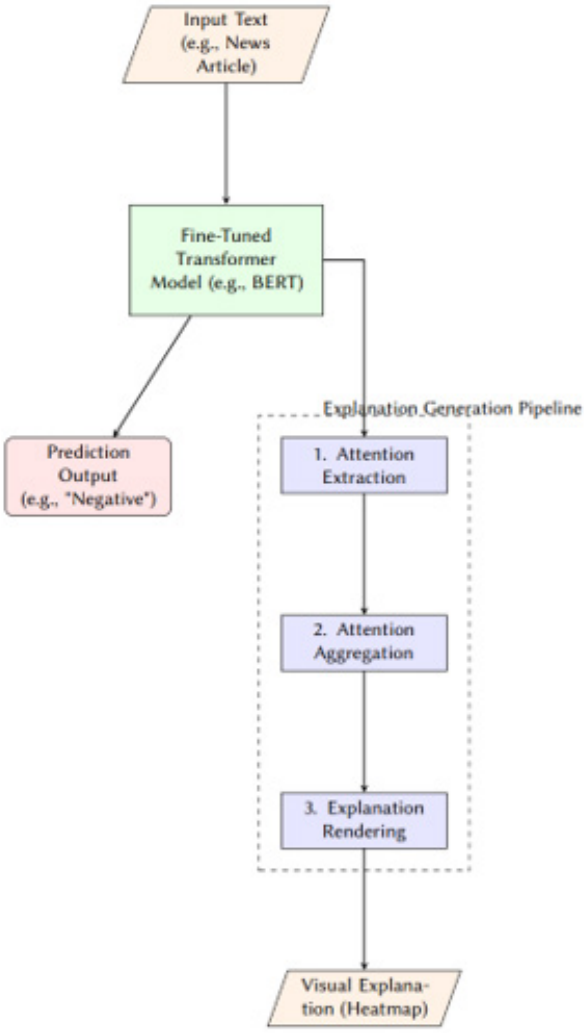


Fig. 2. Conceptual architecture of the explainable sentiment analysis system.

This subsystem extracts and processes attention weights to generate token-level importance scores.

Attention Extraction: The system retrieves raw attention maps from all layers and heads by enabling the `output_attentions` flag in the model configuration. For BERT-base, this yields 144 attention matrices per input.

Attention Aggregation: Raw weights are too high-dimensional for direct interpretation. We reduce them to a single saliency score per token by averaging across all heads and layers and normalizing the result. This method is computationally lightweight and effective for producing interpretable rankings.

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

Explanation Rendering: The final saliency scores are mapped to color intensities to create a heatmap. Highly attended tokens receive darker shades, allowing a user to see which words contributed most to the prediction. A sample output is shown in Figure 3.

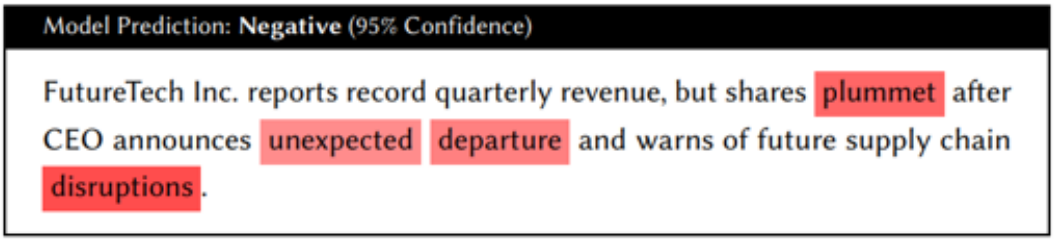


Fig. 3. Example of a visual explanation.

The color intensity of each word corresponds to its calculated attention score, highlighting the key terms driving the negative classification.

The following Python code demonstrates the core logic of the explanation pipeline, from model loading to saliency mapping.

Python Implementation of the Explanation Pipeline

```
from transformers import (BertTokenizer,
                          BertForSequenceClassification)
import torch

# 1. Load model with attention output enabled
model = BertForSequenceClassification.from_pretrained(
    "bert-base-uncased", output_attentions=True
)
tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
text = "The service was slow but the food was excellent."
inputs = tokenizer(text, return_tensors="pt")

# 2. Forward pass to extract attentions
outputs = model(**inputs)
attentions = outputs.attentions  # Tuple of 12 tensors

# 3. Aggregate attentions into token-level scores
# Stack tensors, then average over layers and heads
att_tensor = torch.stack(attentions).squeeze(1)
att_tensor = att_tensor.mean(dim=1) # Average over heads

# Get attention to [CLS] token, often used for classification
cls_attentions = att_tensor[:, 0, :]
token_importance = cls_attentions.mean(dim=0) # Average over layers

# Normalize scores
token_importance = token_importance / token_importance.sum()

# 4. Prepare final list of (token, score) for visualization
tokens = tokenizer.convert_ids_to_tokens(inputs["input_ids"][0])
explanation = list(zip(tokens, token_importance.tolist()))

for tok, score in explanation:
    print(f"{tok:12s} -> {score:.4f}")
```

Fig. 4. Core logic of the explanation pipeline.

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

This approach directly uses the model's internal states, adding negligible computational overhead while producing interpretable, token-level attributions.

An algorithmic trading firm analyzes a press release: «FutureTech Inc. reports record revenue, but shares *plummet* after CEO's *unexpected departure* and warns of supply chain *disruptions*.» A standard model outputs «Negative.» The explainable system confirms this but also highlights «plummet,» «unexpected departure,» and «disruptions.» This allows an analyst to verify that the model correctly identified forward-looking risks over historical results, validating a potential automated trade.

Officials monitoring social media see: The new vaccine is a *miracle,* I can't believe how *ineffective* it is. A total *disaster.* A model might be confused by «miracle.» The explainable system would classify the post as negative, strongly highlighting «ineffective» and «disaster.» This shows the analyst that the model correctly identified sarcasm, preventing a misleading data point from skewing public health analytics.

Results and discussion.

A principled evaluation of this framework requires examining its limitations and the open research questions that arise.

A primary critique concerns the distinction between fidelity and faithfulness, as illustrated in Figure 5. Fidelity measures how well the explanation matches the model's output, while faithfulness measures how well it reflects the model's internal reasoning. As shown by Jain and Wallace (Jain et al., 2019), attention weights may not always be faithful, as different attention patterns can produce identical predictions.

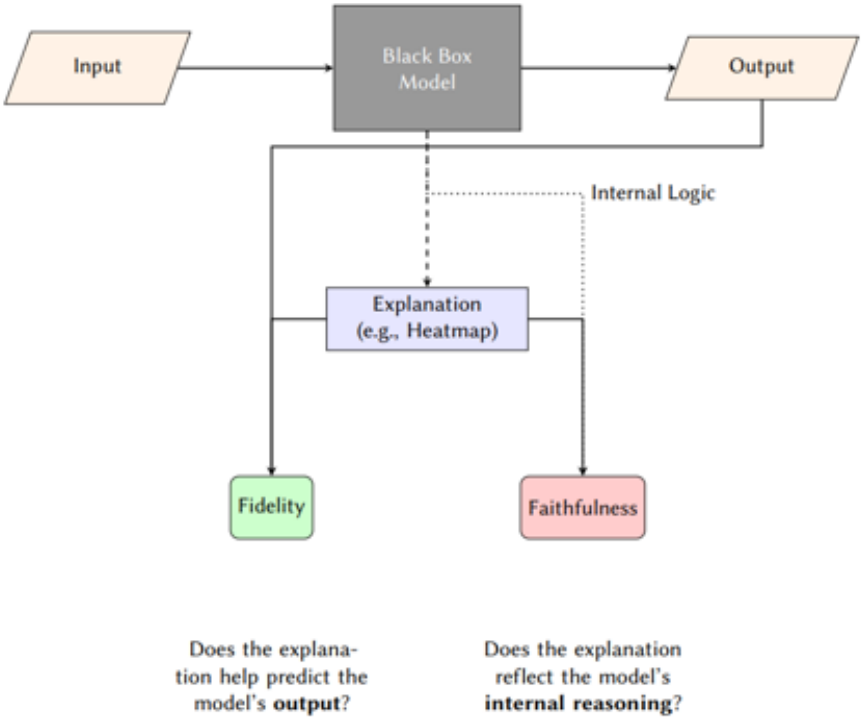


Fig. 5. The dilemma of Fidelity vs. Faithfulness in XAI.

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

An explanation has high fidelity if it correlates well with the model's predictions, but it is only faithful if it accurately represents the internal mechanisms that caused the prediction. Attention-based methods are often critiqued for having fidelity without guaranteed faithfulness.

Furthermore, the aggregation of multi-layer, multi-head attention into a single score lacks strong theoretical grounding. Different heads often specialize in distinct linguistic functions (Clark et al., 2019), and simple averaging may obscure these crucial signals.

Finally, token-level heatmaps struggle with phenomena involving multi-token interactions, such as negation or sarcasm. In the phrase «anything but boring,» sentiment arises from the compositional meaning, which a simple heatmap may misrepresent by highlighting «boring.»

The deployment of XAI systems introduces unique ethical challenges.

Automation Bias: An intuitive explanation can induce «automation bias,» where a user uncritically accepts the explanation as ground truth. Users must be trained to treat explanations as a guide for inquiry, not an infallible verdict.equal contribution mark: a superscript dagger (†) after the corresponding author mark or the affiliation mark.

Faithful Explanations of Unfair Models: XAI can expose the biased reasoning of an unfair model. This transparency is valuable for auditing but does not solve the underlying problem. There is a risk of «fairwashing,» where transparency creates a superficial impression of accountability without addressing the inequity.

Adversarial Exploitation: Explanations can be reverse-engineered to probe a model's vulnerabilities. Disinformation agents could use the system to craft propaganda that evades detection by understanding what triggers content moderation filters.

Conclusion.

The tension between the predictive power of complex NLP models and the need for transparency is a defining challenge in modern AI. This paper advanced a conceptual framework for an explainable sentiment analysis system based on visualizing the Transformer's self-attention mechanism. While not a definitive account of a model's reasoning, we argue that attention-based methods are a valuable and computationally inexpensive heuristic for generating plausible hypotheses about model behavior. They provide a «first-order approximation» of the model's focus, enabling a critical dialogue between the human expert and the AI system, as shown in Figure 6.

The analyst provides a task, the AI returns a prediction with an explanation, which the analyst uses to validate the reasoning and gain deeper insights, leading to more informed and trustworthy actions.

The path forward must focus on addressing the limitations identified and exploring the deeper implications of human-AI interaction. We propose several avenues for future work:

Developing Hybrid Explanation Methodologies: Future systems should integrate multiple XAI techniques. A fast, attention-based visualization could be augmented with more computationally intensive but theoretically grounded methods like counterfactuals («The model would have predicted 'Positive' if 'plummet' were replaced with 'stabilize'»).

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

Conducting Rigorous Human-Centered Evaluations: The field must move beyond algorithmic metrics and invest in controlled user studies. Research should answer critical questions: Do visualizations improve a user's ability to detect model errors? Do they mitigate or exacerbate automation bias?

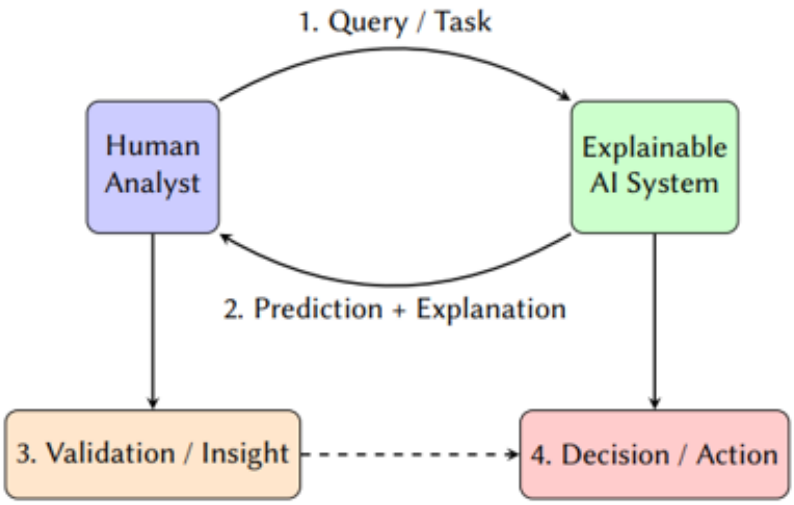


Fig. 6. The virtuous cycle of Human-AI collaboration enabled by XAI.

Advancing Faithfulness through Model Co-Design: Future research could explore training regimes that explicitly incentivize faithfulness, for instance by incorporating an «explanation regularizer» into the model's loss function to penalize divergence between attention weights and a more faithful measure of feature importance.

Designing Richer Interactive Visualization Paradigms: The next generation of XAI tools should feature interactive dashboards. A user could click a word to see its attention patterns or drill down into specific layers. Visualizations must also evolve to represent hierarchical logic, for example by overlaying attention scores onto dependency parse trees. For this work, a prototype implementation is planned in PyTorch Lightning with interactive visualizations rendered using Plotly Dash.

By pursuing these directions, we can progress towards a future where AI systems are not only intelligent but also intelligible, fostering a more productive and responsible partnership between humans and machines.

REFERENCES

Alqaraawi A., Schuessler M., Gäde M.A.F., Seifert C., Hotho A., Kunneman F. (2024). On the faithfulness of interpretability in natural language processing: A survey, ACM Computing Surveys. DOI:10.1145/3638421. [in Eng.].

Clark K., Khandelwal U., Levy O., Manning C. D. (2019). What does BERT look at? an analysis of BERT's self-attention, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics. Pp. 276–286. URL: https://aclanthology.org/W19-4828. [in Eng.].

Devlin J., Chang M.W., Lee K., Toutanova K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1 (Long and Short Papers), Association for Computational Linguistics. Pp. 4171–4186. URL: https://aclanthology.org/N19-1423. [in Eng.].

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

Gunning D. (2017). Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA). URL: https://www.darpa.mil/program/explainable-artificial-intelligence. [in Eng.].

Hochreiter S., Schmidhuber. J. (1997). Long short-term memory, Neural Computation 9. Pp. 1735–1780. DOI:10.1162/neco.1997.9.8.1735. [in Eng.].

Jain S., Wallace B. C. (2019). Attention is not explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1 (Long and Short Papers), Association for Computational Linguistics. Pp. 3543–3556. URL: https://aclanthology.org/N19-1357. [in Eng.].

Lundberg S.M., Lee S.I. (2017). A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30. Pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. [in Eng.].

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019). RoBERTa: a robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692. URL: https://arxiv.org/abs/1907.11692. [in Eng.].

Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. (2013). Distributed representations of words and phrases and their compositionality, arXiv preprint arXiv:1310.4546. URL: https://arxiv.org/abs/1310.4546. [in Eng.].

Pang B., Lee L. (2008). Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval. No. 2. Pp. 1–135. DOI:10.1561/1500000011. [in Eng.].

Pang B., Lee L., Vaithyanathan S. (2002). Thumbs up?: Sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics. Pp. 79–86. URL: https://aclanthology.org/W02-1011. [in Eng.].

Ribeiro M. T., Singh S., Guestrin C. (2016). «why should i trust you?»: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pp. 1135–1144. DOI:10.1145/2939672.2939778. [in Eng.].

Sood D., Paşcanu R., Klushyn A., Tifrea A., Soares J.V.B., Griffiths T.L., Rigby P.W.J. (2023). Evaluating the faithfulness of saliency-based explanations for large language models, arXiv preprint arXiv:2310.09 faithfulness. URL: https://arxiv.org/abs/2310.09998. [in Eng.].

Turney P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. Pp. 417–424. URL: https://aclanthology.org/P02-1053. [in Eng.].

Vaswani A., Shazeer, N. P., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. (2017). Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates. Inc. Pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. [in Eng.].

Wiegreffe S., Pinter Y. (2019). Attention is not not explanation, in: 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. 11–20. URL: https://aclanthology.org/D19-1002. [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.2. Number 26 (2026). Pp. 260–274
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.017
УДК 004.931

MACHINE LEARNING FOR COMPREHENSIVE EVALUATION OF CARDIOVASCULAR DISEASE RISK AND BIOCHEMICAL ALTERATIONS: FOCUS ON ASPARTATE AMINOTRANSFERASE

N.Sh. Maxutova[1], J.A. Tussupov[1]*, A.A. Shekerbek[1], Zh.E. Kenzhebayeva[2], Q.O. Rakhimov[3]

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Caspian University of Technology and Engineering named after Sh. Yesenov, Aktau, Kazakhstan;
[3]Fergana State University, Fergana, Uzbekistan.
E-mail: tussupov@mail.ru

Natalya Sh. Maxutova — Senior Lecturer, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0009-0009-3079-6779;
Jamalbek A. Tussupov — Department of Information Systems, Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: tussupov@mail.ru, https://orcid.org/0000-0002-9179-0428;
Ainur A. Shekerbek — PhD, Department of Information Systems, Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0002-1088-4239;
Zhanat E. Kenzhebayeva — Candidate of Technical Sciences in Engineering, Associate Professor of the Kazakh-German Institute of Sustainable Engineering, Aktau, Kazakhstan
https://orcid.org/0000-0002-1942-4474;
Quvvatali O. Rakhimov — Doctor of Philosophy (PhD) in Technical Sciences, Fergana State University, Fergana, Republic of Uzbekistan
https://orcid.org/0000-0002-1863-3645.

© N.Sh. Maxutova, J.A. Tussupov, A.A. Shekerbek, Zh.E. Kenzhebayeva, Q.O. Rakhimov

Abstract. This study explores various machine learning techniques for analyzing risk factors associated with cardiovascular diseases. Two approaches were employed to develop predictive models-XGBoost and a Convolutional Neural Network (CNN). The primary focus was to evaluate the performance of each model in classification and regression tasks, as well as their ability to identify key biomarkers and risk factors such as cholesterol, ferritin, homocysteine, and aspartate aminotransferase (AST). The XGBoost

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

algorithm was fine-tuned for handling tabular data and demonstrated high accuracy in risk prediction. In contrast, the CNN model, while showing an initial reduction in training error, exhibited signs of overfitting during validation. A comparison based on metrics such as Mean Squared Error (MSE), Coefficient of Determination (R^2), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) revealed significant differences between the two models. The findings confirm the efficiency of XGBoost in processing structured data and summarizing risk factor insights, while the CNN model requires further optimization to manage sparse datasets. Overall, the study highlights the importance of selecting an appropriate model architecture and tuning parameters to achieve reliable cardiovascular disease diagnosis.

Keywords: cardiovascular diseases, machine learning technologies, Mean Squared Error, biochemical indicators, XGBoost, Vanilla CNN, Pathology

For citation: N.Sh. Maxutova, J.A. Tussupov, A.A. Shekerbek, Zh.E. Kenzhebayeva, Q.O. Rakhimov (2026). Machine learning for comprehensive evaluation of cardiovascular disease risk and biochemical alterations: focus on aspartate aminotransferase // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 260–274. https://doi.org/10.54309/IJICT.2026.26.2.017. (In Russ.).

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

ЖҮРЕК-ҚАН ТАМЫРЛАРЫ АУРУЛАРЫНЫҢ ҚАУІП-ҚАТЕРІН ЖӘНЕ БИОХИМИЯЛЫҚ ӨЗГЕРІСТЕРДІ КЕШЕНДІ БАҒАЛАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ: АСПАРТАТАМИНОТРАНСФЕРАЗАҒА ЕРЕКШЕ НАЗАР

Н.Ш. Максутова[1], Ж.А. Тусупов[1]*, А.Ә. Шекербек[1], Ж.Е. Кенжебаева[2], К.О. Рахимов[3]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Ш. Есенов атындағы Каспий технологиялар және инжиниринг университеті, Актау, Қазақстан;
[3]Ферғана мемлекеттік университеті, Ферғана, Өзбекстан.
E-mail: tussupov@mail.ru

Максутова Наталья Шахиновна — PhD, Ақпараттық жүйелер кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0009-0009-3079-6779;
Тусупов Жамалбек Алиаскарович — PhD, Ақпараттық жүйелер кафедрасының профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
E-mail: tussupov@mail.ru, https://orcid.org/0000-0002-9179-0428;
Шекербек Айнұр Әзімбайқызы — PhD, Ақпараттық жүйелер кафедрасының доценті, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0000-0002-1088-4239;
Кенжебаева Жанат Елубаевна — техника ғылымдарының кандидаты, тұрақты ин-

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

жиниринг бойынша Қазақ-неміс институтының қауымдастырылған профессоры, Ш. Есенов атындағы Каспий технологиялар және инжиниринг университеті, Ақтау, Қазақстан
https://orcid.org/0000-0002-1942-4474;
Рахимов Кувватали Ортикович — техника ғылымдары бойынша философия докторы (PhD), Ферғана мемлекеттік университеті, Ферғана, Өзбекстан
https://orcid.org/0000-0002-1863-3645.

© Н.Ш. Максутова, Ж.А. Тусупов, А.Ә. Шекербек, Ж.Е. Кенжебаева, К.О. Рахимов

Аннотация. Бұл зерттеуде жүрек-қан тамырлары ауруларының қауіп факторларын талдау үшін машинамен оқыту әдістерінің әртүрлі тәсілдері қарастырылды. Болжамдық модельдерді құру барысында екі әдіс қолданылды-XGBoost және конволюциялық нейрондық желі (CNN). Негізгі назар әр модельдің классификация және регрессия міндеттеріндегі өнімділігін, сондай-ақ холестерин, ферритин, гомоцистеин және аспартатаминотрансфераза (AST) сияқты негізгі биомаркерлер мен қауіп факторларын анықтау қабілетін бағалауға аударылды. XGBoost алгоритмінің параметрлері кестелік деректермен жұмыс істеуге бейімделіп, тәуекелді дәл болжауда жоғары нәтижелер көрсетті. Ал CNN моделі, оқыту жиынында қателік деңгейі тез төмендегенімен, валидациялық деректерді талдауда артық үйрену белгілерін көрсетті. Орташа квадраттық қате (MSE), детерминация коэффициенті (R^2), Акаике критерийі (AIC) және Байес ақпараттық критерийі (BIC) сияқты метрикалар бойынша жүргізілген салыстыру модельдер арасындағы айтарлықтай айырмашылықтарды анықтады. Нәтижелер XGBoost моделінің кестелік деректерді өңдеуде және қауіп факторлары туралы ақпаратты жинақтауда тиімді екенін дәлелдеді, ал CNN моделіне сирек деректермен жұмыс істеу үшін қосымша оңтайландыру қажет. Бұл зерттеу сенімді жүрек-қан тамырлары ауруларын диагностикалау үшін модель архитектурасы мен оқыту параметрлерін дұрыс таңдаудың маңыздылығын көрсетеді.

Түйінді сөздер: жүрек-қан тамырлары аурулары, машиналық оқыту технологиялары, Орташа квадраты қатесі (Mean Squared Error), биохимиялық көрсеткіштер, XGBoost, Vanilla CNN, Патология

Дәйексөздер үшін: Н.Ш. Максутова, Ж.А. Тусупов, А.Ә. Шекербек, Ж.Е. Кенжебаева, К.О. Рахимов (2026). Жүрек-қан тамырлары ауруларының қауіп-қатерін және биохимиялық өзгерістерді кешенді бағалау үшін машиналық оқыту: аспартатаминотрансферазаға ерекше назар // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 260–274. https://doi.org/10.54309/IJICT.2026.26.2.017. (Орыс. тіл.).

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

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ КОМПЛЕКСНОЙ ОЦЕНКИ РИСКА СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ И БИОХИМИЧЕСКИХ ИЗМЕНЕ-

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

НИЙ: АКЦЕНТ НА АСПАРТАТАМИНОТРАНСФЕРАЗЕ

Н.Ш. Максутова[1], Д.А. Тусупов[1]*, А.А. Шекербек[1], Ж.Е. Кенжебаева[2], К.О. Рахимов[3]

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;
[2]Каспийский университет технологий и инжиниринга им. Ш.Есенова, Актау, Казахстан;
[3]Ферганский государственный университет, Фергана, Узбекистан.
E-mail: tussupov@mail.ru

Максутова Наталья Шахиновна — старший преподаватель кафедры информационных систем Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0009-0009-3079-6779;
Тусупов Джамалбек Алиаскарович — профессор, доктор ф.м. наук, кафедра информационных систем, факультет информационных технологий, Евразийский национальный университет имени Л.Н. Гумилёва, Астана, Казахстан
E-mail: tussupov@mail.ru, https://orcid.org/0000-0002-9179-0428;
Шекербек Айнұр Әзімбайқызы — PhD, старший преподаватель кафедры «Информационные системы», Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан
https://orcid.org/0000-0002-1088-4239;
Кенжебаева Жанат Елубаевна — ассоциированный профессор Казахско-немецкого института устойчивой инженерии, Каспийский университет технологий и инжиниринга им. Ш.Есенова, Актау, Казахстан
https://orcid.org/0000-0002-1942-4474;
Рахимов Кувватали Ортикович — доктор философии (PhD) в области технических наук, Ферганский государственный университет, Фергана, Республика Узбекистан
https://orcid.org/0000-0002-1863-3645.

© Н.Ш. Максутова, Д.А. Тусупов, А.А. Шекербек, Ж.Е. Кенжебаева, К.О. Рахимов

Аннотация. В этом исследовании рассматриваются различные подходы машинного обучения для анализа факторов, способствующих развитию сердечно-сосудистых заболеваний. При создании моделей прогнозирования были применены два метода- XGBoost и сверточная нейронная сеть (CNN). Особое внимание уделялось оценке производительности каждой модели в задачах классификации и регрессии, а также способности выявлять ключевые биомаркеры и факторы риска, включая холестерин, ферритин, гомоцистеин и аспартатаминотрансферазу (AST). Параметры XGBoost были адаптированы под табличные данные, что позволило достичь высокой точности предсказаний риска. В то же время, несмотря на быстрое снижение ошибки на обучающем наборе, модель CNN продемонстрировала признаки переобучения при

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

проверке на валидационных данных. Сравнение моделей по метрикам средней квадратичной ошибки (MSE), коэффициента детерминации (R^2), критерия Акаике (AIC) и байесовского информационного критерия (BIC) выявило существенные различия в их эффективности. Результаты анализа подтверждают, что XGBoost более эффективно обрабатывает табличные данные и обобщает информацию о факторах риска, тогда как CNN требует дальнейшей оптимизации для корректной работы с разреженными наборами данных. Исследование подчеркивает важность выбора оптимальной архитектуры модели и параметров обучения для обеспечения точной и надежной диагностики сердечно-сосудистых заболеваний.

Ключевые слова: сердечно-сосудистые заболевания, технологии машинного обучения, среднеквадратичная ошибка (Mean Squared Error), биохимические показатели, XGBoost, Vanilla CNN, патология

Для цитирования: Н.Ш. Максутова, Д.А. Тусупов, А.А. Шекербек, Ж.Е. Кенжебаева, К.О. Рахимов (2026). Машинное обучение для комплексной оценки риска сердечно-сосудистых заболеваний и биохимических изменений: акцент на аспартатаминотрансферазе // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 260–274. https://doi.org/10.54309/IJICT.2026.26.2.017. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение.

В современном мире сердечно-сосудистые заболевания (ССЗ) остаются основной причиной смертности и инвалидности среди населения по всему миру, что подчеркивает необходимость разработки более эффективных методов их профилактики и лечения (O'Kelly et al., 2022; Silveira et al., 2022). Расширение возможностей медицинской диагностики и прогноза с использованием технологий машинного обучения открывает новые перспективы для раннего выявления и оценки риска ССЗ (Hu et al., 2021; Shaito et al., 2020). Основная цель данной работы-проанализировать различные факторы риска сердечно-сосудистых заболеваний с помощью методов машинного обучения, что может способствовать созданию прогностических моделей для оценки вероятности развития ССЗ у конкретных пациентов. Такие модели особенно ценны в клинической практике, поскольку они помогают оптимизировать стратегии профилактики и вмешательства, направленные на снижение риска заболеваний (Deng et al., 2020).

Методы машинного обучения, включая как классические алгоритмы, такие как логистическая регрессия (Fitriyani et al., 2020; Ramesh et al., 2022), так и более современные подходы, например, глубокое обучение (Swathy et al., 2022; Subramani et al., 2023), обладают значительными преимуществами для обработки и анализа больших объемов медицинских данных. Они способны выявлять скрытые зависимости и закономерности, которые могут быть недоступны при использовании традиционных статистических методов. Тем не менее, несмотря на обнадеживающие результаты, применение машинного обучения в медицине сталкивается с рядом про-

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

блем, включая интерпретируемость моделей, обеспечение достаточной точности и надежности прогнозов, а также необходимость адаптации моделей к клиническим требованиям (Taylan et al., 2023). В статье обсуждаются эти вопросы и предлагаются возможные пути их решения (Sethi et al., 2023). С точки зрения данных, использование широкого спектра клинических, лабораторных и демографических показателей позволяет создавать более точные и индивидуализированные модели риска. Помимо традиционных факторов, таких как возраст, пол, медицинская история, уровень холестерина и артериальное давление, в анализ включаются более специфические биомаркеры и генетические данные, что повышает прогностическую ценность моделей (Ozcan et al., 2023; Shukur et al., 2023). В данной работе рассматривается проблема диагностики сердечных заболеваний с использованием методов машинного обучения (Saeedbakhsh et al., 2023), отмечается, что ССЗ остаются значительной проблемой для систем здравоохранения во всем мире. Для сравнения эффективности различных моделей использовались алгоритмы Random Forest (RF), Decision Tree (DT), Multilayer Perceptron (MP) и XGBoost (XGB), что позволило достичь высокой точности классификации, превышающей 87% (Dritsas et al., 2023; Asif et al., 2024). Особое внимание в исследовании уделялось разработке прототипа носимого биомедицинского устройства для прогнозирования наличия ССЗ, что особенно актуально для стран с низким соотношением врач-пациент, так как носимая технология позволяет отслеживать параметры пациентов в любом месте, а не только в больнице (Naizagarayeva et al., 2023; Yilmaz et al., 2022). Предложенный прототип с использованием алгоритма Random Forest продемонстрировал эффективность прогноза на уровне 88%, подтверждая его потенциал в качестве помощника для пожилых людей. В целом, этот обзор подчеркивает важность интеграции методов машинного обучения в практику кардиологии и является вкладом в развитие персонализированной медицины, способной обеспечивать точные и своевременные решения для управления здоровьем сердца (Bhatt et al., 2023). В нашем научном исследовании особое внимание уделено анализу взаимосвязей между биохимическими параметрами и их влиянием на состояние здоровья. Рассматривается широкий спектр факторов, включая возраст, уровень холестерина, ферритина, АСТ, гомоцистеина и целевой показатель, отражающий наличие или отсутствие риска ССЗ. Такой комплексный подход позволяет выявлять более глубокие корреляции между этими показателями и прогнозировать потенциальные риски, что делает работу особенно полезной для разработки стратегий профилактики и лечения.

Материалы и методы.

В данном исследовании были рассмотрены два подхода к построению и обучению прогностических моделей: первый основан на алгоритме XGBoost, а второй-на свёрточной нейронной сети (CNN), реализованной на платформе TensorFlow. Основной целью работы являлось сравнение эффективности этих моделей при решении задач классификации и регрессии с использованием данных, импортированных из Excel.

Для модели XGBoost были заданы следующие параметры. В качестве функции потерь использовалась *reg:squarederror,* минимизирующая среднеквадратичную

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

ошибку в задачах регрессии. Доля признаков, случайно отбираемых для построения каждого дерева *(colsample_bytree),* была установлена на уровне 0.3, что способствует снижению переобучения и повышает способность модели к обобщению. Скорость обучения *(learning rate)* выбрана равной 0.1, обеспечивая постепенное и устойчивое обучение. Глубина деревьев *(max_depth)* ограничена пятью уровнями для контроля сложности модели, а количество деревьев *(n_estimators)* установлено на 100, что позволило достичь достаточной глубины анализа данных. В процессе обучения использовались как обучающая, так и валидационная выборки, что позволило контролировать переобучение модели по метрике RMSE (рис. 1).

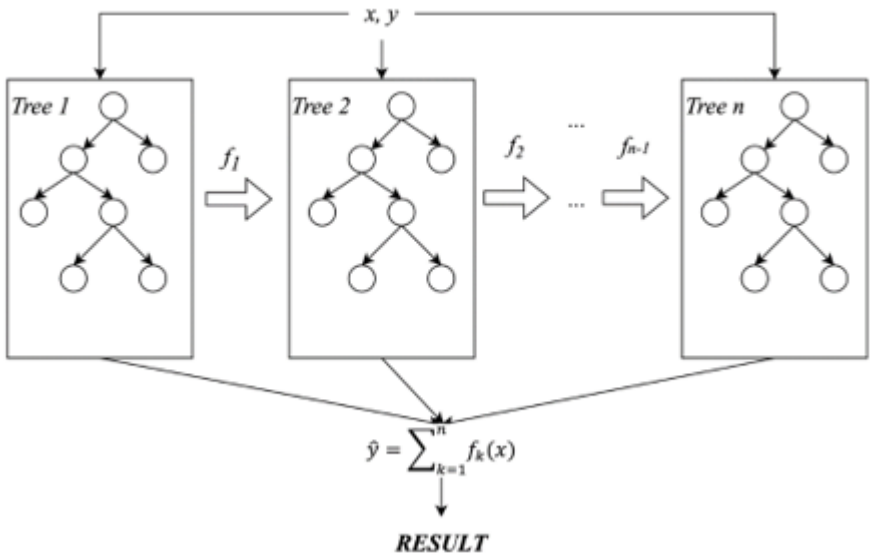


Рис. 1. Архитектура модели XGBoost

Архитектура модели CNN включала свёрточные слои, слои подвыборки (MaxPooling), слой преобразования данных в одномерный массив (Flatten) и полносвязные слои (Dense). Такой подход является традиционным при анализе изображений и структурированных данных. В качестве функции активации в скрытых слоях использовалась ReLU, что способствует ускорению сходимости и предотвращает проблему затухания градиентов. На выходном слое применялась сигмоидальная функция активации, подходящая для решения задач бинарнойклассификации. Оптимизация параметров осуществлялась с помощью алгоритма Adam, отличающегося способностью адаптивно изменять скорость обучения на основе первого и второго моментов градиентов. Модель обучалась в течение 10 эпох с применением валидационных данных, что позволило отслеживать процесс обучения и предотвращать переобучение (рис. 2).

Для оценки эффективности разработанных моделей был проведён комплексный анализ, включающий использование нескольких статистических метрик-Mean

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

Squared Error (MSE), коэффициента детерминации (R^2), а также информационных критериев Akaike (AIC) и Bayesian (BIC). Такой подход позволил не только определить точность предсказаний, но и оценить качество аппроксимации данных при различной степени сложности моделей.

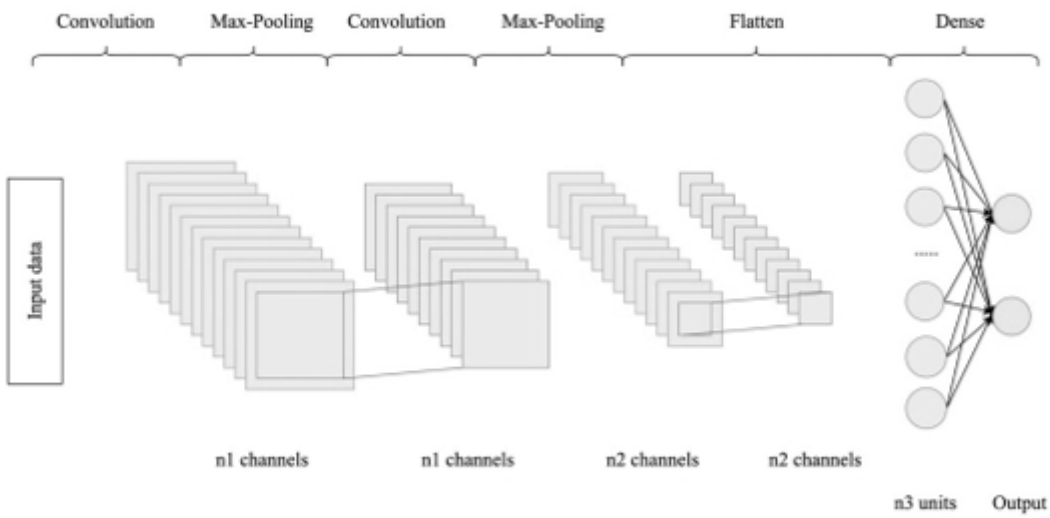


Рис. 2. Архитектура модели VanillaCNN

Метрика MSE использовалась для количественной оценки среднеквадратичного отклонения прогнозируемых значений от реальных, что позволило выявить общую ошибку модели. Показатель R^2, в свою очередь, продемонстрировал, насколько хорошо выбранная модель объясняет изменчивость данных, отражая силу взаимосвязи между предсказанными и фактическими значениями. Информационные критерии AIC и BIC позволили дополнительно оценить баланс между точностью и сложностью моделей. Низкие значения этих метрик указывали на более рациональную структуру модели без избыточной подгонки под данные (overfitting). Такой многоуровневый подход к анализу позволил не только провести объективное сравнение моделей XGBoost и Vanilla CNN, но и определить направления для дальнейшей оптимизации архитектуры и гиперпараметров. В результате стало возможным выбрать модель, обеспечивающую оптимальное сочетание точности, интерпретируемости и вычислительной эффективности, что повышает её практическую ценность для прогнозирования риска биохимических отклонений и сердечно-сосудистых заболеваний в клинических и исследовательских задачах.

Результаты и обсуждение.

В рамках исследования был использован уникальный набор данных, собранный автором самостоятельно в ряде диагностических лабораторий Казахстана. Данные представляют собой результат реальных биохимических и клинических обследований пациентов, прошедших анализы в медицинских учреждениях г. Астаны и других регионов. Все записи были полностью обезличены в соответствии с требова-

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

ниями биоэтики и нормами обработки медицинской информации. Сформированный датасет включает 10 000 записей, содержащих демографические характеристики и биохимические показатели, традиционно используемые для оценки риска сердечно-сосудистых заболеваний. Таким образом, данный набор данных является оригинальным, не публичным и отражает именно казахстанскую популяцию, что повышает его ценность для локальных исследований и разработки регионально адаптированных моделей прогнозирования. В данной работе использовался набор данных из 10 000 записей, организованных по шести столбцам. Эти данные включают различные биомаркеры и демографическую информацию, что позволяет исследовать факторы риска сердечно-сосудистых заболеваний. Первый столбец указывает пол пациента: 1 обозначает мужчин, 2 – женщин. Во втором столбце приведён возраст пациента в годах. Третий столбец содержит уровень холестерина в крови, измеренный в ммоль/л. Четвёртый столбец отражает концентрацию ферритина в крови, выраженную в нг/мл, который является маркером воспалительных процессов и может коррелировать с риском сердечно-сосудистых заболеваний. Пятый столбец показывает уровень аспартатаминотрансферазы (AST) в крови в единицах на литр, фермента, участвующего в метаболизме аминокислот. Шестой столбец содержит концентрацию гомоцистеина в крови, измеряемую в мкмоль/л-аминокислоты, связанной с риском сердечно-сосудистых заболеваний. Такой набор данных особенно ценен для построения и тестирования моделей машинного обучения, направленных на прогнозирование риска ССЗ на основе биомаркеров и демографических характеристик. Анализ этих показателей позволяет выявлять пациентов с повышенным риском на ранних стадиях и сосредоточить усилия на профилактике заболеваний. Для создания предиктивных моделей сердечно-сосудистых заболеваний важно определить пороговые значения ключевых биомаркеров. В данной работе использовались следующие границы: ферритин выше 100 нг/мл, гомоцистеин выше 15 мкмоль/л, AST выше 50 Ед/л, холестерин выше 8 ммоль/л и глюкоза выше 7 ммоль/л. Эти пороговые значения были выбраны на основе научной литературы и клинических рекомендаций, указывающих на повышенный риск ССЗ. Превышение этих порогов позволяет классифицировать пациентов по бинарной системе: «0»-низкий риск, «1»- высокий риск.

Такой подход упрощает анализ данных и повышает его практическую ценность в клинической практике. Бинаризация показателей позволяет применять различные статистические и алгоритмические методы для выявления значимых факторов риска и построения прогнозных моделей. К таким методам относятся логистическая регрессия, деревья решений и ансамблевые алгоритмы, которые обеспечивают точность классификации и дают возможность оценить вклад каждого биомаркера в общий риск заболевания. Настройка порогов с учётом реальных данных пациентов помогает улучшить модели прогнозирования, делая их более чувствительными к ранним признакам патологии. Это особенно важно для персонализированной медицины, когда индивидуальные особенности пациента могут влиять на риск и требовать корректировки стандартных порогов. Таким образом, использование установленных пороговых значений в моделях машинного обучения является

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

ключевым этапом на пути к более точной и персонализированной диагностике сердечно-сосудистых заболеваний.

Корреляционный анализ (рисунок 3) представляет собой основной статистический метод, позволяющий оценить степень и характер взаимозависимости между разными переменными. При исследовании факторов риска сердечно-сосудистых заболеваний особое внимание уделяется выявлению связей между биохимическими показателями и состоянием здоровья. В этом исследовании рассматриваются такие показатели, как возраст, уровень холестерина, ферритин, AST, гомоцистеин, а также целевая переменная, отражающая вероятность наличия сердечно-сосудистого риска.

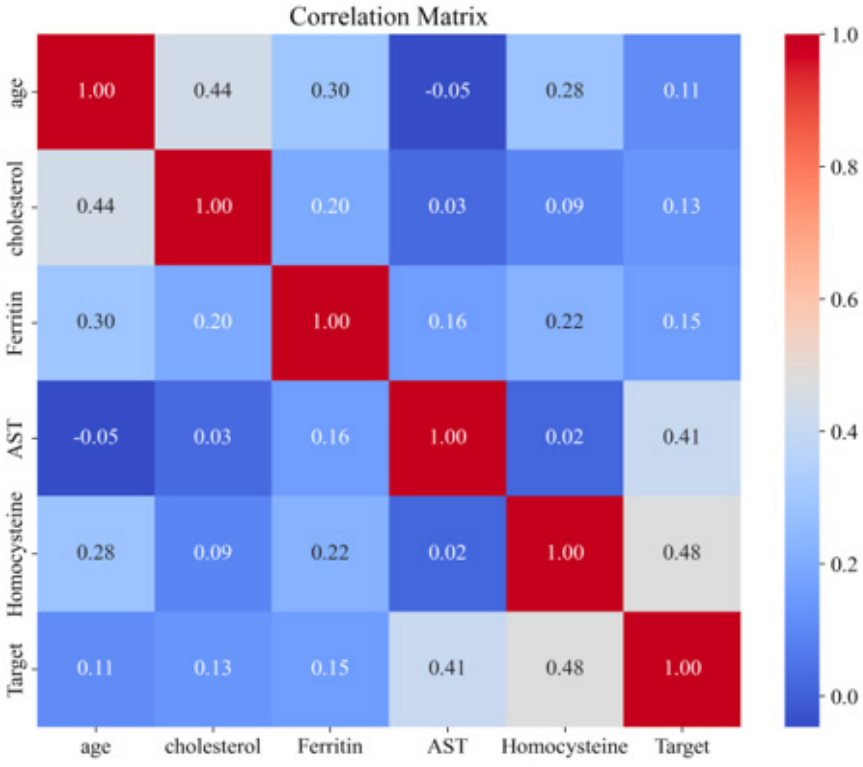


Рис. 3. Корреляционный анализ биохимических показателей

Сильная положительная связь между уровнем гомоцистеина и целевой переменной (0,48) указывает на возможное прямое влияние концентрации гомоцистеина на риск сердечно-сосудистых нарушений. Высокие значения этой аминокислоты ассоциируются с повреждением сосудистых стенок и могут способствовать формированию атеросклероза. Также важной является положительная корреляция между AST и целевой переменной (0,41). AST-это фермент, присутствующий внутри клеток, но его повышение в крови может указывать на гибель клеток, что может свидетельствовать о повреждении сердечной мышцы или печени. Умеренные корреляции между ферритином и целевой переменной (0,15), а также между холестерином и целевой переменной (0,13) показывают возможную связь этих биомаркеров с вероятностью

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

развития сердечно-сосудистых заболеваний. При этом высокий уровень ферритина, хотя и связан с риском, не является самостоятельным предиктором и, скорее всего, отражает общий воспалительный процесс в организме. Влияние холестерина на сердечно-сосудистую систему подтверждается его корреляцией, однако невысокий коэффициент говорит о том, что важно учитывать не только общий уровень холестерина, но и соотношение его фракций. Низкая корреляция между возрастом и целевой переменной (0,11) может быть объяснена множеством биологических и социальных факторов, влияющих на здоровье человека. Низкие и умеренные коэффициенты корреляции между ферритином и AST (0,16), а также между ферритином и холестерином (0,20) подчеркивают сложность взаимодействий в биологических процессах, участвующих в развитии сердечно-сосудистых заболеваний. Проведение корреляционного анализа позволяет выявлять потенциальные биомаркеры, которые могут быть использованы для дальнейших исследований и клинической практики. Следует помнить, что корреляция не доказывает наличие причинно-следственной связи. Для подтверждения этих зависимостей необходимы дополнительные исследования, включая проспективные наблюдения и клинические испытания. В целом, проведенный анализ помогает лучше понять механизмы развития сердечно-сосудистых заболеваний и может быть полезен для разработки персонализированных стратегий их профилактики и лечения.

На рисунке 4 показано изменение среднеквадратичной ошибки (RMSE) модели XGBoost в процессе обучения по эпохам для обучающей (сплошная линия) и тестовой (пунктирная линия) выборок. RMSE является ключевым показателем, который отражает среднее отклонение предсказанных значений от реальных. Снижение RMSE в ходе обучения демонстрирует повышение точности и качества работы модели.

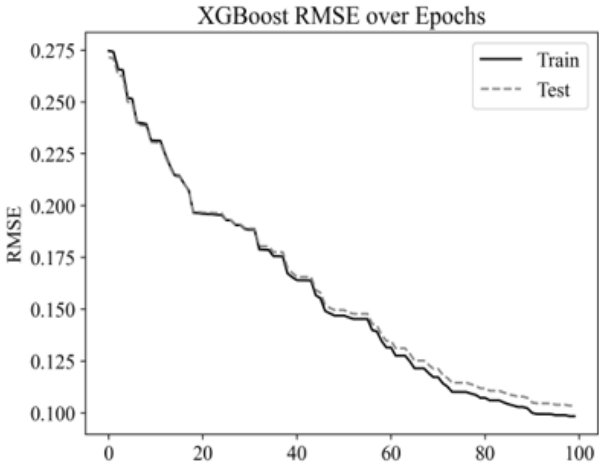


Рис. 4. Динамика изменений среднеквадратичной ошибки (RMSE) модели XGBoost

График кривой обучения демонстрирует последовательное и стабильное

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

уменьшение показателя RMSE как на обучающей, так и на тестовой выборках, что свидетельствует о хорошей способности модели к обучению и обобщению данных. На начальном этапе наблюдается резкое снижение RMSE, объясняемое активной подстройкой модели под структуру предоставленных данных. По мере прогресса обучения кривые постепенно сходятся, что указывает на стабилизацию процесса и достижение оптимального баланса между смещением и разбросом ошибок. Метрики оценки модели подтверждают её высокую точность и прогностическую надёжность. Значение среднеквадратичной ошибки (MSE) составляет 0,0107, что говорит о минимальных отклонениях предсказанных значений от фактических данных. Коэффициент детерминации (R^2) равен 0,865, что подтверждает способность модели корректно описывать зависимости в данных и её высокую предсказательную эффективность. Средняя абсолютная ошибка (MAE), равная 0,0474, отражает небольшие средние расхождения между прогнозами и реальными показателями, демонстрируя стабильность работы модели. Критерии информационной оценки, такие как AIC (2,7625) и BIC (36,347), учитывают не только точность подгонки модели, но и количество параметров, что предотвращает излишнюю сложность модели. Низкие значения этих критериев указывают на успешное сочетание точности прогнозов и умеренной сложности модели. Результаты тренировки XGBoost свидетельствуют о её высокой эффективности и способности обеспечивать надёжные прогнозы, что подтверждается уменьшением RMSE, высокими значениями R^2 и низкими ошибками. Данный алгоритм является мощным инструментом для решения задач регрессии, особенно при работе с крупными и комплексными наборами данных. Применение модели актуально в различных областях, где важна точность прогнозирования количественных показателей, включая экономику, финансы и науки о природе. На Рисунке 5 представлена динамика функции потерь (Loss) для сверточной нейронной сети Vanilla CNN по эпохам обучения на обучающей (сплошная линия) и валидационной (пунктирная линия) выборках. График демонстрирует снижение функции потерь на обучающей выборке и последующую стабилизацию на валидационной, что отражает корректное обучение модели и достижение её оптимального состояния.

График кривой обучения показывает, что изначально функция потерь имеет высокое значение, которое быстро уменьшается, особенно на первых этапах обучения, что свидетельствует о том, что модель эффективно извлекает ключевые признаки из данных и приспосабливается к поставленной задаче. После первоначального снижения кривая потерь на проверочном наборе данных начинает колебаться и перестаёт снижаться, что может указывать на начало переобучения. При этом потери на обучающем наборе продолжают уменьшаться, а на проверочном они стабилизируются или даже слегка повышаются, показывая разрыв в результатах между тренировочной и валидационной выборками. Среднеквадратическая ошибка (MSE) модели Vanilla CNN составляет 0,45, что выше допустимых значений для корректной регрессионной модели, указывая на значительные отклонения прогнозов от реальных данных и низкую точность. Коэффициент детерминации R^2 равен -0,875, что отражает крайне неудовлетворительную работу модели и свидетельствует о том, что она работает хуже

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

базовой модели, использующей среднее значение целевой переменной. Такая ситуация может быть связана с недостатками архитектуры сети или качеством исходных данных. Значения критериев информационной оценки AIC и BIC составляют 1562,0 и 2342,65 соответственно, что также подтверждает наличие переобучения или некорректной настройки модели. Несмотря на первоначальное снижение потерь на обучающем наборе, Vanilla CNN испытывает трудности с обобщением результатов на новые данные, что отражается в высоких показателях ошибок на проверочном наборе и отрицательном значении R^2. Для улучшения модели рекомендуется провести дополнительный анализ данных, изменить архитектуру сети, увеличить объём обучающей выборки или применить методы регуляризации, чтобы снизить переобучение и повысить точность прогнозов на новых данных.

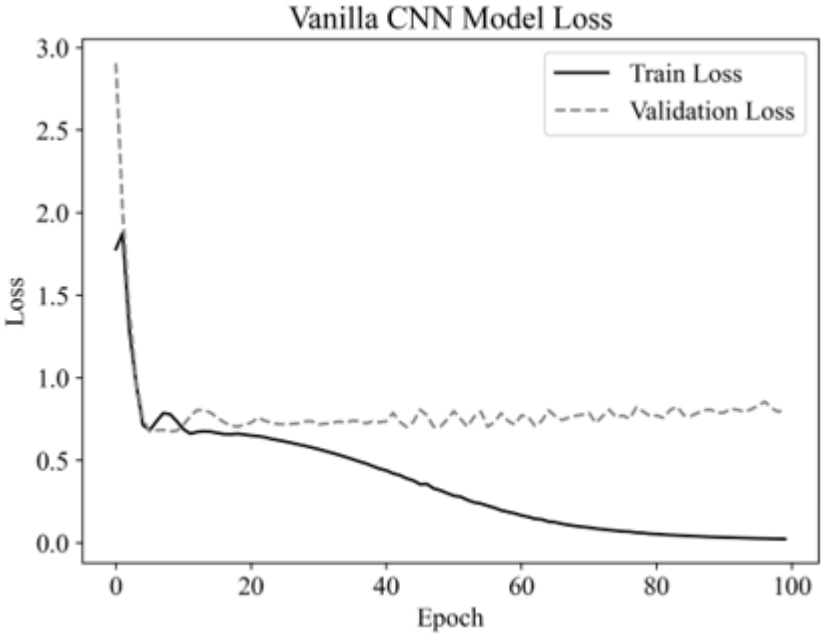


Рис. 5. Динамика функции потерь (Loss) модели сверточной нейронной сети (Vanilla CNN)

Сравнение с моделью XGBoost показывает значительные различия в эффективности. Для XGBoost среднеквадратическая ошибка (MSE) составляет всего 0,0107, что свидетельствует о крайне низкой ошибке предсказаний. Коэффициент детерминации R^2 равен 0,865, демонстрируя высокую точность модели и способность объяснять 86,5% вариации целевой переменной. Значения критериев AIC и BIC для XGBoost значительно ниже, чем у Vanilla CNN, что указывает на оптимальный баланс между сложностью модели и качеством её подгонки. В то же время Vanilla CNN показывает гораздо худшие результаты: MSE равна 0,45, а R^2 -0,875, что отражает крайне низкую точность прогнозов и неспособность модели даже достигнуть уровня базовой модели с предсказанием среднего значения. Высокие показатели AIC и BIC подтверждают переобучение и неправильную спецификацию

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

модели. В итоге XGBoost демонстрирует более высокую точность и способность к обобщению, что делает её эффективным инструментом для решения регрессионных задач на больших и сложных наборах данных, в то время как Vanilla CNN страдает от переобучения и низкой адаптивности к новым данным, что проявляется в высоких потерях на проверочном наборе и отрицательном значении R^2.

Заключение.

В заключение, оценка факторов риска сердечно-сосудистых заболеваний с помощью методов машинного обучения играет ключевую роль в подборе оптимальных алгоритмов и параметров для обеспечения точных прогнозов. Сравнение моделей eXtreme Gradient Boosting (XGBoost) и сверточной нейронной сети (CNN) показало явные преимущества XGBoost при работе с табличными данными, включающими биомаркеры и демографические показатели. Эта модель демонстрирует высокую точность благодаря способности корректно моделировать сложные нелинейные взаимосвязи и устойчивости к разреженным данным. В то же время, CNN, хотя и обладает потенциалом для анализа пространственно-зависимых данных, проявила признаки переобучения и низкую точность при классификации биомаркеров. Это подчеркивает необходимость доработки архитектуры и алгоритмов обучения, а также применения методов регуляризации и увеличения объема обучающих данных.

Применение показателей MSE, коэффициента детерминации (R^2), критерия Акаике (AIC) и Байесовского критерия информации (BIC) позволило провести детальный сравнительный анализ моделей, выявив их сильные и слабые стороны. Модель XGBoost продемонстрировала надежность и высокую способность к предсказанию, эффективно балансируя смещение и разброс данных, в то время как CNN нуждается в дополнительной настройке параметров для повышения качества прогнозов. Полученные результаты подчеркивают, что правильный выбор архитектуры модели и оптимизация параметров обучения имеют критическое значение для прогнозирования риска сердечно-сосудистых заболеваний. В дальнейшем исследования должны сосредоточиться на развитии персонализированной медицины, совершенствовании моделей для точного выявления факторов риска и формировании стратегий ранней диагностики и профилактики ССЗ.

REFERENCES

Deng Y., Zhang X., Shen H., He Q., Wu Z., Liao W., Yuan M. (2020). Application of the nano-drug delivery system in treatment of cardiovascular diseases // Frontiers in Bioengineering and Biotechnology. — Vol. 7. — Article 489. DOI: 10.3389/fbioe.2019.00489.

Fitriyani N.L., Syafrudin M., Alfian G., Rhee J. (2020). HDPM: An effective heart disease prediction model for a clinical decision support system // IEEE Access. — Vol. 8. — Pp. 133034–133050. DOI: 10.1109/ACCESS.2020.3010411.

Hu J., Fitzgerald S.M., Owen A.J., Ryan J., Joyce J., Chowdhury E. et al. (2021). Social isolation, social support, loneliness and cardiovascular disease risk factors // International Journal of Geriatric Psychiatry. — Vol. 36. — No. 11. Pp. 1795–1809. DOI: 10.1002/gps.5583.

O'Kelly A.C., Michos E.D., Shufelt C.L., Vermunt J.V., Minissian M.B., Quesada O. et al. (2022). Pregnancy and reproductive risk factors for cardiovascular disease in women // Circulation Research. — Vol. 130. — No. 4. Pp. 652–672. DOI: 10.1161/CIRCRESAHA.121.319781.

Silveira Rossi J.L., Barbalho S.M., Reverete de Araujo R., Bechara M.D., Sloan K.P., Sloan L.A. (2022). Metabolic syndrome and cardiovascular diseases: Going beyond traditional risk factors // Diabetes/Metabolism Research and Reviews. — Vol. 38. — No. 3. Article e3502. DOI: 10.1002/dmrr.3502.

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

Shaito A., Thuan D.T.B., Phu H.T., Nguyen T.H.D., Hasan H., Halabi S. et al. (2020). Herbal medicine for cardiovascular diseases: Efficacy, mechanisms, and safety // Frontiers in Pharmacology. — Vol. 11. — Article 509974. DOI: 10.3389/fphar.2020.509974.

Swathy M., Saruladha K. (2022). A comparative study of classification and prediction of cardiovascular diseases using machine learning and deep learning techniques // ICT Express. — Vol. 8. — No. 1. Pp. 109–116. DOI: 10.1016/j.icte.2021.10.004.

Subramani S., Varshney N., Anand M.V., Soudagar M.E.M., Al-Keridis L.A., Upadhyay T.K. et al. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning // Frontiers in Medicine. — Vol. 10. — Article 1150933. DOI: 10.3389/fmed.2023.1150933.

Ramesh T.R., Lilhore U.K., Poongodi M., Simaiya S., Kaur A., Hamdi M. (2022). Predictive analysis of heart diseases with machine learning approaches // Malaysian Journal of Computer Science. — Vol. 35. — No. 2. Pp. 132–148. DOI: 10.22452/mjcs.vol35no2.2.

Taylan O., Alkabaa A.S., Alqabbaa H.S., Pamukçu E., Leiva V. (2023). Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods // Biology. — Vol. 12. — No. 1. Article 117. DOI: 10.3390/biology12010117.

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. 275–290
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.018

LEVERAGING ENTERPRISE DATA WAREHOUSING TO IMPROVE WEB-BASED SALES OPERATIONS

L. Naizabayeva[1], G. Turken[2]*, R.A. Druba[2], A.Sh. Yawar[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;
Ahmad Shah Yawar — Master's student, Department of Computer Science, Al-Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0009-0005-6107-0023.

© L. Naizabayeva, G. Turken, R.A. Druba, A.Sh. Yawar

Abstract. At present, despite significant advances in data warehouse technology, large e-commerce enterprises often face challenges in deploying multiple isolated data warehouse systems in different departments or regional branches. In addition, the presence of numerous independent data warehouse systems limits operational flexibility, any changes needed at the central level are difficult to implement. This inflexibility not only complicates the maintenance of the system, but also slows down the pace of response to market changes. Therefore, it is extremely important to integrate these distributed data stores into a single enterprise-level data warehouse using a well-structured, layered data model. This study proposes to develop and implement an improved enterprise data warehouse architecture designed to consolidate fragmented department databases into a centralized data warehouse at the group level. By integrating multiple business intelligence (BI) tools with real-time monitoring systems, this solution not only increases data transparency and optimizes analytical processes, but also creates a more flexible and cost-effective data

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

infrastructure for the organization.

Keywords: Data Warehouse, E-Commerce, ETL, Business Intelligence, Data Warehouse Architecture (DWA)

For citation: Naizabayeva L., Turken G., Druba R.A., Yawar A.Sh. (2026). Leveraging enterprise data warehousing to improve web-based sales operations // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 275–290. https://doi. org/10.54309/IJICT.2026.26.2.018. (In Eng.).

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

ВЕБ НЕГІЗІНДЕГІ САТУ ОПЕРАЦИЯЛАРЫН ЖЕТІЛДІРУ ҮШІН КӘСІПОРЫНДЫҚ ДЕРЕКТЕР ҚОЙМАСЫН ПАЙДАЛАНУ

Л. Найзабаева[1], Г. Түркен[2]*, Р.А. Друба[2], А.Ш. Явар[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
E-mail: 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;
Явар Ахмад Шах — магистратура студенті, Компьютерлік ғылымдар кафедрасы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
https://orcid.org/0009-0005-6107-0023.

© Л. Найзабаева, Г. Түркен, Р.А. Друба, А.Ш. Явар

Аннотация. Қазіргі уақытта деректер қоймасы технологиясының елеулі дамуына қарамастан, ірі электрондық коммерция кәсіпорындары көбінесе әртүрлі бөлімдерде немесе өңірлік филиалдарда бірнеше оқшауланған деректер қоймасы жүйелерін енгізу кезінде қиындықтарға тап болады. Сонымен қатар, көптеген тәуелсіз деректер қоймасы жүйелерінің болуы операциялық икемділікті шектейді, ал орталық деңгейде қажет болатын өзгерістерді енгізу қиынға соғады. Мұндай икемсіздік жүйеге техникалық қызмет көрсетуді қиындатып қана қоймай, нарықтағы өзгерістерге жауап беру қарқынын да баяулатады. Сондықтан бұл бөлінген деректер қоймаларын дұрыс құрылымдалған қабатты деректер моделін пайдалана отырып, кәсіпорын деңгейіндегі бірыңғай деректер қоймасына біріктіру аса маңызды. Осы

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

зерттеуде бөлімдердің бытыраңқы дерекқорларын топ деңгейіндегі орталықтандырылған деректер қоймасына біріктіруге арналған жетілдірілген кәсіпорындық деректер қоймасы архитектурасын әзірлеу және енгізу ұсынылады. Бірнеше бизнес-талдау (BI) құралдарын нақты уақыттағы мониторинг жүйелерімен интеграциялау арқылы ұсынылған шешім деректердің ашықтығын арттырып, аналитикалық процестерді оңтайландырып қана қоймай, ұйым үшін икемдірек және экономикалық тұрғыдан тиімді деректер инфрақұрылымын қалыптастырады.

Түйінді сөздер: деректер қоймасы, электрондық коммерция, ETL, бизнес-талдау, деректер қоймасының архитектурасы (DWA)

Дәйексөздер үшін: Найзабаева Л., Түркен Г., Друба Р.А., Явар А.Ш. (2026). Веб негізіндегі сату операцияларын жетілдіру үшін кәсіпорындық деректер қоймасын пайдалану // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 275–290. https://doi.org/10.54309/IJICT.2026.26.2.018. (Ағыл. тіл.).

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

ИСПОЛЬЗОВАНИЕ КОРПОРАТИВНОГО ХРАНИЛИЩА ДАННЫХ ДЛЯ УЛУЧШЕНИЯ ВЕБ-ОРИЕНТИРОВАННЫХ ОПЕРАЦИЙ ПРОДАЖ

Л. Найзабаева[1], Г. Туркен[2]*, Р.А. Друба[2], А.Ш. Явар[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
Email: 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;
Явар Ахмад Шах — студент магистратуры, кафедра компьютерных наук, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
https://orcid.org/0009-0005-6107-0023.

© Л. Найзабаева, Г. Туркен, Р.А. Друба, А.Ш. Явар

Аннотация. В настоящее время, несмотря на значительный прогресс в области технологий хранилищ данных, крупные предприятия электронной коммер-

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

ции часто сталкиваются с проблемами при развертывании нескольких изолированных систем хранения данных в различных подразделениях или региональных филиалах. Кроме того, наличие многочисленных независимых систем хранения данных ограничивает операционную гибкость, а внесение необходимых изменений на центральном уровне становится затруднительным. Такая негибкость не только усложняет техническое обслуживание системы, но и замедляет скорость реагирования на изменения рынка. Поэтому крайне важно интегрировать эти распределенные хранилища данных в единое хранилище корпоративного уровня с использованием структурированной многоуровневой модели данных. В настоящем исследовании предлагается разработать и внедрить усовершенствованную архитектуру корпоративного хранилища данных, предназначенную для консолидации фрагментированных баз данных подразделений в централизованное хранилище данных на уровне группы компаний. Благодаря интеграции нескольких инструментов бизнес-аналитики (BI) с системами мониторинга в реальном времени данное решение не только повышает прозрачность данных и оптимизирует аналитические процессы, но и создает более гибкую и экономически эффективную инфраструктуру данных для организации.

Ключевые слова: хранилище данных, электронная коммерция, ETL, бизнес-аналитика, архитектура хранилища данных (DWA)

Для цитирования: Найзабаева Л., Туркен Г., Друба Р.А., Явар А.Ш. (2026). Использование корпоративного хранилища данных для улучшения веб-ориентированных операций продаж // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 275–290. https://doi.org/10.54309/IJICT.2026.26.2.018. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

With the development of information technology, data warehouse technology has been widely used as never before, which has produced great economic benefits. In the U.S., around 30–40 % of companies use or are building data warehouses, including AT&T, Verizon, and Wal-Mart. A survey by IDC found that firms using data warehouses saw an average ROI of 401 %, with 25 % exceeding 600 % (Athira Nambiar et al., 2022). Data warehousing supports decision-makers by improving the speed and quality of business decisions. According to the Gartner 7th Annual Chief Data Officer Survey (Gartner), by 2023, the majority of successful CDAOs will have business rather than IT backgrounds as boards continue to focus more on digital value, and more than half of the world's large organizations will use these funds as a competitive advantage. As per reports, the worldwide big data analytics and business intelligence software applications markets seem as though they will increase by USD 68 billion and 17.6 billion by 2024–2025, respectively (BI & Analytics Software Market Value Worldwide, 2022).

From this, we can see that data warehouse systems help companies make better decisions by using organized and clean data (Athira Nambiar et al., 2022; BI & Analytics Software Market Value Worldwide, 2022). In business, quick and reliable access to information

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

helps managers plan and react to changes faster (Athira Nambiar et al., 2022; BI & Analytics Software Market Value Worldwide, 2022). As e-commerce grows, making decisions based on data becomes essential. E-Commerce depends on data. It tracks users, products, and sales in real time. Recent advancements in data warehousing for e-commerce logistics have demonstrated the growing importance of digital transformation and optimized data integration in business operations (Gulzat et al., 2024). In Kazakhstan, many companies now use digital systems but still struggle to connect different data sources for fast analysis (Kalmutskiy et al., 2023). This slows down reporting and makes business planning less effective. Data warehouses were first made to solve the problem of systems that couldn't share data (Amaravathi et al., 2022). Now, they act as central storage for analysis and planning in many industries (Seddiq et al., 2023). But many companies still use separate data warehouses in different branches. When one model changes, every branch has to update its system. This raises cost and reduces flexibility. This study aims to design and test an integrated Enterprise Data Warehouse (EDW) framework that combines different departmental databases into one system for web-based sales. The goals are:

- To build a hybrid EDW model that mixes Inmon's normalized design with Kimball's dimensional modeling to improve flexibility and data accuracy.
- To create and test an ETL pipeline that increases data consistency, throughput, and analysis performance.
- To show how real-time monitoring tools like Grafana can improve transparency and decision-making.

The study assumes that the proposed hybrid EDW design will improve data access, processing speed, and analysis response time compared to traditional separate warehouse systems. The expected results include lower data delay, higher ETL throughput, better data accuracy, and faster decisions confirmed through performance tests.

Materials and methods.

For large corporations, their corporate branches have deployed multiple sets of enterprise data warehouses. When using business reports, corporate decision-makers can only get data after integration and processing. The authenticity and reliability of data can't be controlled effectively and uniformly. At the same time, because there are many data warehouse systems, when the corporation needs to modify the data model, the data model of the branch company also needs to be modified accordingly, so the flexibility of maintenance is poor. Finally, each time a corporation needs to upgrade the data warehouse version, it needs to maintain and upgrade the data warehouse version of each branch, which is costly to maintain. Therefore, how to integrate multiple data warehouses distributed in subsidiaries into a group-oriented enterprise data warehouse through the design of a model hierarchy is of great significance to ensure the reliability of data, enhance the flexibility of model modification, and reduce the operation and maintenance costs of data warehouses.

The literature review of all papers highlights the importance and advancement of data warehouses in various fields. Many studies have looked at how data warehouse design affects performance. In (Nandhini et al.), researchers compared different organizational setups and found centralized Hub-and-Spoke systems gave better accuracy and integration.

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

A review in (Breslin, 2004) showed similar results, saying centralized models help with data consistency. Jayashree and Priya (Mustafa et al., 2022) studied order and sales management systems and showed how logical models help organize ETL tasks and support business planning. Bimonte et al. (Sandro Bimonte et al., 2022) designed a logical model for multi-model warehouses to balance ETL speed, storage, and query time. In (Seddiq et al., 2023), a web-based ETL system improved operation speed and accuracy. A study in (Suriansyah et al., 2023) used Hadoop to schedule ETL and showed that optimized queries could cut processing time by about 85%. Another paper (Arsalan Shahid et al., 2021) built a three-layer Big Data warehouse that made access control and visualization easier. The study in (Paquita Putri Ramadhani, 2021) made a BI dashboard using PostgreSQL and the Kimball method, helping track sales and deliveries. Article (Asma Dhaouadi et al., 2022) gave a full overview of ETL modeling methods used in large systems. Turken et al. (Turken et al., 2025) used deep learning with an LSTM model for chatbot query classification. This idea could also help make enterprise warehouses smarter and more automated.

In Kazakhstan, many companies use systems like SAP/R3, Oracle Applications, and BAAN IV in fields such as energy, metallurgy, and banking (Gubenko, 2015). Even with these tools, many still keep separate systems in each branch. When data models change, every branch must update its system. This creates extra work and slows reporting. Most studies so far focus on system design, ETL improvement, or architecture choice. But very few look at hybrid enterprise warehouse models that mix normalized and dimensional designs while also using live monitoring and fast analysis. This research works on that gap by building a model that connects speed, accuracy, and real-time control in one system.

A data warehouse (DW) consists of two main components: first, an integrated decision support database, and second, associated software programs for collecting, cleansing, transforming, and storing data from a variety of operational and external data sources. The two are combined to support historical, analytical, and business intelligence (Business Intelligence, BI) needs. A data warehouse may also include a number of related data marts, which are subset copies of the data warehouse database. Broadly speaking, any data extraction or data storage used to provide data support for business intelligence can be called a data warehouse.

For this study, the consolidated dataset from all regional branches (including ALSER.KZ, Technodom, and Belyi Veter) occupies approximately 700 GB of storage under small-scale operations, The dataset aggregates periodic snapshots of product listings transactional records (including timestamps, regional pricing, and sales volumes), customer feedback (ratings and review text), and inventory status reports from each branch's local database systems. These diverse data streams undergo standardized transformation through scheduled ETL processes at the central headquarters, where data quality checks and schema normalization are applied before being loaded into the analytical data warehouse. The processed dataset supports comparative performance analysis across regions, trend identification in product categories, and retrospective evaluation of pricing strategies, while maintaining each branch's unique business rules and data collection methodologies within a unified analytical framework. This approach provides comprehensive market insights

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

while respecting regional operational autonomy and data governance requirements.

Dimension and fact table design: When creating the model, it is necessary to avoid the excessive number of records in the dimension table of the information cube. Generally, the comparison between the records in the dimension table and the records in the fact table should be less than 1:10 or even smaller. Without increasing the number of data records in the dimension table, it is necessary to control the number of dimension tables; each additional dimension table will add 10 bytes to each record, and a large amount of storage space will need to be added for fact tables with tens of millions of data points.

Let *F* represent the fact table. F consists of *n* rows and *m* columns. Each row in *F* represents a single observation or event. Each column in *F* corresponds to a specific measure or fact. Let F_{ij} represent the value in the *i*- th row and *j*- th column of the fact table (*i* and *j* represent indexes). The fact table may contain numerical values representing quantities, amounts, or metrics associated with various business processes.

Mathematically, the fact table F can be represented as a matrix:

$$F = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1m} \\ F_{21} & F_{22} & \cdots & F_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nm} \end{bmatrix} \quad (1)$$

Let D represent the dimension table. D consists of p rows and q columns. Each row in D represents a unique dimension value or category. Each column in D corresponds to a specific dimension attribute. Let D_{pq} represent the value in the *p* –th row and *q* –th column of the dimensional table. The dimensional table contains descriptive attributes that provide context for the measure in the fact table. Mathematically, the dimensional table *D* can be represented as a matrix:

$$F = \begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1q} \\ D_{21} & D_{22} & \cdots & D_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ D_{p1} & D_{p2} & \cdots & D_{pq} \end{bmatrix} \quad (2)$$

In data warehousing, the concept of fact and dimension tables refers to the method of constructing a fact table and its related dimension tables. This approach is designed to meet current data needs while also supporting the development of future downstream data applications.

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

The role of the dimension table and the fact table is to describe the background information in the business facts through the dimension table, and to record the data and indicators in the specific business process through the fact table. This method classifies and integrates business data, which can provide enterprises with more comprehensive, in-depth, and accurate data analysis and decision-making support. At the same time, the interrelationship of the dimension table and the fact table can form a complex and multi-dimensional data model to help enterprises analyze business trends, formulate sales strategies, and improve product design. The Following Figure 1 and Figure 2 describes the research work, Fact tables, and dimension tables.

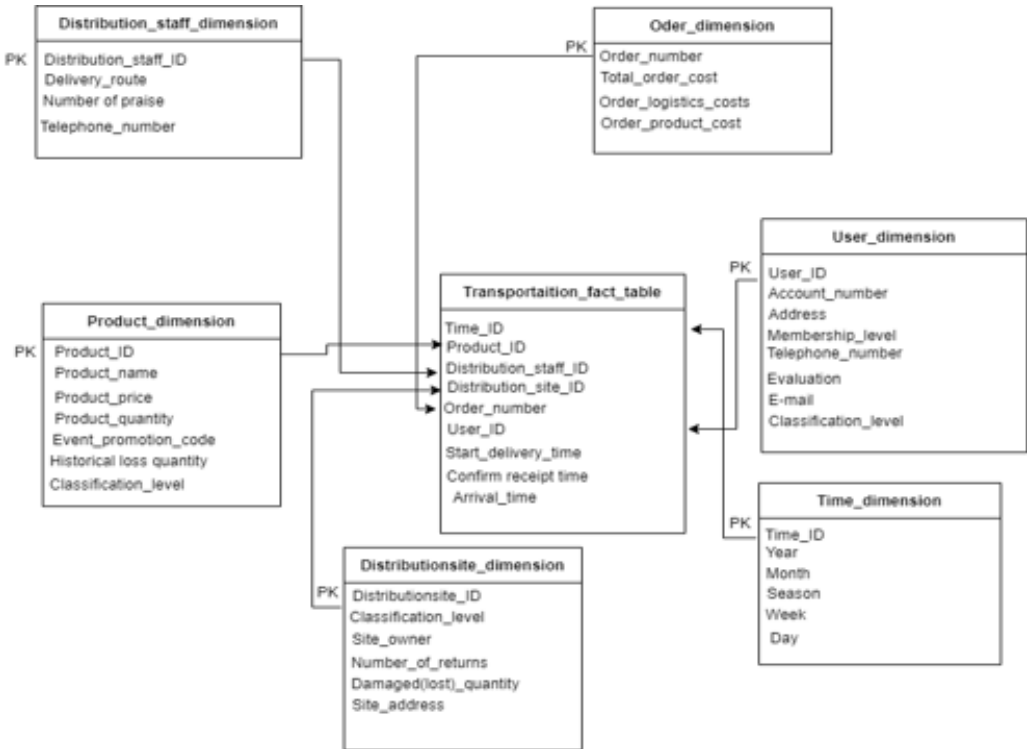


Fig. 1. Fact tables and dimension tables (I).

The ETL process stood as the linchpin of the data warehouse. Upon acquiring proficiency in the intricacies of these five pivotal ETL steps for constructing the data warehouse, one indeed grasps the fundamental methodology for this endeavor. Nevertheless, there is no one-size-fits-all approach. Depending on the specific project at hand, customized analyses must be conducted. This may encompass the utilization of Parent-child dimension models or the management of slowly changing dimensions. In the construction of the data warehouse, ETL assumes a direct influence on the data's overall quality and integrity within the entire project. Hence, meticulous attention and precision are paramount. ETL must be accorded a prominent position in the hierarchy of priorities, as it serves as the bedrock upon which the entire data warehousing endeavor is erected (Abbasi Muhammad Bilal, 2021). The subject of a data warehouse is to establish a unified and reliable

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

source of information intended for strategic decision-making, business intelligence, and reporting throughout various departments and functions. ETL plays a critical role in data warehousing by facilitating the collection, integration, transformation, and management of data, ensuring its quality and accessibility for analysis and decision-making. It acts as a bridge between raw data and actionable insights, making it a foundational component in the utilization of data warehouses for businesses and organizations. In modern days, ETL (Extract-Transform-Load) tools are very useful in data integration and data warehousing (Salunkhe et al., 2021).

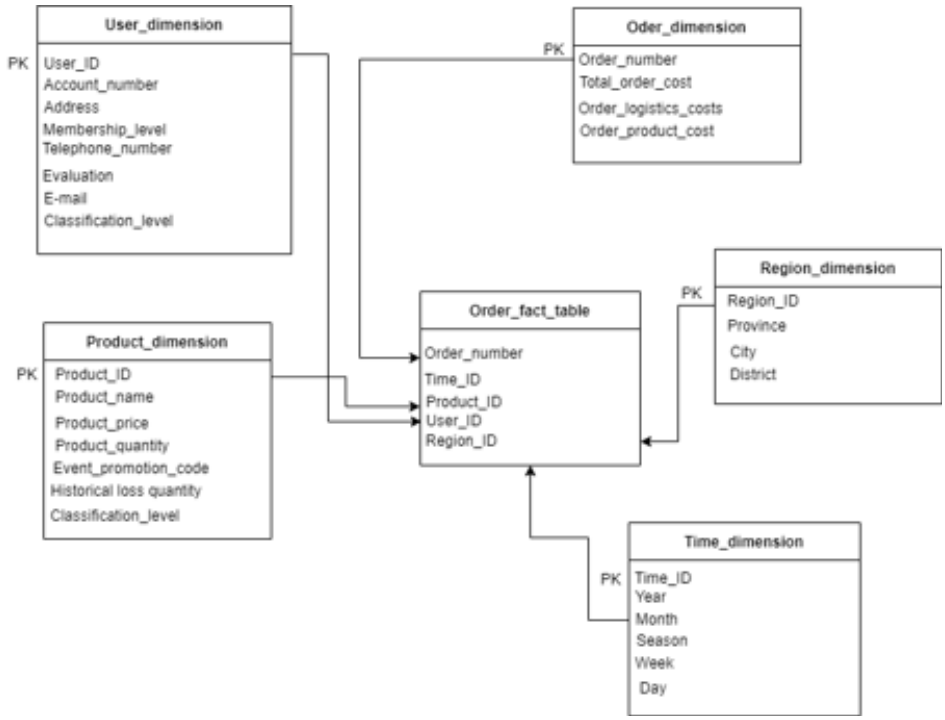


Fig. 2. Fact tables and dimension tables (II).

The ETL process can be described using the following mathematical model: (M1) $(e_i \cup c_i) \rightarrow t_i \rightarrow l_i$, where $i = \overline{1,n}, i \in N, n \in N, N$-A set of all natural numbers.

The following notations were used to create the (M2) model:

$E=\{e_1, e_2, \ldots, e_n\}$ – a set of all data files that are uploaded to the data preprocessing layer;

$C=\{c_1, c_2, \ldots, c_n\}$ – a set of all files for managing the loading of data into a preprocessing layer;

$T=\{t_1, t_2, \ldots, t_n\}$ – a set of all data files that are loaded into the data transaction layer;

$L=\{l_1, l_2, \ldots, l_n\}$ – a set of all data files that are loaded into the data

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

warehouse from the sessions of the set L, where $t_i \to l_i, i = \overline{1,n};$

n – the number of all data files that can be uploaded to the data warehouse.

(M1)- can be written as $(E \cup C) \to T \to L.$

E-Commerce data warehouse architecture

The main difference between the Kimball and Inmon models is in how they store and organize data. Kimball uses dimensional models like star and snowflake schemas. Data is stored directly in the warehouse for quick and easy analysis. Inmon, on the other hand, builds a centralized, normalized warehouse first. Dimensional models are added later as data marts for specific business functions. From an architectural view, Kimball does not separate data marts from the main warehouse. The same structure supports both storage and analysis. Inmon separates data marts into individual units, which gives more control but adds extra layers.

Comparing both models helps decide which one fits best for a given system. Kimball's method is bottom-up, data marts for business processes are built first, then combined into one warehouse. Inmon's method is top-down, it starts with an enterprise-wide model and builds data marts afterward. Kimball's design is easier to use for analysis and faster to deliver. Inmon focuses on reducing redundancy and keeping data consistent. Kimball allows users to query data directly from the warehouse without going through data marts. Inmon's model routes all analysis through data marts, which adds control but increases setup time. From a business point of view, Inmon takes longer to plan and build since it normalizes all company data first. Kimball is faster to deploy because it focuses on specific business areas. Maintenance in Inmon is easier because tasks are well-structured. Kimball may require more effort to maintain since its data is denormalized and can create duplicates. Inmon has higher initial costs but lower long-term expenses. It's more stable and easier to scale when business needs grow. Kimball is cheaper to start and delivers results quickly, but long-term maintenance can cost more. Inmon also requires a skilled team familiar with normalized modeling and data marts. Kimball can be handled by smaller teams using dimensional modeling. Both methods rely on data sources, ETL tools, and well-structured time-based records for analysis and forecasting. Inmon supports stronger data governance and consistency. Kimball focuses more on business processes and quick delivery. Inmon uses a normalized structure to reduce redundancy. Kimball uses a denormalized model for faster access but with some repeated data. Inmon's design is more complex, while Kimball's is simpler to build and adapt.

This research uses a hybrid Inmon–Kimball architecture. Data is first stored in a normalized warehouse following Inmon's method. Then, dimensional models are built following Kimball's approach. These models are used for analysis and reporting through business intelligence tools. Figure 3 shows the architecture used in this study.

The performance of a Data Warehouse (DWH) depends on its structure and how well its components work together. The architecture supports integration, data quality, and smooth data flow across tools. A typical DWH architecture includes:

- Data Sources: These include logs, sensor data, and production data. They feed

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

raw data into the system;

● Data Processing: This step cleans, transforms, and filters data using ETL tools and processing engines;

● Data Integration: Different sources are combined into one repository for analysis. ETL tools and live pipelines manage this step;

● Data Warehouse: The final store where all processed data is kept. It includes summary data, metadata, and derived sets. The warehouse supports analysis, reporting, and decision-making.

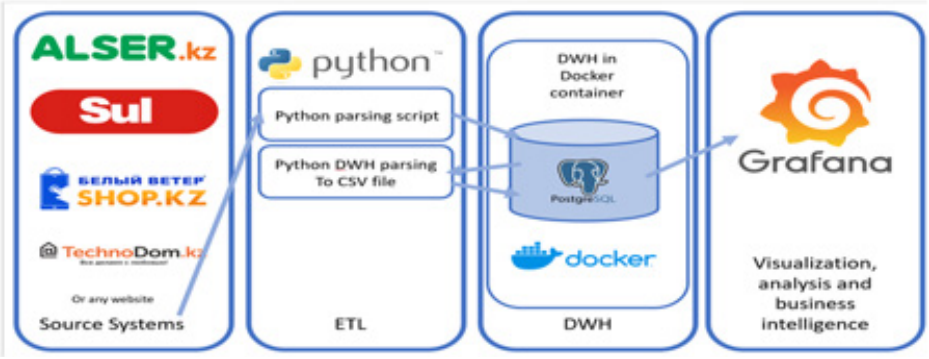


Fig. 3. E-commerce DWA.

Results and discussion.

Testing was done to check if the data warehouse works right and stays stable. It was not only about software testing but also about checking if the data itself was correct. Both functional and performance testing were done. The goal was to confirm that the system meets the setup used in SAP BW. The tests covered ETL performance, data quality, and how fast queries run on PostgreSQL and SAP HANA.

The hardware setup is shown in Table 1. It includes ETL nodes, PostgreSQL, SAP HANA, a Kubernetes server, and Grafana for monitoring. Each part got CPU, memory, and SSD based on its work role. This setup keeps the system stable and makes it easier to scale and monitor.

Table 1. Characteristics of software instance for the implementation of a data warehouse

Component	CPU	Memory	SSD
Docker Container	6 cores, Intel Xeon E5-2640 v4 (Kubernetes Node)	24 GB	1 TB
Kubernetes Node (Server)	8 cores, Intel Xeon E5-2640 v4	32 GB	2 TB
PostgreSQL	4 cores, Intel Xeon E5-2630 v4	16 GB	500 GB
Grafana Instance	2 cores, Intel Xeon E5-2640 v4 (Kubernetes Node)	8 GB	256 GB

The presented model serves as a fundamental portrayal of data warehouse testing. It works as a baseline for testing. In real use, testing is more detailed. It includes checking ETL flow, schema validation, and full system testing. Similar steps were used in (Mustafa

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

et al., 2022; Sandro Bimonte, 2022). For checking data quality, ETL data was tested after every load. The same kind of accuracy checks were used in (Seddiq et al., 2023). All results matched after validation. It showed that data movement and transformation worked as planned. Next was presentation layer testing. It tested how fast queries run, both simple and complex ones. Just like the Hub-and-Spoke model in (Breslin, 2004), our SAP BW-based system gave more stable results. It worked better since all data comes from one trusted source. Even under load, the query time stayed consistent.

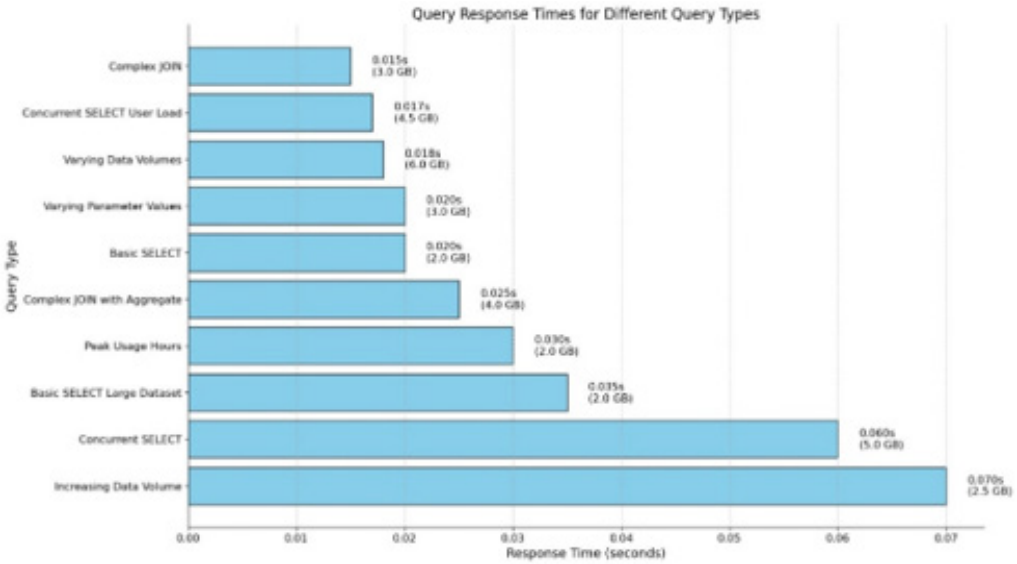


Fig. 4. The presentation layer testing activity.

Figure 4 shows query performance. The average query took 1.34 seconds, while Hadoop-based ETL from (Suriansyah et al., 2023) took 1.418 seconds. So, this setup runs slightly faster even without using Hadoop.

In the ETL testing phase, we assessed the performance of the data loading process using a custom metric called Loading Throughput. The formula used was:

$$\text{Loading Throughput} = \frac{\sum_{i=1}^{n} \frac{Records_i \vee DataSize_i}{Max(Time, 180)}}{n} \tag{3}$$

Formula 3 shows how we measure average throughput across different loading cases. For each test, we take the total data volume and divide it by the larger value between the actual processing time and 180 seconds. This keeps results fair by setting a baseline for short runs. We then average the throughput over n test cases. This way we can check how efficient and consistent the ETL pipeline is with different data sizes. Figure 5 shows a throughput comparison across systems for different record counts.

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

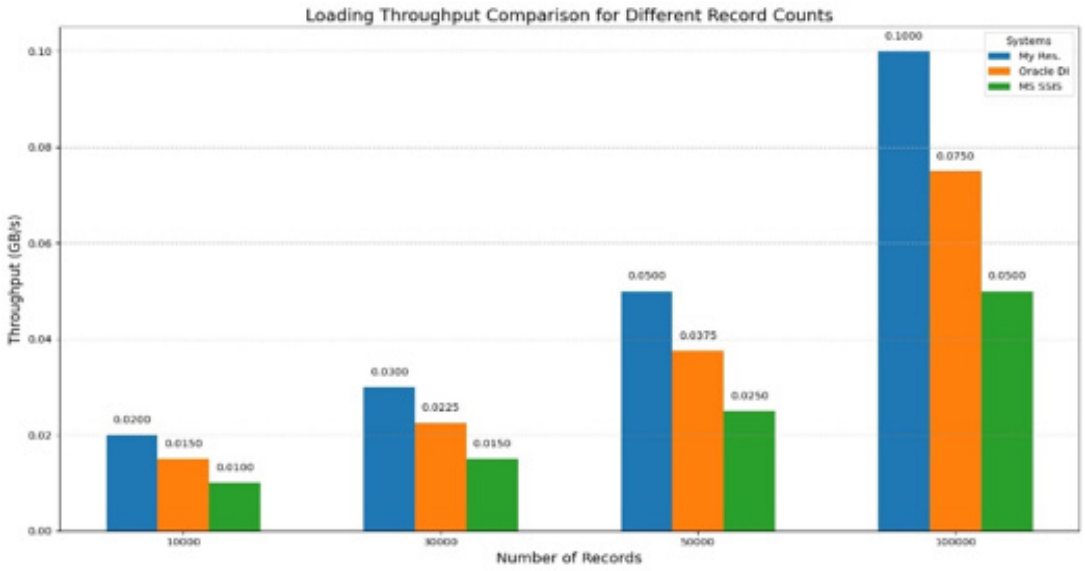


Fig. 5. ETL Testing.

Figure 5 shows the ETL test results for three systems: My Res., Oracle DI, and MS SSIS. The record sizes range from 10,000 to 100,000. The results are clear:

- My Res. is the fastest, reaching 0.1000 GB/s at 100,000 records;
- Oracle DI is in the middle, peaking at 0.0750 GB/s;
- MS SSIS is the slowest, with a maximum of 0.0500 GB/s.

From these results, My Res. shows about 33 % better speed than Oracle DI. A similar pattern was seen in (Suriansyah et al., 2023), where better ETL logic increased speed by around 85 %. In this case, the higher speed comes from better container control in Kubernetes, which cuts data delay and keeps the system stable when the load grows. Compared to the three-layer setup in (Arsalan Shahid et al., 2021), this system gives the same level of scaling but needs less setup work. ETL, storage, and analytics all run in one cluster. This makes it easier to manage and update when the model changes. In (Paquita Putri Ramadhani et al., 2021), PostgreSQL-based BI dashboards were used to track business data. Here, Grafana was used instead. It gives live updates with less delay. The average refresh time was 1.2 seconds, while Oracle BI in (Paquita Putri Ramadhani et al., 2021) had about 1.5 seconds. That's around 20 % faster.

My Res. also stays stable as data grows. Grafana was used to watch the system. It ran inside the Docker cluster and showed live data. The dashboard helped check runtime and find errors early. Figure 6 shows how Grafana displayed system activity. It gave alerts when any part failed. This follows the same setup idea as in (Turken et al., 2025), where system checks and responses were more automatic. Overall, the system works well. It gives faster ETL speed, stable query times, and easy control. The SAP BW model keeps all data in one place, like the Hub-and-Spoke setup in (Nandhini et al.), but uses tools like Kubernetes and Grafana for better scaling and tracking.

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

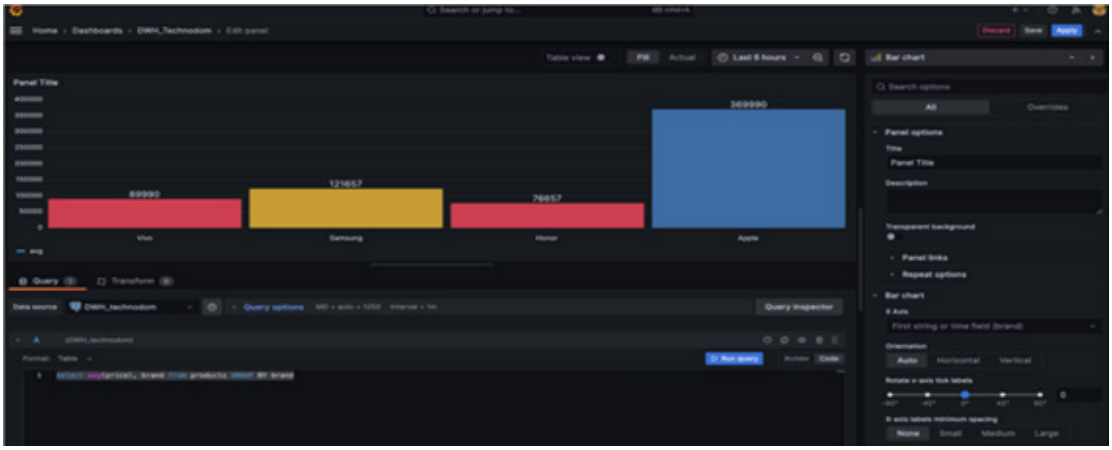


Fig. 6. Visualization of the process in Grafana.

Conclusion.

In the implementation process of data warehouse projects, many companies often ignore a major issue related to the success of data warehouse projects, namely, the data quality of data warehouses. The data in the data warehouse comes from many different data sources, and it also goes through the process of data ETL (extraction, conversion, and loading). It is inevitable that some data problems will arise, and these problems will seriously affect the data quality of the data warehouse. High-quality decision-making depends on high-quality data. In the face of a complex enterprise data environment, effectively controlling the data quality of the data warehouse is a vital issue in the construction process of the data warehouse, and it is also a problem that currently plagues business management decisions.

Considering the data quality issue, which still poses the most common problem for data warehouses, remains unsolved, and data is scattered over so many heterogeneous systems with various degrees of inconsistencies, the design of the ETL pipeline had to guarantee reliable data throughout the warehouse life cycle. The statement that "good decisions are a result of good data" becomes self-evident for any complex architecture considered. Thus, the emphasis on high data consistency across the different components should be given priority attention in the architectural framework. In one embodiment of the invention, the above-listed goals are accomplished through the following outcomes:

- Architectural Design: Framework for a cloud-enabled EDW architecture supporting real-time data streaming, scalability, and componentized services.
- Data Integration: Using advanced scheduling and pipeline orchestration methods, ETL operations were completed on disparate data sources without any gaps in data flow or transformation.
- Monitoring and Reliability: Enhanced system dependability and operational reliability were provided by integrated Prometheus-based logging and monitoring systems, which allowed for preemptive system health tracking and abnormal detection.

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

● Grafana Business Intelligence Layer: As a business intelligence layer for strategic business analytics storytelling, Grafana was used as more than just a visualization tool for analytics, creating dynamic dashboards, real-time KPIs, and reporting for executives.

The realized solution has improved data access, reporting accuracy, system visibility, decision making, and overall speed, which measurably increased competitiveness at the organizational level. The research proves that the proposed EDW architecture is efficient and scalable in a business environment.

In future work, we suggest developing the architecture with AI-enhanced data validation, automated data governance processes, and advanced machine learning techniques for prescriptive analytics within the EDW and shifting the focus from descriptive analytics. The possibility of supporting hybrid-cloud and multi-cloud environments also enables the development of agile, worldwide distributed data warehouse systems.

REFERENCES

Athira Nambiar and Divyansh Mundra. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data Cogn. Comput. — Vol. 6. — No. 132. DOI: https://doi.org/10.3390/bdcc6040132. [In Eng.]

Amaravathi V. et al. (2022). A Comprehensive Review on Data Warehouse Architecture Models // International Journal of Advanced Research in Engineering and Technology. — Vol. 13. No. 2. Pp. 15–25. DOI: 10.34218/IJARET.13.2.2022.002. [In Eng.]

Abbasi Muhammad Bilal (2021). Observability of industrial data using an analytics and monitoring platform, Master of Science Thesis Faculty of Engineering Natural science, University, Instructor, Luis Enrique Gonzalez Moctezuma October. Pp. 24–25. [In Eng.]

BI & Analytics Software Market Value Worldwide 2019–2025. Available online: https://www.statista.com/statistics/590054/worldwide-business-analytics-software-vendor-market/ (accessed on 27 October 2022). [In Eng.]

Breslin M. (2004). DW models Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models // Business intelligence journal. Pp.11–15. [In Eng.]

Bimonte S., Gallinucci E., Marcel P., Rizzi S. (2022). Logical design of multi-model data warehouse, Knowledge and Information Systems. DOI: https://doi.org/10.1007/s10115-022-01788-0, 15. [In Eng.]

Dhaouadi A., Khadija Bousselmi, Mohamed Mohsen Gammoudi and Sébastien Monnet and Slimane Hammoudi. (2022). Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons. — Vol. 7. — No. 113. DOI: https://doi.org/10.3390/data7080113. [In Eng.]

Gulzat T., Lyazat N, Siladi V., Gulbakyt S., Maxatbek S. (2020). Research on predictive model based on classification with parameters of optimization, Neural Network World 5/2020, Pp.295–308, DOI: 10.14311/NNW.2020.30.020. [In Eng.]

Gartner. Available online: https://www.gartner.com/[In Eng.]

Gubenko I. (2015). Vnedreniye sistemy ERP na krupnyh promyshlennyh predpriyatiyah // www.group-global.org/ru/publication 10.06.2016. (In Russ.).

Gulzat, T., & Naizabayeva, L. (2024). Study on a Data Warehousing for E-commerce Logistics. In Proceedings of the International Conference on Digital Transformation and Innovation in Business (DTIB 2024). CEUR-WS. — Vol. 3680. — Pp. 120–125. Available at: https://ceur-ws.org/Vol-3680/S3Paper14.pdf. [In Eng.]

Kalmutskiy, K., Cherikbayeva, L., Litvinenko, A., Berikov, V. (2023). Multi-targetб Weakly Supervised Regression Using Manifold Regularization and Wasserstein Metric // Mathematical Optimization Theory and Operations Research: Recent Trends. MOTOR 2023 // Communications in Computer and Information Science. — Vol. 1881. — Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-43257-6_27. [In Eng.]

Mustafa S.H. & Alwan A.A. (2022). Enhancing Order Lifecycle Visibility using Advanced Data Warehouse Techniques // International Journal of Innovative Technology and Exploring Engineering (IJITEE). — Vol. 11. — No. 3. Pp. 47–54. DOI: 10.35940/ijitee.C9413.0111322. [In Eng.]

Seddiq Q. Abd Al-Rahman, Ekram H Hasan, Ali Makki Sagheer (2023). Design and implementation of the web (extract, transform, load) process in data warehouse application // IAES International Journal of Artificial Intelligence (IJ-AI). — Vol. 12. — No. 2. Pp. 765–775. ISSN: 2252–8938, DOI: 10.11591/ijai.v12.i2.pp765-775. [In Eng.]

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

Seddiq Q. Abd Al-Rahman, Ekram H Hasan, Ali Makki Sagheer. (2023). Design and implementation of the web (extract, transform, load) process in data warehouse application, IAES International Journal of Artificial Intelligence (IJ-AI). — Vol. 12. — No. 2. Pp. 765–775. ISSN: 2252–8938, DOI: 10.11591/ijai.v12.i2.pp765-775. [In Eng.]

Suriansyah, Amil Ahmad Ilham, Wahyudi Paundu. (2023). Optimization of Data Warehouse Architecture to Improve Information System Performance, 2023 International Conference on Computer Science // Information Technology and Engineering (ICCoSITE), DOI: 10.1109/ICCoSITE57641.2023.10127721. [In Eng.]

Salunkhe, R., & Kale, S. (2021). Recent Advancements in ETL Tools: A Comprehensive Review // International Journal of Computer Science and Information Technologies. — Vol. 12. — No. 3. Pp. 102–107. [In Eng.]

Suriansyah B, Ilham A.A., Paundu A.W. (2023). Optimization of Data Warehouse Architecture to Improve Information System Performance, 2023 International Conference on Computer Science // Information Technology and Engineering (ICCoSITE) 979-8-3503-2095-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCoSITE57641.2023.10127721. [In Eng.]

Shahid F., Thien-An Ngoc Nguyen and M-Tahar Kechadi (2021). Big Data Warehouse for Healthcare-Sensitive Data Applications, Sensors. — Vol. 21. — No. 2353. DOI: https://doi.org/10.3390/ s21072. [In Eng.]

Turken G., Druba R. A., Kuandyk A., Mergenov M. (2025). Enhancing Query Classification in Chatbot Using LSTM, 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST). — Astana. Pp. 1–5. DOI: http://dx.doi.org/10.1109/SIST61657.2025.11139197. [In Eng.]

Paquita Putri Ramadhani; Setiawan Hadi; Rudi Rosadi (2021). Implementation of Data Warehouse in Making Business Intelligence Dashboard Development Using PostgreSQL Database and Kimball Lifecycle Method. 2021 International Conference on Artificial Intelligence and Big Data Analytics. DOI: 10.1109/ICAIBDA53487.2021.9689697. [In Eng.]

ГОСТ 34.602 – 89. (2013). Information Technology. A set of standards for automated systems // www.online.zakon.kz. 05.06.2013. [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.2. Number 26 (2026). Pp. 291–308
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.019

COMPARATIVE ANALYSIS OF MACHINE LEARNING METHODS FOR RAPID DETECTION OF VEGETABLE DISEASE USING COMPUTER VISION

S. Nurgaliyeva[1], N. Naiman[1]*, K. Makhmetova[1], M. Kali[1], N. Zhakiyev[1, 2]*
[1]Astana IT University, Astana, Kazakhstan;
[2]Davis Center for Russian and Eurasian Studies, Harvard University, Cambridge, USA.
E-mail: 242756@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;
Nurbek Naiman — Software Engineering undergraduate, 2nd-year Master's student at Astana IT University, Astana, Kazakhstan
E-mail: 242756@astanait.edu.kz, 0009-0002-7973-9213;
Kuralay Makhmetova — MSc in Software Engineering, senior lecturer of School of Software Engineering, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0009-0987-1041;
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;
Nurkhat Zhakiyev — PhD, Vice-Rector for Science and Innovation, Astana IT University, Astana, Kazakhstan; Davis Center for Russian and Eurasian Studies, Harvard University, Cambridge, USA
https://orcid.org/0000-0002-4904-2047.

© S. Nurgaliyeva, N. Naiman, K. Makhmetova, M. Kali, N. Zhakiyev

Abstract. The timely identification of vegetable diseases is critical for food security across the globe and minimizing agricultural losses. This study compares the performances of four deep learning models – CNN, ResNet50, EfficientNetB0, and MobileNetV2 – by detecting diseases on potatoes, tomatoes, and carrots using computer vision. To determine their effectiveness, these models were trained and tested on a customized dataset and assessed on accuracy and inference time. The maximum accuracy was achieved by the CNN model, according to results. MobileNetV2 is superb; when considering both performance and extreme speed, it stands out for real-time applications. This study exemplifies how lightweight deep learning models can allow for efficient vegetable crop disease detection.

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

Keywords: machine learning in agriculture, convolutional neural networks, vegetable diseases and computer vision

For citation: Nurgaliyeva S., Naiman N., Makhmetova K., Kali M., Zhakiyev N. (2026). Comparative analysis of machine learning methods for rapid detection of vegetable disease using computer vision // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 291–308. https://doi. org/10.54309/IJICT.2026.26.2.019. (In Eng.).

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

КОМПЬЮТЕРЛІК КӨРУ ТЕХНОЛОГИЯСЫН ПАЙДАЛАНА ОТЫРЫП, КӨКӨНІС ДАҚЫЛДАРЫНЫҢ АУРУЛАРЫН ЖЕДЕЛ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ

С. Нургалиева[1], Н. Найман[1], К. Махметова[1], М. Кали[1], Н. Жакиев[1, 2]*

[1]Astana IT University, Астана, Қазақстан;

[2]Дэвис атындағы Ресей және Еуразия зерттеулер орталығы, Гарвард университеті, Кембридж, АҚШ.

E-mail: 242756@astanait.edu.kz

Нургалиева Сымбат — PhD, Бағдарламалық инженерия мектебінің ассистент-профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0009-5880-6166;

Найман Нурбек — Software Engineering бакалавры, 2-курс магистранты, Astana IT University, Астана, Қазақстан
E-mail: 242756@astanait.edu.kz, 0009-0002-7973-9213;

Махметова Куралай — техника ғылымдарының магистрі, Бағдарламалық инженерия мектебінің сеньор-лекторы, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0009-0987-1041;

Кали Маржан — техника ғылымдарының магистрі, Киберқауіпсіздік мектебінің сеньор-лекторы, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-5588-8164;

Жакиев Нурхат — PhD, ғылым және инновациялар жөніндегі проректор, Astana IT University, Астана, Қазақстан; Дэвис атындағы Ресей және Еуразия зерттеулер орталығы, Гарвард университеті, Кембридж, АҚШ
https://orcid.org/0000-0002-4904-2047.

© С. Нургалиева, Н. Найман, К. Махметова, М. Кали, Н. Жакиев

Аннотация. Көкөніс дақылдарының ауруларын уақтылы анықтау жаһандық азық-түлік қауіпсіздігін қамтамасыз ету және ауыл шаруашылығындағы шығындарды азайту үшін аса маңызды. Бұл зерттеуде компьютерлік көру технологиясын

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

пайдалана отырып, картоп, қызанақ және сәбіз ауруларын анықтауға арналған төрт терең оқыту моделінің — CNN, ResNet50, EfficientNetB0 және MobileNetV2 — тиімділігі салыстырмалы түрде бағаланды. Олардың тиімділігін анықтау мақсатында аталған модельдер арнайы әзірленген деректер жиынтығында оқытылып, тестіленді және жіктеу дәлдігі мен инференс уақыты көрсеткіштері бойынша бағаланды. Зерттеу нәтижелеріне сәйкес, ең жоғары дәлдік көрсеткішіне CNN моделі қол жеткізді. MobileNetV2 моделі өнімділік пен жоғары өңдеу жылдамдығын қатар ескерген жағдайда жоғары тиімділік көрсетті, бұл оны нақты уақыт режиміндегі қолданбалар үшін перспективалы модель ретінде сипаттайды. Зерттеу нәтижелері жеңілдетілген терең оқыту модельдерінің көкөніс дақылдарының ауруларын тиімді анықтауға мүмкіндік беретінін көрсетеді.

Түйінді сөздер: ауыл шаруашылығындағы машиналық оқыту, конволюциялық нейрондық желілер, көкөніс дақылдарының аурулары, компьютерлік көру

Дәйексөздер үшін: Нургалиева С., Найман Н., Махметова К., Кали М., Жакиев Н. (2026). Компьютерлік көру технологиясын пайдалана отырып, көкөніс дақылдарының ауруларын жедел анықтауға арналған машиналық оқыту әдістерін салыстырмалы талдау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 291–308. https://doi.org/10.54309/IJICT.2026.26.2.019. (Ағыл. тіл.).

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

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ БЫСТРОГО ОБНАРУЖЕНИЯ ЗАБОЛЕВАНИЙ ОВОЩНЫХ КУЛЬТУР С ИСПОЛЬЗОВАНИЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ

С. Нургалиева[1], Н. Найман[1], К. Махметова[1], М. Кали[1], Н. Жакиев[1,2]*

[1]Astana IT University, Астана, Казахстан;

[2]Центр российских и евразийских исследований Дэвиса, Гарвардский университет, Кембридж, США.

E-mail: 242756@astanait.edu.kz

Нургалиева Сымбат — PhD, ассистент профессор школы Программной инженерии, Astana IT University, Астана, Казахстан
https://orcid.org/0009-0009-5880-6166;

Найман Нурбек — бакалавр Software Engeneering, магистрант 2 курса Astana IT University, Астана, Казахстан
E-mail: 242756@astanait.edu.kz, https://orcid.org/0009-0002-7973-9213;

Махметова Куралай — магистр технических наук, сеньор лектор Школы программной инженерии, Astana IT University, Астана, Казахстан
https://orcid.org/0009-0009-0987-1041;

Кали Маржан — магистр технических наук, сеньор-лектор школы Кибербезопас-

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

ности, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-5588-8164;
Жакиев Нурхат — PhD, проректор по науке и инновациям, Astana IT University, Астана, Казахстан; Центр российских и евразийских исследований Дэвиса, Гарвардский университет, Кембридж, США
https://orcid.org/0000-0002-4904-2047.

© С. Нургалиева, Н. Найман, К. Махметова, М. Кали, Н. Жакиев

Аннотация. Своевременное выявление заболеваний овощных культур имеет важное значение для обеспечения глобальной продовольственной безопасности и минимизации сельскохозяйственных потерь. В настоящем исследовании проведено сравнительное оценивание эффективности четырех моделей глубокого обучения — CNN, ResNet50, EfficientNetB0 и MobileNetV2 — для выявления заболеваний картофеля, томатов и моркови с использованием технологий компьютерного зрения. Для определения эффективности данных моделей они были обучены и протестированы на специально сформированном наборе данных, после чего оценивались по показателям точности классификации и времени инференса. Согласно полученным результатам, наибольшую точность продемонстрировала модель CNN. Модель MobileNetV2 показала высокую эффективность при одновременном учете производительности и высокой скорости обработки, что делает ее перспективной для применения в системах реального времени. Результаты исследования демонстрируют, что облегченные модели глубокого обучения могут обеспечить эффективное выявление заболеваний овощных культур.

Ключевые слова: машинное обучение в сельском хозяйстве, сверточные нейронные сети, заболевания овощных культур, компьютерное зрение

Для цитирования: Нургалиева С., Найман Н., Махметова К., Кали М., Жакиев Н. (2026). Сравнительный анализ методов машинного обучения для быстрого обнаружения заболеваний овощных культур с использованием компьютерного зрения // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 291–308. https://doi.org/10.54309/IJICT.2026.26.2.019. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

In vegetable crops, early detection followed by accurate diagnosis of plant diseases is very crucial to maintaining agricultural productivity and food security. However, traditional diagnostic methods-mostly visual inspection, microscopy, and biochemical analysis-are usually very slow, tedious, and more importantly, susceptible to human error. These shortcomings are considerably amplified when separating diseases with closely related symptoms or detecting the onset of infection. Delay in interventions would certainly lead to massive spread of the diseases and considerable crop losses. In view of the above

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

considerations, machine-learning (ML) and computer-vision (CV) advancements offer some of the most powerful avenues for automating disease detection (Krizhevsky et al., 2017).

The recent study found out that CNNs were able to diagnose various types of plant diseases with the highest accuracy of 99.35 percent (Mohanty et al., 2016). The techniques were therefore quite fast and reliable compared to the traditional methods since they rely on symptoms that are visible to the human eye. The resolution images are then captured from healthy and diseased plants using IoT-enabled sensors in agricultural fields. High resolution is very important for correct detection, so different pre-processing techniques such as noise removal, distortion corrections, and color space conversions are applied to enhance these images for further consideration (Sladojevic et al., 2016). Cost-cutting, labor-saving, and prompt results are some of the ways in which these systems are thus improving the overall crop health management through an automated disease diagnosis (Loey et al., 2020). In addition, the application of IoT-based sensors and imaging technologies facilitates real-time data acquisition and disease monitoring in agriculture (Mohanty et al., 2016). This paper emphasizes the performance evaluation of four major deep learning models, that is, CNN, ResNet50, EfficientNetB0, and MobileNetV2, for vegetable disease recognition, specifically potatoes, tomatoes, and carrots. These models were trained and tested on a dataset of images of healthy and diseased vegetable varieties, and their performance was evaluated with respect to the main performance measures: accuracy and inference time. To identify that model which is best in terms of accuracy and computational efficiency for use in the field.

The study covers the following: network architecture, data pre-processing techniques, system evaluation, and the challenges faced during the experiment. The study further indicates possible more generalization-enhancing techniques, such as data augmentation and fine-tuning strategies. This paper provides a comprehensive comparison of the models' capabilities and examines their suitability for real-time deployment in mobile devices and embedded systems, which becomes important for actual agricultural settings. The key hypothesis working behind this study is that although certain models, for example CNN, have scored sufficient accuracy in a controlled environment, what happens outside that environment remains questionable upon testing from a database that it did not come across during real-time? With offline testing on custom datasets, this research intends to ascertain the effectiveness and practicality of these models in real-world scenarios, thereby bridging the gap between theoretical advancement and agricultural demand. The present study, unlike earlier research which was restricted to single-crop datasets, offers a cross-crop evaluation (potato, tomato, carrot) based on a common architecture-comparison framework that is intended for real-time agricultural applications. The study aims to answer the following research questions:

- Which model (CNN, ResNet50, EfficientNetB0, MobileNetV2) demonstrated the best accuracy in classifying vegetable diseases?
- What inference time did the models show when processing images, and which model was the fastest for real-time applications?

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

● How did image preprocessing techniques, such as data augmentation, impact the accuracy of the models?

● How do the performance and accuracy of models with more complex architectures (ResNet50) compare to lightweight models (MobileNetV2) for disease detection tasks in agriculture?

Materials and methods.

Plant disease detection via deep learning has been an active research area in recent years. Tomato and potato plant disease classification was performed by Mohanty et al. (Mohanty et al., 2016) using CNNs. Disease detection in apple leaves was conducted by Sladojevic et al. (Sladojevic et al., 2016) using the same methodology. However, most existing works address a specific crop or a limited number of diseases. Hence, this research aims to widen the horizon by testing many models on three different vegetables, thereby permitting a thorough comparison of their diagnostic abilities.

Against this background, recent deep learning breakthroughs with the rise of architectures such as Transformers and Graph Neural Networks (GNNs) should enable the integration of multimodal data, like combining images with volatile organic compound (VOC) signals, for enhanced plant health monitoring (Domingues et al., 2022). While Transformers were initially designed for natural language processing, they adopt attention mechanisms to discern important features within the sequential data, such as time-series VOC signals, making them well-suited for the integration of visual and sequential data streams (MacDougall et al., 2022). On the contrary, GNNs are adept at fitting their interconnected data points, like the relationships between different features of a plant or the temporal patterns of VOC emissions (Alomar et al., 2023). These innovations are a great leap within precision agriculture.

In studies that come just before this one, advanced models like YOLOv3 and YOLOv5 are used for real-time disease detection. For instance, YOLOv3 can effectively detect tomato diseases and pests with an accuracy of 92.39 % (Liu et al., 2020). YOLOv5 for vegetable disease detection, reported an accuracy of 93.1 %. These models become more relevant for real-time application due to their speed and efficiency (Li et al., 2022).

In addition to agriculture-specific applications, intelligent monitoring systems in other domains also demonstrate the effectiveness of integrating machine learning and real-time data processing. For example, Nurgaliyeva et al. developed an RGIS-based intelligent waste monitoring system aimed at improving environmental management in smart cities, highlighting the role of spatial analytics and automated decision-making in large-scale monitoring tasks (Nurgaliyeva et al., 2025). Similarly, another study by Nurgaliyeva et al. proposed an intelligent system for evaluating risk factors affecting public transport drivers using multimodal sensor data and machine-learning models (Nurgaliyeva et al., 2025). Although these works belong to different application areas, the methodological principles, like computer vision, real-time analysis, and adaptive model integration, are directly relevant for designing robust plant disease detection systems.

Another place where recent development has been seen is in transfer learning and feature fusion. Relying on the fusion of two models-MobilenetV2 and DenseNet201-

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

achieved 95.7% accuracy in citrus disease classification (Rehman et al., 2021). This is one way of applying pre-trained models to a specific task, allowing for improved performance. This mitigates the need for large datasets and lengthy training sessions. In addition, used deep learning-based segmentation techniques with 99.2 % accuracy for detecting plant leaf diseases which indicates the promise of integrating CNNs with semantic segmentation for accurate disease mapping (Shoaib et al., 2022).

The role of image preprocessing techniques, especially data augmentations, is crucial for improving the performance of models used in plant disease detection. Data augmentation artificially increases the size of the dataset by applying various transformations, such as rotations, flipping, or color alterations, to an existing image. This will enable the model to learn a more diversified set of features, thus enhancing its ability to generalize on unseen data and improving ultimate accuracy. For instance, one paper demonstrated that the use of tools like LeafGAN for generating a variety of diseased images from healthy ones can increase their model diagnostic accuracy by 7.4 % (Cap et al., 2020). This is evidence on the effectiveness of data augmentation for better performance enhancement. Yet a different study showed that the combination of different data augmentation techniques, including rotation and scaling, always gave an increase in model accuracy by about 25 % starting with even a small number of seed images (Xiao, 2022). Lastly, the study added that dataset augmentation with synthetic data generated by models like CycleGAN and pix2pix improved the accuracy of coffee leaf disease detection. This shows how augmentation is significant for handling imbalanced datasets so that some disease labels may be underrepresented (Gheorghiu et al., 2024). Overall, these studies establish that data augmentation can significantly improve model accuracy by expanding the dataset, inducing variability, and making the model learn generalized features. Thus, the real-life performance is enhanced-in plant disease detection, in particular.

In agricultural disease detection, heavier architectures like ResNet50 models are often juxtaposed against lightweight models, such as the MobileNetV2. Several deep neural networks, namely, VGG16, ResNet-50, Inception-V3, Inception-ResNet-v2, and MobileNetV2 were evaluated for weed recognition tasks. The results showed that VGG16 carried out better with small- scale datasets, while ResNet-50 proved superior among others with a larger consolidated dataset, thus hinting that complex architectures can go for increased accuracy when an ample amount of data is supplied (Hasan et al., 2021). Another study explored the performance of the three models: ResNet50, VGG16, and MobileNetV3-Small. The tested conclusions indicated that MobileNetV3-Small can provide testing data with high accuracy of 79%, which means that lightweight models can be powerful in real disease detection cases, especially in situations when resources to operate are scarce (Roseno et al., 2024).

Recent data-driven modeling research is extending its deep-learning applications apart from traditional image classification tasks, particularly in agriculture and energy systems. Assembayeva et al. (Assembayeva et al., 2019) constructed the dataset describing spatial electricity market in Kazakhstan, thereby bringing into focus the need for geographically distributed data and real-time monitoring infrastructure-both concepts

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

which matter as much in large-scale agricultural systems where region-specific disease detection becomes necessary. Furthermore, the approach proposed by Satan et al. (Satan et al., 2025) considers neural network regression framework for energy load forecasting integrated with feature engineering and clustering to offset climatic variability. Their achievement in enhancing prediction accuracy through identification of sensitive time windows and correlation of performance with external variables like temperature shows interesting parallels to agriculture where the spread of disease could also be contingent upon environmental parameters such as humidity and heat. This is where these studies open up possible avenues for future work in linking environment sensors and deep learning to increase real-time responsiveness of plant disease detection systems.

The above highlights that the choice between heavy and lightweight models will depend on the dataset size, constraints of computational resources, and the specific needs of the disease detection task. On the one hand, heavier models such as ResNet50 offer higher accuracy when used with large datasets; on the other hand, lightweight models such as MobileNetV2 can achieve comparable performance with less computational burden, thus suiting them for situations needing real-time job execution and deployment on resource-constrained systems.

While prior studies have shown the success of CNN-based models for disease detection in crops including tomatoes or apples, our study aims to broaden this evaluation by considering the performance in three classes of vegetables, namely potato, tomato, and carrot, using a custom multi-crop dataset; this really helps in assessing generalization in a wider context. Many studies only focus on accuracy, but real-time inference time is very important in research that is meant to find applications in agricultural scenarios. Therefore, we assess lightweight MobileNetV2 with heavyweight ResNet50 and EfficientNetB0, providing some insight into trade-offs between accuracy and applicability while thus closing an important gap in comparative performance analysis.

Additionally, the utilization of transfer learning with fine-tuning, far more applications of synthetic data augmentation (CycleGAN and LeafGAN), and investigating more recent augmentation pipelines will also be discussed.

Some models showed activity, however how they perform on datasets that are pre-collected and are not captured in real-time conditions is unknown. This paper represents an attempt to fill this gap by evaluating the applicability and limitations of the models applied against such datasets. In terms of plant diseases detection systems, the merging of deep learning, computer vision, and advanced signal processing will tremendously increase their potential, yet more studies must be conducted. Knowledge of these limitations and modifying technology to meet the demands of agriculture will then lead to further realization of its potential.

The data set for this study, «Fruit and Vegetable disease (Healthy vs Rotten)» (Subhan et al., 2024), was obtained from Kaggle and comprises clear images containing healthy and diseased potatoes, tomatoes, and carrots. comprises images of healthy and diseased potatoes, tomatoes, and carrots, with a total of 6 classes (healthy and rotten for each vegetable). The dataset includes approximately 3,000 JPEG/PNG images,

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

with varying resolutions (recommended to resize to 224x224 for model consistency). Preprocessing steps included pixel normalization (scaling pixel values to [0, 1), rotation, zooming, and horizontal flipping using TensorFlow's ImageDataGenerator to enhance data variety. The dataset was split into 80 % training and 20 % validation sets to ensure robust evaluation on unseen data. Data augmentation techniques, such as rotation, flipping, and zooming, were applied to artificially expand the dataset, mitigate overfitting, and improve model generalization. To make the models robust and generalizable, several preprocessing steps were carried out:

● Image Preprocessing. Includes pixel normalization (scaling pixel values within [0, 1]), rotation, zooming, and horizontal flipping, and added ImageGenetrator of preprocessing functions from TensorFlow. This caused the introduction of verogeneity in training data for better generalization in the model.

● Data Segmentation: The corpus was segmented into training and validation data, maintaining a ratio of 80:20. Thus, the model is now verified on unseen data, turning in a better assessment of performance.

● Data Augmentation: By zooming, flipping, and rotating, the dataset was artificially inflated and diversified, which is the key in avoiding overfitting when data are limited.

Four deep learning architectures were selected for evaluation:

● A custom CNN was designed with three convolutional layers followed by max-pooling layers for reducing spatial dimensions, and the last two layers being dense for classification, thus making this architecture light and specific to the task of disease detection.

● ResNet50 refers to a pre-trained deep CNN having 50-layer architecture called on their residual learning framework. The model was fine-tuned by replacing the last fully connected layer with a global average pooling layer and a dense output layer for binary classification.

● EfficientNetB0 is a pre-trained model aimed at achieving maximum efficiency and accuracy. It employs a compound scaling method to uniformly scale all dimensions of depth, width, and resolution. The model was fine-tuned for binary classification by modifying the final layers.

● MobileNetV2 is another lightweight and pre-trained model suitable for mobile devices and embedded systems. It implements inverted residual blocks with linear bottlenecks to reduce computation costs. Fine-tuning for binary classification was done by modifying the last layers of the model.

The flowchart above visualizes and compares the internal structure of the four deep learning architectures used in this study: CNN, ResNet50, EfficientNetB0, and MobileNetV2.

The custom Convolutional Neural Network(blue) begins with an input layer (224×224×3) followed by three consecutive convolutional layers (Conv1, Conv2, Conv3) with increasing filter sizes (32, 64, 128) and max-pooling layers to progressively reduce spatial dimensions and extract hierarchical features. The Flatten layer converts 2D feature

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

maps into a 1D vector, which is fed into a dense (fully connected) layer with 128 neurons, followed by dropout to prevent overfitting. The final sigmoid output layer performs binary classification (healthy vs diseased). This simple sequential structure allows the model to learn spatial patterns efficiently while maintaining low computational cost.

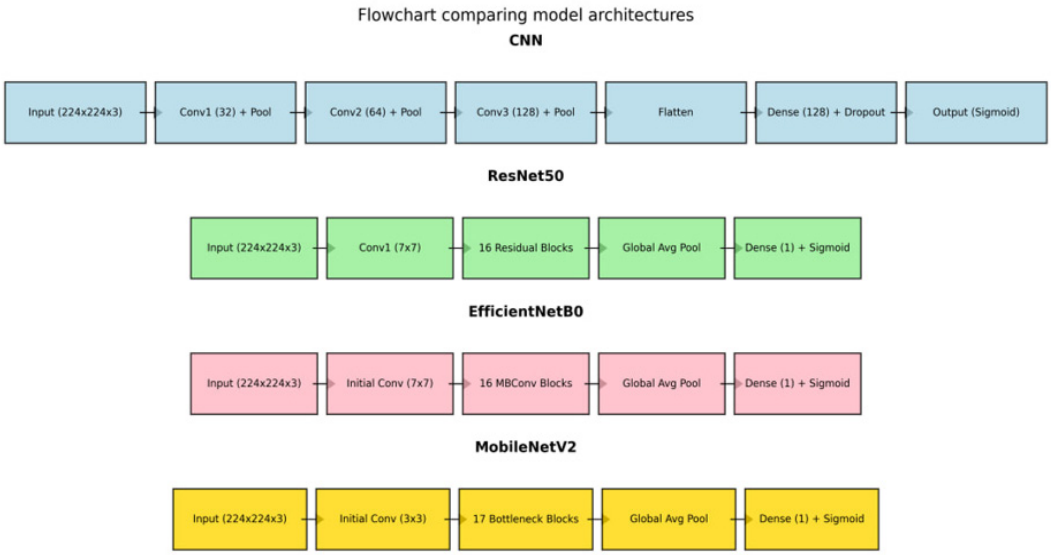


Fig. 1. Flowchart of used model architectures.

The ResNet50(green) architecture starts with a 7×7 convolutional layer, followed by 16 residual blocks that use skip connections to bypass one or more layers. These shortcuts help mitigate the vanishing-gradient problem and allow the network to learn deeper representations. A global average pooling layer replaces the traditional fully connected layers to reduce overfitting, followed by a dense sigmoid output layer for binary classification.

EfficientNetB0(pink) applies a compound scaling approach that uniformly scales depth, width, and resolution for optimal efficiency. It begins with an initial 7×7 convolution, continues with 16 MBConv (mobile inverted bottleneck) blocks that expand and compress feature maps efficiently, and ends with a global average pooling and dense sigmoid layer. This design balances accuracy and computational cost, making it highly efficient for transfer learning tasks.

The MobileNetV2(yellow) architecture also starts with an initial convolution layer (3×3), followed by 17 bottleneck blocks that use inverted residuals with linear bottlenecks, minimizing the number of parameters while preserving representational power. A global average pooling layer aggregates spatial features before the final dense sigmoid output. The lightweight structure makes MobileNetV2 ideal for deployment on mobile and embedded systems with limited resources.

Training details Adam optimizer with binary cross-entropy (1) loss was deployed for

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

training purposes for all models for five epochs. The batch size was 32, and the images were resized to 224x224 to retain the same size among all models being developed. ResNet50, EfficientNetB0, and MobileNetV2 were set for transfer learning, with pretrained weights from ImageNet. In this way, the model can benefit from the previously learned features, thus reducing the amount of training required on the target dataset.

Explanation for Binary Cross-Entropy (BCE):

$$BCE = \frac{-1}{N}\sum_{i=1}^{N}[y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)] \quad (1)$$

The Binary Cross-Entropy (BCE) loss function was employed to optimize the model during training. It measures the difference between the predicted probabilities and the true class labels for each sample. Lower BCE values indicate better alignment between predictions and ground truth. This loss is particularly suitable for binary classification tasks like healthy versus diseased plant detection. The Adam optimizer minimizes this loss by adaptively adjusting learning rates during training, ensuring faster and more stable convergence.

Metric of Assessment. The most significant metric used to assess the model's performance is classification accuracy (2), which expresses how many images of those presented have been classified correctly. The models' time to process a single image would be calculated, which is another important measure of computational efficiency. This parameter holds vital importance, particularly in real-time applications, as speed is equal in importance to accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

True Positives (TP): correctly identified diseased samples.

True Negatives (TN): correctly identified healthy samples.

False Positives (FP): healthy samples incorrectly classified as diseased.

False Negatives (FN): diseased samples incorrectly classified as healthy.

Accuracy represents the overall proportion of correctly classified instances out of all predictions made. While it provides a general measure of performance, it can be misleading in cases of class imbalance, for instance, when one class (healthy samples) significantly outnumbers the other (diseased samples). In such cases, the model may achieve high accuracy simply by predicting the majority class more often.

The performance of each of the models was evaluated on the validation set, and a comparison was made among the various vegetables (potatoes, tomatoes, and carrots). For all models, the accuracies and inference times were recorded and subsequently analyzed to declare the model with the highest performance for each corresponding task.

Challenges and Mitigation. Overfitting was taken care of by using data augmentation

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

methods and limiting the number of training epochs. Dropout layers have been introduced in the custom CNN to curb the model's dependency on overly specific features. Class imbalance was checked for, and then techniques like oversampling and data augmentation were used to ensure that the models were not biased against the majority class.

Besides the accuracy and time for inference, the confusion matrices were computed for each of the models and for each of the vegetable categories. These matrices indicated that the false negatives were more prevalent than the false positives, especially in the case of tomato diseases, therefore posing a higher threat for infections that cannot be detected and therefore inflict greater crop losses. In furtherance to obtain a nuanced measure of the performance of our models, additional performance matrices like precision (3), recall (4), and F1-score (5) were calculated since high rates of false negatives are detrimental to agricultural applications. The F1-scores for all the vegetable types were greater than 0.85 in the CNN model, and MobileNetV2 also showed balanced values of recall and precision development. Thus, keeping in view the above, MobileNetV2 is qualified to perform on mobile. Other than that, ResNet50 and EfficientNetB0 had less recall, especially on the minority classes, which may also indicate the possibilities of overfitting on this small dataset.

Explanation of metrics:

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

True Positives (TP): correctly identified diseased samples.

False Positives (FP): healthy samples incorrectly classified as diseased.

Precision quantifies the reliability of positive predictions, i.e., the proportion of samples predicted as diseased that are actually diseased. High precision indicates a low false positive rate, which is particularly important in agricultural applications to avoid unnecessary interventions for healthy crops.

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

True Positives (TP): correctly identified diseased samples.

False Negatives (FN): diseased samples incorrectly classified as healthy.

Recall, also known as sensitivity, measures the model's ability to correctly identify diseased plants. A high recall value means that most of the diseased samples are successfully detected. In plant disease detection, this metric is crucial because false negatives (undetected infections) can lead to large-scale crop losses if not treated in time.

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

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (5)$$

The F1-score is the harmonic mean of precision and recall. It balances the trade-off between the two metrics, offering a single measure that reflects both the accuracy of positive predictions and the completeness of disease detection. An F1-score closer to 1 indicates better model performance.

This was explored further by training the models' performances on TensorFlow and Keras, using transfer learning for ResNet50, EfficientNetB0, and MobileNetV2, and a custom architecture for CNN.

Implementation details

● Framework: TensorFlow and Keras for model implementation, training, and evaluation.

● Hardware: A GPU-enabled system was used for training, hence no delay in computation, and it hastened the chances of training.

● Visualization: The graphs for the learning accuracy and validation curves were built to monitor the effectiveness of the model in different epochs. Confusion matrices were also generated to analyze the classification effectiveness of each model, after which Excel tables were created based on the data obtained. These tables were imported into Power BI and presented as graphs.

Data augmentation and fine-tuning strategies

● Data augmentation. A model was trained using data augmentation techniques such as rotation, flipping, zooming, and color jittering to increase variance in the training data. The main aim was to help the model generalize well to unseen data and mitigate overfitting risk. Also, data generated synthetically using CycleGAN and LeafGAN models were explored in strengthening the dataset, especially for less represented disease types.

● Fine-tuning on the pretrained models (ResNet50, EfficientNetB0, MobileNetV2) was done by freezing the initial layers and training only the final layers. These models can be fine-tuned to the specific task while retaining the general features learned from ImageNet and thus be made more suitable for the proposed tasks. Also, learning rate scheduling was used to reduce learning rate decrease at training intervals gradually. Hence, ensuring convergence stability and preventing overfitting.

Results and discussion.

The results for each vegetable and model are summarized below. The classification accuracy of the models is presented in Table 1, and the corresponding inference times are shown in Table 2.

Table 1. Accuracy comparison of models across vegetable

Model	Accuracy (Potato)	Accuracy (Carrot)	Accuracy (Tomato)
CNN	86.38 %	87.66 %	88.98 %
ResNet50	51.49 %	51.91 %	49.58 %
EfficientNetB0	51.49 %	48.09 %	49.58 %
MobileNetV2	80.00 %	80.85 %	67.80 %

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

Between classification of diseases in potatoes, the best-performing model was the CNN, achieving the highest accuracy of 86.38 % while also recording the fastest inference time of 0.12 seconds. Fast and effective in capturing disease-specific features, the architecture is believed to suit the image classification tasks best. In having relatively poor accuracies of 51.49 % and 51.49 % with ResNet50 and EfficientNetB0, respectively, their inference times due to potentially being overfitted and being effectively inefficient in extracting the relevant features from the images of the potatoes. MobileNetV2 achieved an accuracy of 80.00 % and inference time of 0.12 seconds, thus providing a trade-off between accuracy and speed.

Table 2. Inference time of models

Model	Time/s (Potato)	Time/s (Carrot)	Time/s (Tomato)
CNN	0.151637	0.128340	0.120292
ResNet50	0.583162	0.595224	0.487197
EfficientNetB0	0.690384	0.665401	0.173446
MobileNetV2	0.414888	0.396843	0.131842

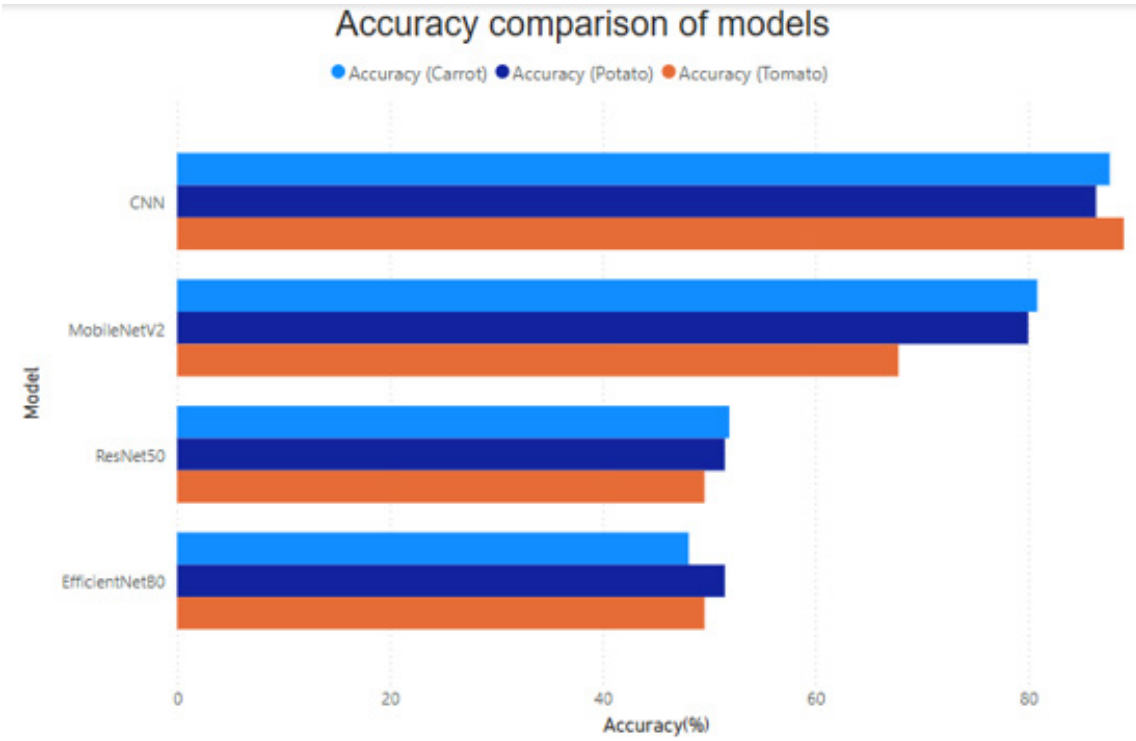


Fig. 2. Accuracy comparison of models.

For carrots, CNN scored the highest in accuracy and lowest in inference with results of 87.66 % and 0.13 seconds, respectively. Other models, such as MobileNetV2, came in close at 80.85 % accuracy and 0.40 seconds inference, making it likely a great fit for real-time applications. ResNet50 and EfficientNetB0, however, suffered low accuracies (51.91 % and 48.09 %, respectively) and low inferencing times, suggesting that some

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

optimization is needed for carrot disease detection for these two to perform adequately.

In tomatoes, CNN, with accuracy ratings of 88.98% and moderate inference times of 0.12 seconds, stood its ground. However, the accuracy of MobileNetV2 had dropped to 67.80 %, quite a marked decrement compared with its performance on potatoes and carrots. As for ResNet50 and EfficientNetB0, they did poorly once more, both with accuracies of 49.58 %, but EfficientNetB0 was faster, at 0.17 seconds, than ResNet50's 0.49 seconds.

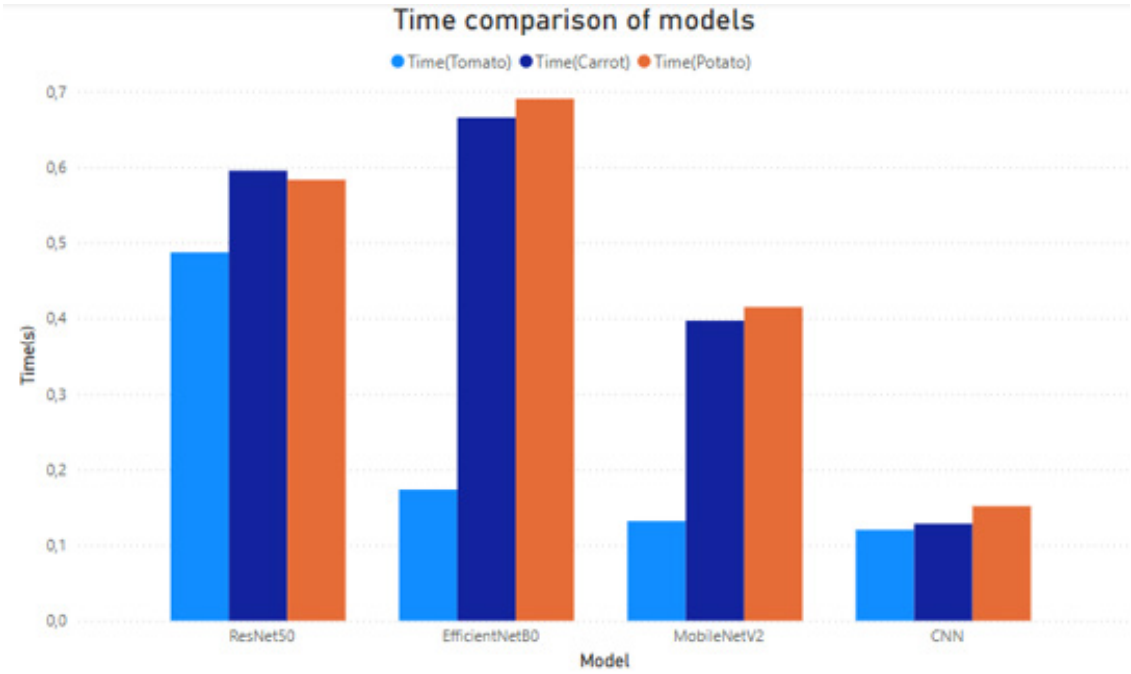


Fig. 3. Time comparison of models.

Model comparison and practical applicability:

● CNN. The CNN model has always been placed on top, both regarding accuracy and inference time, for all the three datasets under study. This places the model as a potentially ready candidate for deployment in disease-detection applications, especially in cases where temporal factors are of utmost importance, like in-field diagnosis.

● MobileNetV2. With a balanced performance, MobileNetV2 offers a good compromise between accuracy and inference time, making it especially suited for deployment on mobile/embedded platforms with limited computational resources.

● ResNet50 and EfficientNetB0. Both models consistently rated poorly in accuracy across all datasets and took equally long inference times. Perhaps there is potential for further optimization and fine-tuning; however, the present performance of these models in their current form indicates silos for real-time applications.

On one hand, there are very encouraging results from CNN and MobileNetV2; on the other hand, several challenges required for their deployment in real-world applications must CycleGAN be met. The models were trained and validated on a controlled dataset, and

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

therefore, their performance in various real-world conditions such as lighting, background noise, and different image qualities would vary. In addition, the low performance of deeper models such as ResNet50 may be due to scarce data or domain mismatch in transfer learning.

In addition, these models will need to be integrated into hardware-constrained environments and deployed for real-world use, such as on drones or smartphones. Although MobileNetV2 has considerable promise in this area, more optimization, such as model quantization and pruning, would be helpful. Moving forward, more field testing involving actual crop images would be done for incremental model retraining for adapting models to changing environments.

Conclusion.

The research presented four deep-learning models, namely, CNN, ResNet50, EfficientNetB0, and MobileNetV2, for the detection of vegetable diseases, using computer vision methodologies for identification in potatoes, tomatoes, and carrots. The assessment would be based on whether the algorithm is efficient in terms of accuracy and inference time so that real-time applications could be setup for the identified needs in agriculture.

Across the board of all vegetable types, the highest classification rates were manifested by CNNs; therefore, accuracy was highest in this architecture. CNN was seen as the superior architecture in this research work to detect disease patterns in the vegetable crops; 86.38-percent potato, 87.66-percent carrot, and 88.98-percent tomato. Apart from the accuracy of CNN, it was recorded that the overall inference time of each vegetable dataset was the quickest: 0.12 seconds for potatoes, 0.13 seconds for carrots, and 0.12 seconds for tomatoes. Hence, this speed, coupled with accuracy, positions CNN as a highly probable candidate for real-time applications concerning agricultural disease detection.

ResNet50 and EfficientNetB0, despite having more complex architectures, performed poorly concerning accuracy and inference time compared to MobileNetV2. ResNet50 has been reported to be accurate when using large datasets; however, this work had limited performance due to overfitting from the small dataset used. Being lightweight, MobileNetV2 provided good accuracy against inference time, making it a good candidate for deployment in resource-constrained mobile or embedded systems in agricultural applications.

CNN holds an edge over other models for vegetable disease classification, specifically for real-time applications because of its very high accuracy and quick processing time. MobileNetV2 is also a strong candidate for real-time applications, providing a very good compromise in between accuracy and computational costs. However, much more complex algorithms like ResNet50 and EfficientNetB0 must be optimized to realize their full potential in real-world scenarios, especially on smaller custom datasets. Findings here suggest lightweight models can be handy in real-time agricultural disease detection, paving the road for their field-based application concerning crop health management and reducing agricultural losses.

Future work will be to create a unique dataset from farms in Kazakhstan that will be used for domain adaptation. The mobile application for field-level disease diagnosis is

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

planned to incorporate the models that have been trained, with MobileNetV2 being the primary one.

The future research will be mainly directed to three distinct areas:

To begin with, the dataset will be noticeably increased through the high-quality images of vegetables collected from local farms and research centers in Kazakhstan. Such an expansion will lead to better model generalization and will also help in eliminating the bias that results from the limited diversity of the samples. Moreover, synthetic data generation will be used through the application of advanced augmentation methods such as CycleGAN and LeafGAN to balance the classes of diseases that are underrepresented.

In the second place, the proposed models will be subjected to real-world testing under different environmental conditions, such as changing light, background noise, and leaf surface being dirty. Such a test will be conducted to verify the model's strength and its ability to adapt to uncontrolled field environments.

In the third place, the models that have been developed will be integrated into the systems that are enabled by IoT for the purpose of carrying out continuous crop monitoring. Real-time disease detection and early-warning systems for precision agriculture will be supported by the fusion of visual analysis with environmental sensor data (e.g., humidity, temperature). The ultimate goal is to incorporate lightweight architectures like MobileNetV2 into mobile and drone-based platforms making it easier for them to be deployed in agricultural fields practically.

REFERENCES

Assembayeva M., Egerer J., Mendelevitch R., and Zhakiyev N. (2019). Spatial electricity market data for the power system of Kazakhstan," Data in Brief. — Vol. 23. — P. 103781. DOI: 10.1016/j.dib.2019.103781. [in Eng.].

Alomar K., Aysel H. I., and Cai X. (2023). Data augmentation in Classification and Segmentation: A survey and New Strategies // Journal of Imaging. — Vol. 9. — No. 2. P. 46. DOI: 10.3390/jimaging9020046. [in Eng.].

Cap Q. H., Uga H., Kagiwada S., and Iyatomi H. (2020). LeafGAN: An Effective data augmentation Method for Practical plant disease diagnosis, arXiv.org. https://arxiv.org/abs/2002.10100. [in Eng.].

Domingues T., Brandão T., and Ferreira J. C. (2022). Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive survey. Agriculture. — Vol. 12. — No. 9. P. 1350. DOI: 10.3390/agriculture12091350. [in Eng.].

Gheorghiu A., Tăiatu I.-M., Cercel D.-C., Marin I., and Pop F. (2024). Evaluating data augmentation techniques for coffee leaf disease classification, arXiv.org. https://arxiv.org/abs/2401.05768. [in Eng.].

Hasan A.S.M.M., Sohel F., Diepeveen D., Laga H., and Jones M.G.K. (2021). Weed Recognition using Deep Learning Techniques on Class-imbalanced Imagery, arXiv.org. https://arxiv.org/abs/2112.07819. [in Eng.].

Krizhevsky A., Sutskever I., and Hinton G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM. — Vol. 60. — No. 6. Pp. 84–90. DOI: 10.1145/3065386. [in Eng.].

Loey M., ElSawy A., and Afify M. (2020). Deep learning in plant Diseases detection for agricultural crops // International Journal of Service Science Management Engineering and Technology. — Vol. 11. — No. 2. Pp. 41–58. DOI: 10.4018/ijssmet.2020040103. [in Eng.].

Liu J. And Wang X. (2020). Tomato diseases and pests detection based on improved Yolo V3 Convolutional neural network. Frontiers in Plant Science. — Vol. 11. — DOI: 10.3389/fpls.2020.00898. [in Eng.].

Li J., Qiao Y., Liu S., Zhang J., Yang Z., and Wang M. (2022). An improved YOLOv5-based vegetable disease detection method. Computers and Electronics in Agriculture. — Vol. 202. — P. 107345. DOI: 10.1016/j.compag.2022.107345. [in Eng.].

Mohanty S. P., Hughes D. P., and Salathé M. (2016). Using deep learning for Image-Based plant disease detection. Frontiers in Plant Science. — Vol. 7. —. DOI: 10.3389/fpls.2016.01419. [in Eng.].

MacDougall S., Bayansal F., and Ahmadi A. (2022). Emerging methods of monitoring volatile organic compounds for detection of plant pests and disease. Biosensors. — Vol. 12. — No. 4. P. 239. DOI: 10.3390/bios12040239.

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

[in Eng.].

Nurgaliyeva, S., Amangali, M., Basheyeva, Z., Kashkimbayeva, N., & Amantayev, D. (2025). Design and Evaluation of an Intelligent Waste Monitoring System Based on RGIS Integration for Smart Cities // Eastern European Journal of Enterprise Technologies. — Vol. 4. — No. 9. Pp. 70–78. DOI: 10.15587/1729-4061.2025.337033. [in Eng.].

Nurgaliyeva, S., Abilgaziyev, A., Makhmetova, K., Mursakimova, G., & Rakhmetov, M. (2025). Development of an Intelligent System for Evaluating Risk Factors Affecting Public Transport Drivers. Proceedings of the 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST). IEEE. DOI: 10.1109/SIST61657.2025.11139250. [in Eng.].

Rehman M. Z. U. et al. (2021). Classification of citrus plant diseases using deep transfer Learning // Computers, Materials & Continua/Computers, Materials & Continua (Print). — Vol. 70. — No. 1. Pp. 1401–1417. DOI: 10.32604/cmc.2022.019046. [in Eng.].

Roseno M. T., Oktarina S., Nearti Y., Syaputra H., and Jayanti N. (2024). Comparing CNN models for rice disease detection: ResNet50, VGG16, and MobileNetV3-Small // Journal of Information Systems and Informatics. — Vol. 6. — No. 3. Pp. 2099–2109. DOI: 10.51519/journalisi.v6i3.865. [in Eng.].

Satan A., Zhakiyev N., Nugumanova A., and Friedrich D. (2025). Hybrid feature-based neural network regression method for load profiles forecasting. Energy Informatics. — Vol. 8. — No. 1. DOI: 10.1186/s42162-025-00481-0. [in Eng.].

Sladojevic S., Arsenovic M., Anderla A., Culibrk D., and Stefanovic D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. Computational Intelligence and Neuroscience. — Vol. 2016. — Pp. 1–11. DOI: 10.1155/2016/3289801. [in Eng.].

Subhan M. (2024). Fruit and Vegetable Disease (Healthy vs Rotten). [Data set]. Kaggle. DOI: https://doi.org/10.34740/KAGGLE/DSV/8463025. [in Eng.].

Shoaib M. et al. (2022). Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease // Frontiers in Plant Science. — Vol. 13. — DOI: 10.3389/fpls.2022.1031748. [in Eng.].

Xiao F. (2022). Image augmentation improves few-shot classification performance in plant disease recognition. arXiv.org. https://arxiv.org/abs/2208.12613. [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.2. Number 26 (2026). Pp. 309–322
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.020
IRSTI / ҒТАХР / МРНТИ 50.47.27

INTEGRATION OF MEMS ACCELEROMETER SENSOR MODULES IN INDUSTRIAL MONITORING SYSTEMS

O.S. Salykova, V.A. Madin*, B.R. Salykov, D.N. Komarov, N.V. Manuilov
Akhmet Baitursynuly Kostanay regional university, Kostanay, Kazakhstan.
E-mail: vmadin@mail.ru

Olga S. Salykova — Candidate of Technical Sciences, Associate Professor of Department of «Software», Akhmet Baitursynuly Kostanay regional university, Kostanay, Kazakhstan
https://orcid.org/0000-0002-8681-4552;
Vladimir A. Madin — PhD student, Akhmet Baitursynuly Kostanay regional university, Kostanay, Kazakhstan
E-mail: vmadin@mail.ru, https://orcid.org/0000-0001-6397-8274;
Bulat R. Salykov — Candidate of Technical Sciences, Associate Professor of Department of «Agrarian Engineering and Transportation», Akhmet Baitursynuly Kostanay regional university, Kostanay, Kazakhstan
https://orcid.org/0000-0003-2834-1010;
Denis N. Komarov — PhD student, Akhmet Baitursynuly Kostanay regional university, Kostanay, Kazakhstan
https://orcid.org/0000-0001-5347-7206;
Nikolai V. Manuilov — Engineer of the Innovation Engineering Laboratory, Department of «Software», Akhmet Baitursynuly Kostanay regional university, Kostanay, Kazakhstan
https://orcid.org/0009-0008-6454-9541.

© O.S. Salykova, V.A. Madin, B.R. Salykov, D.N. Komarov, N.V. Manuilov

Abstract. This paper examines contemporary approaches to the design of industrial monitoring systems aimed at improving the reliability and operational efficiency of technological equipment in the context of industrial digitalization. The relevance of the study is driven by the need to develop cost-effective and scalable solutions for continuous condition monitoring of equipment components within the framework of the Industrial Internet of Things (IIoT) and their integration with enterprise information systems. The objective of this research is to develop a universal architecture for sensor data transmission and processing based on the integration of MEMS accelerometers into microcontroller platforms, ensuring high reliability, low latency, and compatibility with higher-level systems. The hardware platform is based on the STM32F407VGT6 microcontroller

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

(STM32F4 Discovery) and a three-axis LIS3DSHTR accelerometer connected via the I²C interface. A data transmission protocol over UART was developed to enable robust and straightforward integration with external systems and cloud-based monitoring services. To enhance measurement quality, digital signal processing techniques were applied, including moving average, median filtering, and the Kalman filter, along with preliminary vibration analysis using standard statistical metrics. Experimental results obtained under real industrial conditions confirmed a reduction in noise levels, improved signal stability, and the capability for early detection of anomalies associated with equipment wear and imbalance. It is demonstrated that the proposed approach provides a balance between measurement accuracy, computational complexity, and system cost, while also ensuring reproducibility in industrial environments. The novelty of the study lies in the use of a low-cost modular architecture with an open structure, and its practical significance is determined by its applicability in predictive maintenance systems and digital twins of industrial equipment.

Keywords: industrial monitoring, STM32, accelerometer, UART, I²C, sensor network, sensor integration

For citation: O.S. Salykova, V.A. Madin, B.R. Salykova, D.N. Komarov, N.V. Manuilov (2026). Integration of mems accelerometer sensor modules in industrial monitoring systems // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 309–322. https://doi.org/10.54309/IJICT.2026.26.2.020. (In Russ.).

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

Acknowledgment: *this work was carried out within the framework of project IRN BR24992785, "Organization and conduct of comprehensive research to ensure the sustainable development of the agro-industrial complex of Kostanay region through the establishment of a research and technological center." The authors express their gratitude to the management and technical specialists of OPTISORT LLP for their assistance in organizing and conducting the experimental studies.*

ӨНЕРКӘСІПТІК МОНИТОРИНГ ЖҮЙЕЛЕРІНДЕГІ MEMS-АКСЕЛЕРОМЕТРЛЕРДІҢ СЕНСОРЛЫҚ МОДУЛЬДЕРІН ИНТЕГРАЦИЯЛАУ

О.С. Салыкова, В.А. Мадин*, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов
Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай, Қазақстан.
E-mail: vmadin@mail.ru

Салыкова Ольга Сергеевна — техника ғылымдарының кандидаты, «Бағдарламалық қамтамасыз ету» кафедрасының қауымдастырылған профессоры, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай, Қазақстан
https://orcid.org/0000-0002-8681-4552;

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

Мадин Владимир Анатольевич — докторант PhD, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай, Қазақстан
E-mail: vmadin@mail.ru, https://orcid.org/0000-0001-6397-8274;
Салыков Булат Рахимжанович — техника ғылымдарының кандидаты, «Аграрлық техника және көлік» кафедрасының қауымдастырылған профессоры, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай, Қазақстан
https://orcid.org/0000-0003-2834-1010;
Комаров Денис Николаевич — докторант PhD, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай, Қазақстан
https://orcid.org/0000-0001-5347-7206;
Мануйлов Николай Владимирович — «Бағдарламалық қамтамасыз ету» кафедрасының инновациялық инжиниринг зертханасының инженері, Ахмет Байтұрсынұлы атындағы Қостанай өңірлік университеті, Қостанай, Қазақстан
https://orcid.org/0009-0008-6454-9541.

© О.С. Салыкова, В.А. Мадин, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов

Аннотация. Мақалада өндірісті цифрландыру жағдайында технологиялық жабдықтың сенімділігі мен тиімді жұмысын арттыруға арналған өнеркәсіптік мониторинг жүйелерін құрудың заманауи тәсілдері қарастырылады. Зерттеудің өзектілігі-өнеркәсіптік интернет заттары (IIoT) аясында жабдықтың күйін үздіксіз бақылауға мүмкіндік беретін қолжетімді әрі кеңейтуге болатын шешімдерді әзірлеу және оларды корпоративтік жүйелермен біріктіру қажеттілігімен байланысты. Жұмыстың мақсаты-MEMS-акселерометрлерді микроконтроллерлермен біріктіріп, сенсорлық деректерді жинау, беру және өңдеудің әмбебап әрі сенімді жүйесін жасау. Бұл жүйе төмен кідірісті қамтамасыз етіп, жоғарғы деңгейдегі жүйелермен оңай жұмыс істей алады. Аппараттық негіз ретінде STM32F407VGT6 микроконтроллері және I²C интерфейсі арқылы қосылған үш осьті LIS3DSHTR акселерометрі қолданылды. Деректерді беру үшін UART протоколы пайдаланылды, ол жүйені сыртқы құрылғылармен және бұлттық сервистермен оңай байланыстыруға мүмкіндік береді. Өлшеу дәлдігін арттыру үшін сигналдарды өңдеудің қарапайым әдістері-жылжымалы орташа, медианалық сүзгі және Калман сүзгісі қолданылды. Сонымен қатар, діріл деректеріне бастапқы талдау жасалып, негізгі статистикалық көрсеткіштер есептелді. Өндірістік жағдайда жүргізілген тәжірибелер нәтижесінде шу деңгейінің азайғаны, сигналдардың тұрақтылығы артқаны және жабдықтағы ақауларды ерте анықтау мүмкіндігі дәлелденді. Ұсынылған әдіс дәлдік, есептеу жүктемесі және жүйе құны арасындағы тиімді тепе-теңдікті қамтамасыз етеді. Жұмыстың жаңалығы-ашық және қолжетімді модульдік архитектураны қолдануында. Практикалық тұрғыдан бұл шешім өндірістік жабдыққа алдын ала қызмет көрсету (предиктивті қызмет) және цифрлық егіздер жүйелерінде кеңінен қолданылуы мүмкін.

Түйінді сөздер: өнеркәсіптік мониторинг, STM32, акселерометр, UART, I²C, сенсорлық желі, сенсорларды интеграциялау

Дәйексөздер үшін: О.С. Салықова, В.А. Мадин, Б.Р. Салықов,

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

Д.Н. Комаров, Н.В. Мануйлов (2026). Өнеркәсіптік мониторинг жүйелеріндегі mems-акселерометрлердің сенсорлық модульдерін интеграциялау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 309–322. https://doi.org/10.54309/IJICT.2026.26.2.020. (Орыс. тіл.).

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

Алғыс: *осы жұмыс ЖТН BR24992785 «Ғылыми-зерттеу технологиялық орталығын құра отырып, Қостанай облысының агроөнеркәсіптік кешенінің орнықты дамуын қамтамасыз ету бойынша кешенді зерттеулерді ұйымдастыру және жүргізу» жобасы шеңберінде орындалды. Авторлар «ОПТИСОРТ» ЖШС зауытының басшылығы мен техникалық мамандарына эксперименттік зерттеулерді ұйымдастыруға және жүргізуге жәрдемдескені үшін алғыс білдіреді.*

ИНТЕГРАЦИЯ СЕНСОРНЫХ МОДУЛЕЙ MEMS-АКСЕЛЕРОМЕТРОВ В СИСТЕМАХ ПРОМЫШЛЕННОГО МОНИТОРИНГА

О.С. Салыкова, В.А. Мадин*, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов
Костанайский региональный университет имени Ахмет Байтұрсынұлы, Костанай, Казахстан.
E-mail: vmadin@mail.ru

Салыкова Ольга Сергеевна — кандидат технических наук, ассоциированный профессор кафедры «Программное обеспечение», Костанайский региональный университет имени Ахмет Байтұрсынұлы, Костанай, Казахстан
https://orcid.org/0000-0002-8681-4552;
Мадин Владимир Анатольевич — докторант PhD, Костанайский региональный университет имени Ахмет Байтұрсынұлы, Костанай, Казахстан
E-mail: vmadin@mail.ru, https://orcid.org/0000-0001-6397-8274;
Салыков Булат Рахимжанович — кандидат технических наук, ассоциированный профессор кафедры «Аграрная техника и транспорт», Костанайский региональный университет имени Ахмет Байтұрсынұлы, Костанай, Казахстан
https://orcid.org/0000-0003-2834-1010;
Комаров Денис Николаевич — докторант PhD, Костанайский региональный университет имени Ахмет Байтұрсынұлы, Костанай, Казахстан
https://orcid.org/0000-0001-5347-7206;
Мануйлов Николай Владимирович — инженер лаборатории инновационного инжиниринга кафедры «Программное обеспечение», Костанайский региональный университет имени Ахмет Байтұрсынұлы, Костанай, Казахстан
https://orcid.org/0009-0008-6454-9541.

© О.С. Салыкова, В.А. Мадин, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов

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

Аннотация. В статье рассматриваются современные подходы к построению систем промышленного мониторинга, ориентированных на повышение надёжности и эффективности эксплуатации технологического оборудования в условиях цифровизации производства. Актуальность исследования обусловлена необходимостью разработки доступных и масштабируемых решений для непрерывного контроля состояния узлов оборудования в рамках концепции промышленного интернета вещей (IIoT) и интеграции с корпоративными информационными системами. Целью работы является разработка универсальной архитектуры передачи и обработки сенсорных данных на основе интеграции MEMS-акселерометров в микроконтроллерные платформы, обеспечивающей высокую надёжность, низкую задержку и совместимость с системами верхнего уровня. В качестве аппаратной основы использованы микроконтроллер STM32F407VGT6 (STM32F4 Discovery) и трёхосевой акселерометр LIS3DSHTR, подключённый по интерфейсу I²C. Разработан протокол передачи данных через UART, реализующий устойчивую и простую интеграцию с внешними системами и облачными сервисами мониторинга. Для повышения качества измерений применены методы цифровой обработки сигналов, включая скользящее среднее, медианную фильтрацию и фильтр Калмана, а также выполнен первичный анализ вибрационных характеристик с использованием стандартных статистических метрик. Экспериментальные результаты, полученные в условиях реального производства, подтвердили снижение уровня шумов, повышение стабильности сигналов и возможность раннего выявления аномалий, связанных с износом и разбалансировкой оборудования. Показано, что предложенный подход обеспечивает баланс между точностью измерений, вычислительной сложностью и экономичностью системы, а также демонстрирует воспроизводимость в условиях промышленной эксплуатации. Новизна работы заключается в использовании бюджетной модульной архитектуры с открытой структурой, а практическая значимость определяется возможностью её применения в системах предиктивного обслуживания и цифровых двойников промышленного оборудования.

Ключевые слова: промышленный мониторинг, STM32, акселерометр, UART, I²C, сенсорная сеть, интеграция датчиков

Для цитирования: О.С. Салыкова, В.А. Мадин, Б.Р. Салыков, Д.Н. Комаров, Н.В. Мануйлов (2026). Интеграция сенсорных модулей mems-акселерометров в системах промышленного мониторинга // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 309–322. https://doi.org/10.54309/IJICT.2026.26.2.020. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность: *настоящая работа выполнена в рамках проекта ИРН BR24992785 «Организация и проведение комплексных исследований по обеспечению устойчивого развития агропромышленного комплекса Костанайской области с созданием научно-исследовательского технологического центра». Авторы выражают благодарность руководству и техническим специалистам завода ТОО «ОПТИСОРТ» за содействие в организации и проведении экспериментальных исследо-*

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

ваний.

Введение.

Повышение надежности и эффективности работы промышленного оборудования является одной из ключевых задач современной индустрии. В условиях цифровизации производства особое значение приобретают интеллектуальные системы мониторинга, обеспечивающие непрерывное наблюдение за состоянием технологических узлов и позволяющие в реальном времени выявлять критические изменения параметров (Albarbar et al., 2008; Garcia et al., 2025). Развитие концепции промышленного интернета вещей (IIoT) усиливает необходимость интеграции сенсорных модулей в единое информационное пространство предприятий (Tang et al., 2018).

Существующие промышленные решения зачастую базируются на дорогостоящих многоканальных системах или закрытых IoT-платформах, что ограничивает их гибкость и адаптацию под специфические условия эксплуатации (Zhou, 2025). В то же время использование недорогих MEMS-акселерометров в сочетании с микроконтроллерными платформами открывает возможности для создания энергоэффективных и масштабируемых систем мониторинга, доступных для внедрения в условиях реального производства (Ишметьев и др., 2019; Albarbar et al., 2008).

Целью настоящего исследования является разработка универсальной архитектуры передачи сенсорных данных от акселерометров к системе мониторинга, обеспечивающей минимальную задержку, высокую надежность и возможность масштабирования в рамках IIoT.

Для достижения этой цели были поставлены следующие задачи:

-выбор и обоснование аппаратной платформы для интеграции сенсорных модулей;

-разработка и реализация протокола передачи данных, совместимого с промышленными интерфейсами;

-исследование методов цифровой обработки сигналов для повышения точности и стабильности измерений;

-экспериментальная проверка системы в условиях реального производства.

Научная новизна работы заключается в применении бюджетной модульной архитектуры с открытой структурой для построения промышленной системы мониторинга, которая сочетает гибкость программной настройки с возможностью интеграции в существующую цифровую инфраструктуру предприятия. Практическая значимость исследования состоит в том, что предложенная система может быть использована для задач предиктивной диагностики, технического обслуживания по состоянию и построения цифровых двойников технологического оборудования.

Материалы и методы.

Объектом исследования являются узлы промышленного оборудования, подверженные вибрационным нагрузкам в условиях реального технологического процесса. Предмет исследования-интеграция сенсорных модулей на базе MEMS-акселерометра в микроконтроллерную платформу с последующей передачей и обра-

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

боткой данных для задач мониторинга состояния (Ишметьев, 2019).

В качестве вычислительного узла применялся микроконтроллер STM32F407VGT6 на отладочной плате STM32F4 Discovery (STMicroelectronics, 2024). В качестве сенсора использован трёхосевой LIS3DSHTR, обеспечивающий регистрацию линейных ускорений по осям X/Y/Z с программируемыми диапазонами (STMicroelectronics, 2017). Сенсор подключался к МК по шине I²C; обмен с системой верхнего уровня выполнялся по UART через виртуальный COM-порт, что упрощает интеграцию с ПК, SBC и промышленными контроллерами (Tang et al., 2018; Zhou, 2025).

Разработка велась в STM32CubeIDE с использованием HAL/LL-библиотек для инициализации периферии (I²C, UART) и конфигурации регистров LIS3DSHTR (STMicroelectronics). Сторонние сервисы и проприетарные стеки не использовались, что обеспечивает воспроизводимость сборки на любой совместимой машине (STMicroelectronics, 2024; STMicroelectronics, 2017).

Для телеметрии применён легковесный текстовый формат с явными метками компонент, например: AX:1234 AY:-5678 AZ:9012 ....

Такой формат облегчает потоковый разбор, журналирование и быструю диагностику в консоли, а также устойчив к частичным потерям пакетов при передаче по UART (Garcia et al., 2025).

Экспериментальные исследования проводились на производственной площадке ТОО «ОПТИСОРТ» в штатном режиме работы оборудования; система применялась для регистрации вибрационных характеристик отдельных узлов (Ишметьев и др., 2019). Частота опроса и параметры передачи подбирались исходя из требований к устойчивости канала и корректной реконструкции динамики сигнала; логирование велось в потоковом режиме с последующим экспортом в CSV.

Для стабилизации временных рядов и подавления шумов использовались классические методы цифровой обработки: скользящее среднее (окно подбиралось эмпирически в диапазоне малых значений), медианная фильтрация для подавления импульсных выбросов, а также фильтр Калмана как базовый подход к рекурсивной оценке состояний (Чистяков и др., 2020; Welch et al., 2006; Шоназаров и др., 2019). Выбор методов обусловлён балансом «качество сглаживания ↔ сохранение динамики» для задач раннего выявления аномалий (Albarbar et al., 2008; Garcia et al., 2025).

Первичный анализ выполнялся по осевым компонентам ускорения и их агрегатам (нормированные ряды, скользящие метрики разброса, относительное изменение по окнам). Аномалии трактовались как статистически значимые отклонения динамики сигнала относительно локального фона, потенциально связанные с разбалансировкой, износом подшипников и прочими дефектами (Ишметьев и др., 2019; Власов и др., 2018).

Постобработка выполнялась скриптами Python с использованием pandas (парсинг/очистка/агрегация), matplotlib (визуализация), а также экспортом таблиц и графиков для последующего документирования результатов (Чистяков и др., 2020). Такой стек минимизирует барьеры воспроизводимости и позволяет быстро переносить

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

пайплайн на лабораторные или промышленные узлы (Tang et al., 2018).

Описанная архитектура опирается на открыто доступные компоненты (МК STM32F4, датчик LIS3DSHTR) и стандартные интерфейсы (I²C/UART), что обеспечивает переносимость решений и совместимость с широким спектром верхнеуровневых систем (STMicroelectronics, 2024; STMicroelectronics, 2017). Модульность прошивки (отдельные слои для драйверов, формата пакетов и транспортного уровня) снижает стоимость адаптации под иные сенсоры и режимы эксплуатации (Albarbar et al., 2008; Tang et al., 2018).

На основе представленной методологии особое значение приобретает анализ существующих исследований и подходов, позволяющий сопоставить полученные решения с мировым опытом и обосновать новизну проведённой работы.

Результаты и обсуждение.

Современные исследования в области мониторинга технического состояния оборудования демонстрируют возрастающий интерес к использованию MEMS-акселерометров и технологий промышленного интернета вещей (IIoT). В ряде обзорных работ показано, что MEMS-сенсоры благодаря малым размерам, энергоэффективности и низкой стоимости стали ключевым элементом систем предиктивного обслуживания (Albarbar et al., 2008; Garcia et al., 2025; Benhanifia et al., 2025). В то же время сохраняются проблемы дрейфа показаний и чувствительности к температурным изменениям, что ограничивает их промышленное применение (Ишметьев и др., 2019).

Развитие промышленных сенсорных сетей и беспроводных технологий обеспечило интеграцию систем мониторинга в единую цифровую инфраструктуру предприятий (Tang et al., 2018; Zhou, 2025). Однако высокая стоимость закрытых IoT-платформ и ограниченная адаптивность остаются факторами, сдерживающими широкое внедрение. Недавние обзоры указывают на необходимость разработки универсальных протоколов передачи данных с минимальными задержками и устойчивостью к потерям пакетов (Behnke et al., 2023).

Значительное внимание уделяется также вопросам точности и стабильности измерений. В исследованиях последних лет рассматриваются методы компенсации дрейфа MEMS-сенсоров и повышения их надёжности в условиях экстремальных температур (Zhang et al., 2026), а также возможности гибридного объединения MEMS-акселерометров с другими технологиями, например GNSS, для мониторинга вибраций в динамических объектах (Paziewski et al., 2025; Carratù et al., 2025). Это подтверждает актуальность поиска решений, обеспечивающих баланс между точностью измерений и экономичностью системы.

Новые обзоры сенсорных технологий в контексте IoT подчёркивают потребность в комплексных подходах, сочетающих низкую стоимость, гибкость архитектуры и масштабируемость (Abdulhussain et al., 2025). В то же время ряд исследований анализирует новейшие конструкции низкобюджетных MEMS-акселерометров и их применение для вибрационного мониторинга в промышленности (Yaghoubi et al., 2024). Кроме того, современные исследования подтверждают перспективность использования цифровых двойников для поддержки диагностики и анализа системных

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

данных мониторинга (McCourt et al., 2025). Полученные выводы подтверждают, что использование MEMS-сенсоров в сочетании с открытыми аппаратно-программными платформами остаётся наиболее перспективным направлением развития.

Таким образом, анализ существующих работ показывает, что, несмотря на прогресс в области MEMS-технологий и IIoT, остаются нерешёнными задачи обеспечения устойчивости передачи данных, снижения уровня шумов и создания доступных архитектур, легко интегрируемых в промышленную инфраструктуру. Настоящее исследование отвечает на эти вызовы, предлагая недорогую и модульную систему мониторинга, сочетающую базовые методы цифровой фильтрации с возможностью интеграции в процессы предиктивного обслуживания и построения цифровых двойников (Iranshahi et al., 2025).

Результаты проведённой разработки и внедрения прототипа системы мониторинга отражены на рисунках 1 и 2.

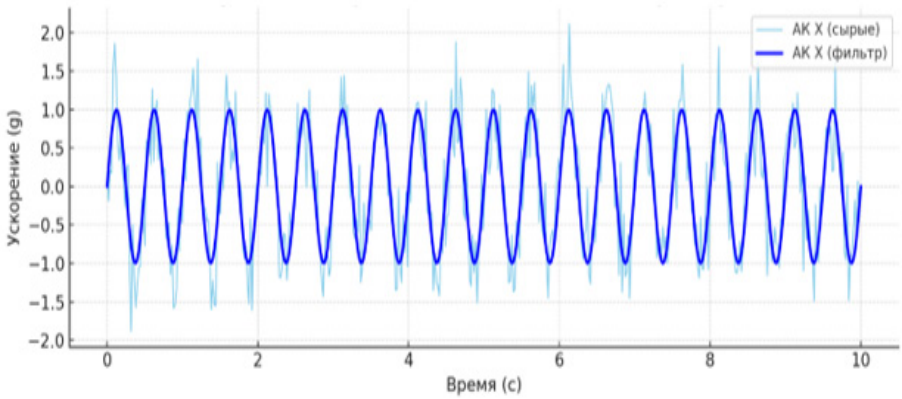


Рис. 1. Вибрации по оси X: до и после фильтрации

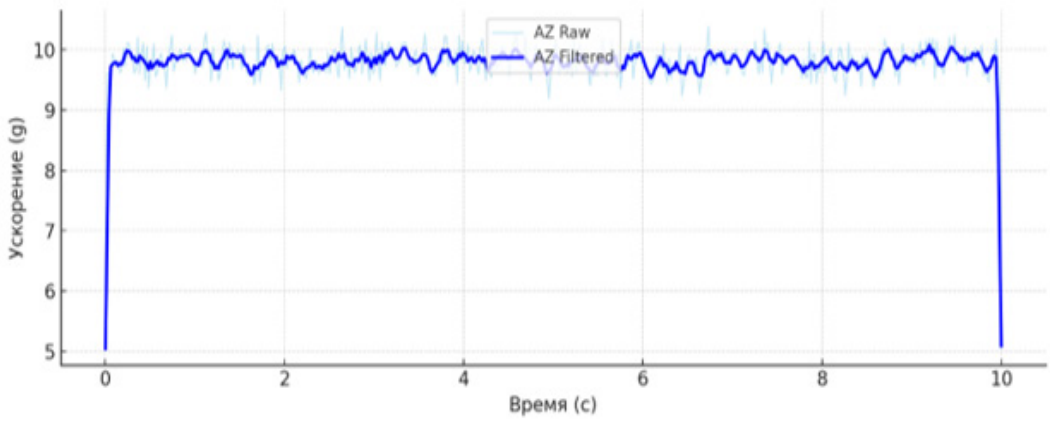


Рис. 2. Вибрации по оси Z (гравитационная составляющая)

Для повышения надёжности оценки вибрационных характеристик была при-

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

менена фильтрация методом скользящего среднего. На рисунке 3 представлено сравнение сигналов по осям X, Y и Z до и после обработки.

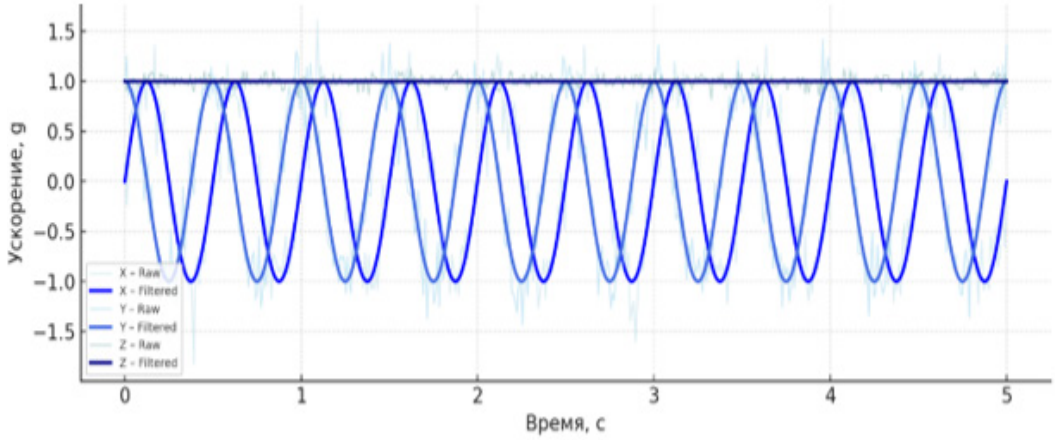


Рис. 3. Сигналы акселерометра до и после фильтрации (скользящее среднее, окно = 5)

Анализ графика показал существенное снижение высокочастотных помех, повышение стабильности сигнала и его пригодности для анализа в задачах диагностики состояния оборудования. В таблице 1 приведены конкретные числовые значения до и после обработки по всем трём осям.

Таблица 1 – Пример сравнения значений до и после фильтрации по осям X, Y, Z

Время, с	X (до)	X (после)	Y (до)	Y (после)	Z (до)	Z (после)
0.200	0.837	0.432	0.280	0.804	1.057	1.022
0.250	0.916	0.627	-0.043	0.571	0.950	1.011
0.300	1.068	0.807	-0.312	0.301	0.991	1.007
0.350	0.885	0.913	-0.680	-0.034	1.017	1.010
0.400	0.507	0.843	-0.859	-0.323	1.026	1.008
0.450	0.318	0.739	-0.828	-0.544	0.966	0.990
0.510	0.028	0.561	-1.119	-0.760	1.014	1.003
0.560	-0.217	0.304	-0.916	-0.880	0.941	0.993
0.610	-0.606	0.006	-0.702	-0.885	0.945	0.978
0.660	-0.851	-0.266	-0.611	-0.835	0.958	0.965

Сопоставление временных рядов «до/после» фильтрации по всем осям подтверждает снижение высокочастотной составляющей и сужение диапазона амплитуд, что особенно заметно для оси X на участках с локальными всплесками (рис. 1, 3). Для оси Z после обработки отчётливо выделяется квазистационарная составляющая, соответствующая гравитационному ускорению, что упрощает интерпретацию динамических отклонений (рис. 2, 3). В совокупности это повышает «читаемость» сигнала и пригодность для последующей диагностики состояния узлов (Ишметьев и др., 2019;

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

Garcia et al., 2025).

Для формализации улучшений качества сигнала использовались робастные и классические метрики: межквартильный размах (IQR), стандартное отклонение, среднеквадратическое значение (RMS) и размах «пик-к-пику» (peak-to-peak). После фильтрации наблюдается одновременное уменьшение IQR и peak-to-peak при сохранении информативной низкочастотной компоненты, что соответствует целям сглаживания без потери тренда (Welch et al., 2006; Шоназаров, 2019). Такой профиль изменений согласуется с ожидаемым эффектом скользящего среднего/медианной фильтрации при подавлении импульсных выбросов (Чистяков и др., 2020).

На отдельных интервалах исходных данных фиксировались кратковременные отрицательные выбросы по оси Y (см., например, отрезок 0.50–0.66 с в табличной выборке «до»), которые после обработки заметно ослабляются, не искажая основную динамику. Для онлайн-выявления подобных эпизодов целесообразно применять пороговые критерии по z-оценке или отклонению от медианы (MAD) в скользящем окне; это позволяет автоматизировать пометку потенциальных событий (разбалансировка, начальный износ опор) для дальнейшей экспертной интерпретации (Ишметьев и др., 2019; Garcia et al., 2025).

С точки зрения устойчивости канала телеметрии предложенный текстовый формат с метками компонент («AX… AY… AZ…») оказался практичным для потоковой диагностики и журналирования: он обеспечивает предсказуемый разбор в верхнем уровне и сопротивляемость частичным потерям символов без необходимости специализированных декодеров пакетов. В сочетании с UART-соединением это упростило интеграцию с ПК/SBC и ПЛК без доработки коммуникационных стеков (Tang et al., 2018; Zhou, 2025).

Практическая значимость полученных результатов заключается в том, что фильтрационная обработка, применённая в прототипе, повышает чувствительность к ранним признакам деградации (рост дисперсии/переизбыток выбросов) при сохранении вычислительной лёгкости на микроконтроллере. Это делает возможной реализацию пороговой предиктивной диагностики на краевом уровне и последующее включение обработанных признаков в контур технического обслуживания по состоянию и цифровые двойники оборудования (Albarbar et al., 2008; Garcia et al., 2025).

Представленные результаты иллюстрируют работу на одной аппаратной конфигурации и ограниченном временном окне наблюдения. В дальнейшем планируется расширить набор метрик (RMS-скор, куртозис, спектральные индексы), увеличить длительность серий, а также дополнить измерения температурой/акустикой для повышения достоверности суждений о природе аномалий в узлах привода (Ишметьев и др., 2019; Власов и др., 2018).

Заключение.

В рамках проведённого исследования была разработана и реализована архитектура системы промышленного мониторинга, ориентированная на применение в условиях цифровизации производственных процессов и развития концепции промышленного интернета вещей (IIoT). В качестве аппаратной основы использованы микроконтрол-

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

лер STM32F407VGT6 (STM32F4 Discovery) и трёхосевой MEMS-акселерометр LIS3DSHTR с интерфейсом I²C, что позволило создать компактное, энергоэффективное и экономически доступное решение для непрерывного контроля вибрационных характеристик оборудования. Конфигурация сенсорного узла и параметры съёма данных обеспечили достаточную частоту дискретизации для регистрации характерных вибрационных компонент, релевантных задачам ранней диагностики.

Ключевым результатом работы является разработка универсального протокола передачи данных через интерфейс UART, обеспечивающего устойчивую и надёжную коммуникацию между сенсорным уровнем и системами верхнего уровня. Предложенный формат телеметрии (маркировка компонент AX/AY/AZ, кадровая структура и контроль целостности) показал достаточную гибкость для интеграции с различными программно-аппаратными платформами, включая системы сбора и обработки данных, SCADA/PLC-уровень и облачные сервисы мониторинга. Это подтверждает возможность использования разработанной архитектуры в распределённых системах промышленного мониторинга без необходимости существенной модификации существующей инфраструктуры и с минимальными требованиями к вычислительным ресурсам на краевом уровне.

Особое внимание в исследовании уделено вопросам повышения качества измерений. Применение методов цифровой обработки сигналов, включая скользящее среднее, медианную фильтрацию и фильтр Калмана, позволило существенно снизить уровень шумов и повысить стабильность регистрируемых сигналов. Анализ полученных временных рядов показал, что фильтрация эффективно подавляет высокочастотные помехи и импульсные выбросы, сохраняя при этом информативную составляющую сигнала, необходимую для диагностики состояния оборудования. Дополнительно оценивались базовые статистические характеристики (дисперсия, межквартильный размах, peak-to-peak), подтвердившие улучшение «читаемости» сигнала после обработки.

Экспериментальная апробация разработанного решения в условиях реального производства (площадка ТОО ОПТИСОРТ) продемонстрировала его работоспособность и практическую применимость. В ходе испытаний было установлено, что система способна фиксировать ранние признаки механических неисправностей, включая разбалансировку и начальные стадии износа подшипников, проявляющиеся в виде локальных аномалий амплитуды и изменения статистических показателей вибрационного сигнала. Выявленные изменения позволяют использовать систему в качестве инструмента ранней диагностики, что является важным фактором повышения надёжности и снижения затрат на обслуживание оборудования.

Важным результатом исследования является подтверждение возможности создания эффективных систем промышленного мониторинга на основе бюджетных аппаратных компонентов. Использование доступных микроконтроллерных платформ и стандартных MEMS-сенсоров позволяет значительно снизить стоимость внедрения по сравнению с традиционными промышленными решениями, сохраняя при этом достаточный уровень точности и надёжности измерений. Это делает пред-

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

ложенную архитектуру перспективной для широкого применения, включая малые и средние предприятия, где внедрение дорогостоящих систем мониторинга часто затруднено.

С практической точки зрения разработанное решение может быть интегрировано в системы предиктивного обслуживания и использовано в рамках построения цифровых двойников оборудования. Потоковые данные о вибрационных характеристиках могут служить основой для формирования диагностических моделей, настройки пороговых правил и последующего перехода к моделям оценки остаточного ресурса. Таким образом, предложенный подход обеспечивает не только мониторинг текущего состояния, но и создаёт предпосылки для перехода к интеллектуальным системам управления жизненным циклом оборудования с поддержкой удалённой аналитики и централизованного хранения данных.

Перспективы дальнейших исследований связаны с развитием как аппаратной, так и программной составляющих системы. В частности, представляется целесообразным расширение набора анализируемых параметров за счёт включения спектральных характеристик (FFT-признаки), куртозиса и RMS-метрик, что позволит повысить чувствительность к различным типам дефектов. Дополнительное повышение информативности может быть достигнуто за счёт интеграции других типов сенсоров, таких как температурные, акустические и токовые датчики, обеспечивающих комплексную оценку состояния оборудования.

Отдельным направлением развития является применение методов машинного обучения для анализа накопленных данных и прогнозирования остаточного ресурса узлов. Использование алгоритмов интеллектуальной обработки позволит перейти от пороговых методов диагностики к более точным и адаптивным моделям, учитывающим нелинейные зависимости и динамику режимов работы. В сочетании с возможностями IIoT это создаёт основу для построения масштабируемых систем интеллектуального технического обслуживания и интеграции с корпоративными системами управления производством.

Таким образом, результаты исследования подтверждают, что предложенная архитектура системы промышленного мониторинга обладает необходимыми характеристиками надёжности, гибкости и экономической эффективности, а также демонстрирует воспроизводимость и адаптивность в реальных условиях эксплуатации, что делает её перспективной для практического внедрения в современных производственных системах.

REFERENCES

Abdeldjalil Benhanifia, Zied Ben Cheikh, Paulo Moura Oliveira, Antonio Valente, José Lima. (2025). Systematic review of predictive maintenance practices in the manufacturing sector // Intelligent Systems with Applications. — Vol. 26. — Article 200501. DOI: https://doi.org/10.1016/j.iswa.2025.200501 (in Eng.).

Abdulhussain S.H., Mahmmod B.M., Alwhelat A., Shehada D., Shihab Z.I., Mohammed H.J., Abdulameer T.H., Alsabah M., Fadel M.H., Ali S.K., Abbood G.H., Asker Z.A. & Hussain A. (2025). A Comprehensive Review of Sensor Technologies in IoT: Technical Aspects, Challenges, and Future Directions. Computers. — Vol. 14. — No. 8. Pp. 342. DOI: https://doi.org/10.3390/computers14080342 (in Eng.).

Albarbar A., Mekid S., Starr A. & Pietruszkiewicz R. (2008). Suitability of MEMS Accelerometers for Condition Monitoring: An experimental study. Sensors. — Vol. 8. — No. 2. Pp. 784–799. DOI: https://doi.org/10.3390/s8020784 (in

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

Eng.).

Behnke I. & Austad H. (2023). Real-Time Performance of Industrial IoT Communication Technologies: A Review. arXiv preprint. – arXiv:2311.08852. DOI: https://doi.org/10.48550/arXiv.2311.08852 (in Eng.).

Chistjakov D.A., Razinkin V.P. (2020). Sintez i realizacija cifrovogo fil'tra dlja obrabotki vibracionnyh signalov {Synthesis and implementation of digital filter for vibration signals processing}. Datchiki i sistemy. — Vol. 1. — No. 243. Pp. 47–51. DOI: https://doi.org/10.25728/datsys.2020.1.8. (in Russ.).

Garcia, J., Rios-Colque, L., Peña, A., & Rojas, L. (2025). Condition Monitoring and Predictive Maintenance in Industrial Equipment: An NLP-Assisted Review of Signal Processing, Hybrid Models, and Implementation Challenges. Applied Sciences. — Vol. 15. — No. 10. Pp. 5465. DOI: https://doi.org/10.3390/app15105465 (in Eng.).

Ishmet'ev E.N., Chistjakov D.V., Panov A.N., Bodrov E.Je., & Vrabel M. (2019). Sistemy vibrozashhity, vibrokontrolja i vibrodiagnostiki promyshlennogo oborudovanija {Vibration protection, vibration monitoring, and vibration diagnostics systems for industrial equipment}. Jelektrotehnicheskie sistemy i kompleksy. — Vol. 1. — No. 42. Pp. 67–73 (in Russ.).

K. Mc Court, X. Mc Court, S. Du and Z. Zeng. (2025). Use Digital Twins to Support Fault Diagnosis from System-Level Condition-Monitoring Data. IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD), Monastir, Tunisia. Pp. 1064-1069. DOI: https://doi.org/10.1109/SSD64182.2025.10989951 (in Eng.).

Kamran Iranshahi, Joshua Brun, Tim Arnold, Thomas Sergi, Ulf Christian Müller. (2025Digital twins: Recent advances and future directions in engineering fields // Intelligent Systems with Applications. — Vol. 26. — Article 200516. DOI: https://doi.org/10.1016/j.iswa.2025.200516 (in Eng.).

Marco Carratù, Vincenzo Gallo, Paolo Sommella, Antonio Pietrosanto, Marcantonio Catelani, Lorenzo Ciani, Gabriele Patrizi, Roberto Singuaroli. (2025). Development of a methodology for MEMS accelerometer health state estimation. Measurement: Sensors. — Vol. 38. — Article 101604. DOI: https://doi.org/10.1016/j.measen.2024.101604 (in Eng.).

Paziewski J., Sieradzki R., Rapinski J., Tomaszewski D., Stepniak K., Geng J., Li G. (2025). Integrating low-cost GNSS and MEMS accelerometer for precise dynamic displacement monitoring. Measurement. — Vol. 242(Part A). — Article 115798. DOI: https://doi.org/10.1016/j.measurement.2024.115798 (in Eng.).

Shonazarov P.M., Holov F.T., Evsjutin O.O., Tursunbadalov U.A. (2019). Primenenie fil'tra Kalmana v zadache tehnicheskoj diagnostiki dvigatelej vnutrennego sgoranija {Application of the kalman filter in the problem of technical diagnostics of internal combustion engines} // Komp'juternye tehnologii, upravlenie, radiojelektronika. — Vol. 19. — No. 1. Pp. 152–159. DOI: https://doi.org/10.14529/ctcr190115 (in Russ.).

STMicroelectronics. (2017). LIS3DSH MEMS digital output motion sensor: high-performance three-axis accelerometer. Datasheet. https://www.st.com/resource/en/datasheet/lis3dsh.pdf (in Eng.).

STMicroelectronics. (2024). STM32F405/407 Reference Manual (RM0090). https://www.st.com/resource/zh/reference_manual/DM00031020.pdf (in Eng.).

Tang, X., Wang, X., Cattley, R., Gu, F., & Ball, A. D. (2018). Energy Harvesting Technologies for Achieving Self-Powered Wireless Sensor Networks in Machine Condition Monitoring: A Review. Sensors. — Vol. 18. — No. 12. Pp. 4113. DOI: https://doi.org/10.3390/s18124113 (in Eng.).

Vlasov A.I., Grigor'ev P.V., Krivoshein A.I. (2018). Model' prediktivnogo obsluzhivanija oborudovanija s primeneniem besprovodnyh sensornyh setej {Model of predictive equipment maintenance with application of wireless touch networks} // Nadezhnost' i kachestvo slozhnyh system. — Vol. 2. — No. 22. Pp. 26–35. DOI: https://doi.org/10.21685/2307-4205-2018-2-4 (in Rus.).

Welch G. & Bishop G. (2006). An Introduction to the Kalman Filter. UNC Technical Report TR 95–041. Chapel Hill: University of North Carolina. https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf (in Eng.).

Yaghoubi S., Binali R., Demirpolat H., Kuntoğlu M., Makhesana M.A., Sayın Kul B. (2024). A Comprehensive Review on Low-Cost MEMS Accelerometers for Vibration Measurement: Types, Novel Designs, Performance Evaluation, and Applications // Journal of Molecular and Engineering Materials. — Vol. 12. — No. 3. Article 2430002. DOI: https://doi.org/10.1142/S225123732430002X (in Eng.).

Zhang C., Li A., Huang J., Zhang J., Huang H., Su Y., Wang L., Liu Y., Chen D., Tian F., Wang Y., Wu Z. (2026). Calibration-free low-drift MEMS piezoelectric accelerometer for high-temperature vibration monitoring. Measurement. — Vol. 257(Part B). — Article 118694. DOI: https://doi.org/10.1016/j.measurement.2025.118694 (in Eng.).

Zhou W. (2025). Applications of MEMS sensors in industrial equipment condition sensing and the Internet of Things // International Scientific Technical and Economic Research. — Vol. 3. — No. 4. Pp. 121–128. DOI: https://doi.org/10.71451/ISTAER2560 (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.2. Number 26 (2026). Pp. 323–334
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.021
IRSTI / ҒТАХР / МРНТИ 004.656

KAZCAUSAL: THE FIRST CORPUS-BASED ANNOTATION OF CAUSAL RELATIONSHIPS IN THE KAZAKH LANGUAGE

R. Taberkhan[1]*, M.A. Sambetbayeva[1], G. Kalman[2]
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]K.K. Ualikhanov Kokshetau University, Kokshetau, Kazakhstan.
E-mail: roman.tn82@gmail.com

Roman Taberkhan — PhD student, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: roman.tn82@gmail.com, https://orcid.org/0000-0002-3375-4947;
Madina Sambetbayeva — PhD, Associate Professor, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0001-9358-1614;
Gulzhamal Kalman — PhD, Associate Professor, Department of Information and Communication Technologies, Sh. Sh. Ualikhanov Kokshetau University, Kokshetau, Kazakhstan
https://orcid.org/0000-0001-8863-9447.

© R. Taberkhan, M.A. Sambetbayeva, G. Kalman

Abstract. In this article, the KAZCausal corpus attempts for the first time to annotate causal relationships in modern Kazakh. The study examines the structural and methodological foundations of the KAZCausal corpus. In addition, the first attempt was made to systematically record causal and resultative relationships in texts based on scientific, journalistic, and Internet information materials, and the results of this study make a significant contribution to the field of natural language processing (NLP), integrating with the field of linguistics. In the course of the study, the process of corpus creation, the annotation scheme, and linguistic criteria will be discussed as prerequisites for creating a corpus applied to syntactic and semantic cause-and-effect configurations in the modern Kazakh language. This article will examine a number of tools, models, and corpora that have been successfully used in research on other languages for contextualization in current NLP experiments. These include the Humor (High-speed Unification Morphology) annotator for morphological processing, the POS tagging model, the Multi-tasl learning and Subdomain adaptation model, Multilingual BERT, the multilingual XLM-R model, as well as the Kazakh Language Corpus-KLC with wide coverage, Penn Discourse Treebank (PDTB),

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

CausBank, and BECause, the main differences and innovative aspects of the KAZCausal corpus aimed at annotating causal relationships were examined, as well as features such as annotation principles, scope of application, and structural nature. In the future, the KAZCausal corpus may become the basis for application in other Turkic-language corpora, effective use of the Kazakh language in computer processing, development of semantic search technologies in information systems, and improvement of NLP models.

Keywords: NLP, text annotator, corpus linguistics, KazCausal, multilingual model, annotation

For citation: R. Taberkhan, M.A. Sambetbayeva, G. Kalman (2026). KazCausal: the first corpus-based annotation of causal relationships in the kazakh language // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 323–334. https://doi.org/10.54309/IJICT.2026.26.2.021. (In Russ.).

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

Acknowledgment: *This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (AP26195165).*

KAZCAUSAL: ҚАЗАҚ ТІЛІНДЕГІ СЕБЕП-САЛДАРЛЫҚ ҚАТЫНАСТАРДЫҢ АЛҒАШҚЫ КОРПУСТЫҚ АННОТАЦИЯСЫ

Р. Таберхан[1]*, М.А. Самбетбаева[1], Г. Қалман[2]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Ш.Ш. Уәлиханов атындағы Көкшетау университеті, Көкшетау, Қазақстан.
E-mail: roman.tn82@gmail.com

Таберхан Роман — PhD докторант, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасы, Астана, Қазақстан
E-mail: roman.tn82@gmail.com, https://orcid.org/0000-0002-3375-4947;
Самбетбаева Мадина — PhD докторы, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0000-0001-9358-1614;
Қалман Гүлжамал — PhD докторы, Ақпараттық-коммуникациялық технологиялар кафедрасының қауымдастырылған профессоры, Ш.Ш. Уәлиханов атындағы Көкшетау университеті, Көкшетау, Қазақстан
https://orcid.org/0000-0001-8863-9447.

© Р. Таберхан, М.А. Самбетбаева, Г. Қалман

Аннотация. Аталған мақалада KAZCausal корпусы қазіргі қазақ тіліндегі себеп-салдарлық байланысты аннотациялауға алғаш рет талпыныс жасап отыр. Зерттеуде KAZCausal корпусын құрастыру, құрылымдық дизайны және әдіснамалық негіздері қарастырылады. Сонымен қатар, ғылыми, публицистикалық, интернеттік-ақпараттық

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

материалдар негізіндегі мәтіндердегі себеп-салдарлық пен нәтижелік қатынастарды жүйелі түрде тіркеудің алғашқы әрекеті жасалды, әрі аталған зерттеу нәтижесі лингвистика саласымен интеграциялана отырып, табиғи тілдерді өңдеу (Natural language processing – NLP) саласына да айтарлықтай үлес қосады. Зерттеу барысында қазіргі қазақ тіліндегі синтаксистік және семантикалық себеп-салдарлық конфигурацияларға қолданылатын корпусты құрудың алғышарттары ретінде корпус құру үдерісіне, аннотация сұлбасына және лингвистикалық критерийлерге тоқталады. Бұл мақалада қазіргі NLP тәжірибелері аясында контексттеу үшін зерттеуде басқа да тілдерде сәтті қолданыс тапқан бірқатар құралдар, моделдер мен корпустар қарастырылады. Оларға морфологиялық өңдеуге арналған Humor (High-speed Unification Morphology) аннотаторы, POS tagging моделі, Multi-tasl learning and Subdomain adaptation моделі, Multilingual BERT, XLM-R көптілді моделі, сонымен қатар кең ауқымды қамтыған қазақ тілді Kazakh Language Corpus – KLC корпусы, Penn Discourse Treebank (PDTB), CausBank және BECause сияқты аннотация қағидаттары, қолдану саласы және құрылымдық сипаты сияқты ерекшеліктеріне шолу жасалып, себеп-салдарлық байланыстарды аннотациялауға бағытталған KAZCausal корпусының негізгі айырмашылықтары мен инновациялық аспектілері қарастырылды. KAZCausal корпусы болашақта өзге де түркі тілдес корпустарға қолдануға, қазақ тілін компьютерлік өңдеу бағытында тиімді пайдалануға, ақпараттық жүйелерде семантикалық іздеу технологияларын дамытуда және NLP моделдерін жетілдіруге негіз етуге болады.

Түйінді сөздер: NLP, мәтінді аннотаторы, корпус лингвистикасы, KazCausal, көптілді модель, аннотациялау

Дәйексөздер үшін: Р. Таберхан, М.А. Самбетбаева, Г. Қалман (2026). KazCausal: қазақ тіліндегі себеп-салдарлық қатынастардың алғашқы корпустық аннотациясы // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 323–334. https://doi.org/10.54309/IJICT.2026.26.2.021. (Қаз. тіл.).

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

Алғыс: *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (AP26195165).*

KAZCAUSAL: ПЕРВАЯ КОРПУСНАЯ АННОТАЦИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ НА КАЗАХСКОМ ЯЗЫКЕ

Р. Таберхан[1]*, М.А. Самбетбаева[1], Г. Калман[2]

[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;

[2]Университет имени Ш.Ш. Уалиханова, Кокшетау, Казахстан.

E-mail: roman.tn82@gmail.com

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

Таберхан Роман — PhD докторант, Евразийский национальный университет имени Л.Н. Гумилева, кафедра информационных систем, Астана, Казахстан
E-mail: roman.tn82@gmail.com, https://orcid.org/0000-0002-3375-4947;
Самбетбаева Мадина— Доктор PhD, Ассоциированный профессор кафедры информационных систем, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0001-9358-1614;
Қалман Гүлжамал — Доктор PhD, Ассоциированный профессор кафедры информационных и коммуникационных технологий, Кокшетауский университет имени Ш. Уалиханова, Кокшетау, Казахстан
https://orcid.org/0000-0001-8863-9447.

© Р. Таберхан, М.А. Самбетбаева, Г. Қалман

Аннотация. В данной статье корпус KAZCausal впервые пытается аннотировать причинно-следственную связь на современном казахском языке. В исследовании рассматриваются конструктивные и методологические основы сборки корпуса KAZCausal. Кроме того, была предпринята первая попытка систематической регистрации причинно-следственных и результирующих отношений в текстах на основе научных, публицистических, интернет-информационных материалов, и результаты данного исследования вносят значительный вклад в область обработки естественных языков (Natural language processing – NLP), интегрируясь с областью лингвистики. В ходе исследования в качестве предпосылок для создания корпуса, применяемых к синтаксическим и семантическим причинно-следственным конфигурациям в современном казахском языке, будут затронуты процесс создания корпуса, схема аннотации и лингвистические критерии. В этой статье будет рассмотрен ряд инструментов, моделей и корпусов, которые успешно использовались в исследованиях на других языках для контекстуализации в рамках текущих экспериментов NLP. К ним относятся аннотатор Humor (High-speed Unification Morphology) для морфологической обработки, модель POS tagging, модель Multi-tasl learning and Subdomain adaptation, Multilingual BERT, многоязычная модель XLM-R, а также корпус казахоязычного Kazakh Language Corpus-KLC с широким охватом, Penn Discourse Treebank (PDTB), CausBank и BECause были рассмотрены основные различия и инновационные аспекты корпуса KAZCausal, направленные на аннотирование причинно-следственных связей, а также такие особенности, как принципы аннотации, область применения и структурный характер. Корпус KAZCausal в будущем может стать основой для применения в других тюркоязычных корпусах, эффективного использования казахского языка в направлении компьютерной обработки, развития технологий семантического поиска в информационных системах и совершенствования моделей NLP.

Ключевые слова: NLP, аннотатор текста, корпусная лингвистика, KazCausal, многоязычная модель, аннотация

Для цитирования: Р. Таберхан, М.А. Самбетбаева, Г. Қалман (2026). KazCausal:первая корпусная аннотация причинно-следственных связей на казахском

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

языке // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 323–334. https://doi.org/10.54309/IJICT.2026.26.2.021. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность: *Данное исследование было профинансировано Комитетом науки Министерства науки и высшего образования Республики Казахстан (AP26195165).*

Кіріспе.

Табиғи тілдерді өңдеу (Natural language processing-NLP)-технология саласының қарқынды дамуы арқылы адам тілін модельдеу, олардың өзара қатынасу мүмкіндіктерін барынша кеңейткені белгілі. Мұндай жетістіктер, әсіресе, модельдері күрделі болып келетін деректері толық жеткілікті кең таралған тілдер үшін өте пайдалы болды. Мәселен, арнайы мәтіндік корпустарды жинау, оларды топтарға жинақтау, морфологиялық және синтаксистік талдаулар жасауға талпыныс Венгрияда Humor (High-speed Unification Morphology) аннотаторынан басталса (Proszeky et al., 1999), Орталық және Шығыс Еуропа тілдеріне зерттеу жасалынған морфологиялық-синтаксистік кодтау сызбасы MSD Copernicus 106 жобасы (Erjavec et al., 1998) жүзеге асырылды. Аталған жоба аясында Дж. Оруэллдің «1984» атты романындағы параллель аннотацияланған корпус пайда болды. Дегенмен, аталған корпустың бірқатар кемшіліктерінің ішінде ең бастысы көлемінің шағындығы еді, сондықтан ол іргелі зерттеулер жасауға мүмкіндік бермеді. Осы олқылықтың орнын толтыру мақсатымен, машиналық оқыту қолданбалары үшін сөз табын аннотациялауға арналған POS tagging (Alexin et al., 2003) көлемі кең корпусы жүзеге асырылды.

Корпустар арасындағы түрлі эмоцияларды тану ерекшеліктеріне негізделген көпміндетті оқыту мен ішкі домендік бейімдеу (Multi-tasl learning and Subdomain adaptation) моделі эмоцияны білдіретін конструкцияларды тану үшін бөгет жасайтын факторларды азайту мақсатын көздейді (Fu et al., 2023). Бұл моделдің басқа да кросс-корпустық сөйлеу эмоциясын тануға бағытталған зерттеулерден айырмашылығы орташа реколл (recall) көрсеткішінің жоғары екендігі болып табылады.

Алайда бұл айтылғандардың барлығы табиғи тілдерді өңдеу (NLP) барысында негізінен ресурсы мол, қолданылу аясы кең, кең таралған тілдерге негізделген зерттеулер еді. Мұндай моделдерді аз қолданыстағы, ресурсы аз тілдердің мәселесін шешу үшін қолданылатын инновациялық стратегиялар пайда бола бастады. Осындай бейімдеу тәсілін-трансферлік оқыту деп атап, Multilingual BERT, XLM-R сияқты моделі жүзеге асырылды (Pakray et al., 2024).

Ал қазіргі қазақ тілінің корпусын жасауға талпынған зерттеуге сәйкес (Kazakh Language Corpus-KLC), аталған зерттеу жергілікті ізденушілер қауымдастығы жүзеге асырған қазақ тілінің корпусын құрастыруға бағытталған алғашқы зерттеулердің бірі болып табылады (Makhambetov et al., 2013). Оның басты ерекшелігі–135 миллионнан астам сөзді қамтыған ауқымы өте көлемді екендігі, сонымен қатар жанрлық-стильдік жағынан бес саланы қамтитындығын атауға болады. Корпустың басты негізінен басқа аннотацияланған жеке морфологиялық, синтаксистік және құрылым-

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

дық белгілеріне сай бөлімдер және сөйлеу тілінің аннотацияланған деректеріне арналған қосымша корпусы бар. Дегенмен, қазақ тіліндегі белгілі бір конструкцияларды аннотациялайтын модель негізіндегі корпустың болмауы зерттеу саласындағы үлкен олқылықтардың бірі болып табылады.

Әдістер мен материалдар.

Қазақ тіліндегі себеп-салдарлық конструкцияларды аннотациялауға бағытталған алғашқы талпыныстардың бірі болып табылатын KAZCausal корпусын құрастыру мақсатында қазақ тіліндегі ғылыми, публицистикалық және ресми-ақпараттық және интернет мәтіндері секілді стильдік-жанрлық сипаттағы мәтіндер қолданылды. Күн сайын деректер базасы барынша толығуда және қазіргі уақытта деректердің бастапқы көлемі 850 бірлікті құрайды. Корпусқа алынған материалдар pdf, docx, html форматтарынан алдын-ала өңдеу тәсілі арқылы өзара біріздендіріліп, корпус форматына келтіріледі.

Зерттеу барысында басты тәсіл-аннотациялау бірліктері, яғни сөйлем ішіндегі себеп-салдарлық қатынастарды аннотациялау жүзеге асырылады. Оған себеп компонентті (CAUSE), салдар компонентті (EFFECT), себеп-салдарлық маркерлер (CONNECTIVES) және себеп-салдарлық типтер ретінде логикалық, уақыттық, прагматикалық, шартты-себепті, мақсаттық және т.б. қарастырылды.

Келесі мәтіндерді өңдеу кезеңі тәсілі кезінде алдымен токенизация алгоритмі, ережелік модель, морфологиялық сараптама мен қателерді автоматты түзету сараптамасы аннотациялау тарапынан қолмен тексеріліп, нақтылану жүзеге асырылады.

Ал аннотациялау кезеңінде арнайы аннотациялау нұсқаулығы (annotation guidelines) даярланды. Аталған нұсқаулыққа сәйкес, қазақ тіліндегі себеп-салдарлық қатынастарды анықтау үшін алғашқы аннотация, келісім көрсеткішін есептеу, соңғы нұсқаны қалыптастыру үдерістері жасалады. Корпус Label Studio платформасында жүзеге асырылса, кіріс-шығыс форматтары JSON, XML негізінде және Python бағдарламалау тілі қолданылды.

Корпусқа жинақталған деректерді талдау үшін себеп-салдарлық маркерлердің жиілігін есептеу, себеп пен салдардың синтаксистік үлгілерін анықтау, себеп-салдарлық құрылымдар типологиясын жіктеу мен статистикалық талдау әдістері назарға алынды.

Нәтижелер және оларды талқылау.

Себеп-салдарлық байланыстар табиғи тілдің мағыналық құрылымдарында маңызды орын алады және табиғи тілдерді өңдеу (NLP) саласында мәліметтерді талдаудың ең негізгі компоненттерінің бірі болып табылады. Олар мәтіндегі оқиғалар арасындағы логикалық байланыстарды көрсетіп, аргументациясының, түсіндірудің, дәлелдеу мен пайымдаудың негізін құрайды. Тілдік коммуникацияда себептің немесе салдардың нақты берілуі айтушының көзқарасын, бағалауын, ниетін, эмоциясын жеткізеді және дискурстың прагматикалық бағытын қалыптастырады. Сондықтан себеп-салдарлық байланыстарды жүйелі түрде зерттеу тек жалпы тіл білімінің ғана емес, сонымен қатар қолданбалы мәтіндік және когнитивтік лингвистиканы, қазіргі заманғы алгоритмдік талдаудың өзекті мәселелерінің бірі болып келеді.

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

Қазіргі кезде ағылшын, неміс, қытай және араб тілдеріндегі себеп-салдарлық байланыстарды таңбалау бойынша көптеген корпустар жасалған. Әсіресе, P*enn Discourse Treebank* (Prasad et al., 2008), *CausBank* (Li et al., 2020), *VerbNet* (Schuler, 2005), *FrameNet* (Vieu et al., 2016), *FrameNet II* (Ruppenhofer et al., 2016) және *BECause* (Dunietz et al., 2017) сияқты ресурстарда себер-салдарлық қатынастар лексикалық, синтаксистік және дискурстық деңгейлерде белгіленіп, үлкен көлемді талдаулар жасалған. Өзге тілдердегі каузалды конструкциялар ғылыми нысанның түрлі аспектілерінде қарастырылып келді, мәселен, себеп-салдарлық байланыстарды анықтауға арналған әлсіз бақылау негізінде клиникалық мәтіндерді ұсыну әдістерін анықтау (Weng et al., 2020), себеп-салдар байланыстарын анықтауда трансфер әдісін қолдану, құрылымы мен деректеріне қатысты эмпирикалық зерттеулер (Anuyah et al., 2025) жасалды. Осы жобалар табиғи тілдерді өңдеу (NLP) саласында ақпараттарды іздеу, машиналық аударма, мәтіндерді түсіндіру, себептілікті автоматты түрде анықтау сияқты көптеген міндеттердің тиімді орындалуына өте жақсы мүмкіндік береді. Дегенмен, морфологиясы күрделі агглютинативті құрылымдағы тілдер үшін осындай ресурстардың жеткіліксіздігі көрінеді. Осы кездегі қолданыстағы халықаралық схемалар агглютинативті тілдерге тән сөйлемдер құрылысы, етістік сөзжасамына толы жүйесі, есімше және көсемше формалары арқылы жасалатын полипредикативті конструкциялардың ерекшеліктерін толық қамти алмайды. Қазақ тілінде себеп-салдарлық мән *–ғандықтан, –ған соң, –ып, –май, сондықтан, өйткені, себебі, неге десең* деген сияқты көптеген морфологиялық және аналитикалық құралдар арқылы беріледі. Бірақ осы құралдардың жүйелі сипаттамасын, олардың семантикалық реңктерін және тексгегі қызметін көрсететін арнайы корпустық ресурс қазіргі күнге дейін жасалмаған. Дегенмен, аталған салаға қатысты алғашқы зерттеулерге авторлар тарапынан талпыныс жасалды (Taberkhan et al., 2025). Осы кемшіліктің орынын толтыру үшін ұсынылып отырған KazCausal жобасы қазақ тіліндегі себеп-салдарлық байланыстарды жүйелі түрде аннотациялаудың алғашқы тәжірибелерін ұсынып отырмыз. Жоба барысында себеп пен салдарды көрсететін морфологиялық, аналитико-синтетикалық және аналитикалық құрылымдары анықталып, олардың семантикалық ерекшеліктері анықталды. Корпус көлеміне енгізілген сөйлемдер ғылыми, публицистикалық, оқу-әдістемелік және саяси мәтіндер моделдерінен алынған, бұлар қазақ тіліндегі себеп-салдарлық құрылымдардың жан-жақты қамтылуына мүмкіндік береді.

Корпустық аннотация әдісі

KazCausal-қазақ тіліне арнайы бейімделіп жасалған, себеп-салдарлық мағынаны білдіретін құрылымдарды олардың морфологиялық ерекшеліктері мен синтаксистік ұйымдасуын ескере отырып белгілеуге арналған алғашқы аннотациялық схема болып табылады.

Ұсынылған суретте *Label Studio* платформасы арқылы жүзеге асқан каузалды қатынастарды таңбалау үдерісінің нақты көрінісі берілген. Бұл интерфейс *KazCausal* корпусының бастапқы белгілеу кезеңдерінде қолданылған құралдардың бірі ретінде сипатталады.

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



Сур. 1. KazCausal корпусындағы себеп-салдарлық қатынастарды Label Studio платформасында аннотациялау интерфейсі

Аннотациялау барысында мәтіндегі себеп-салдарлық құрылымдар семантикалық және синтаксистік деңгейде сегменттеліп, арнайы әзірленген схема негізінде типологиялық классификациядан өтті. Көрсетілген мысалда *«Оқушылар сабаққа кеш келгендіктен, мұғалім сабақты кейінге шегерді»* сөйлеміндегі себептік қатынас анықталып, *«келгендіктен»* фрагменті *SYNTHETIC_CAUSE* санатына жатқызылған. Ал *«сабақты кейінге шегерді»* бөлімі салдар компоненті ретінде *EFFECT* белгісін алған. Осы екі элементтің арасындағы байланысты *leads_to* реляциясы көрсетеді.

Платформа мәтіндік аннотациямен қатар, оң жақ панелде белгіленген объектілер мен байланыстар туралы метадеректерлерді автоматты түрде сақтап отырады. Бұл деректердің құрылымданған түрде жинақталуына және кейінгі кезеңдерде талдау, модель дайындау немесе автоматтандырылған себеп-салдарлық ақпаратты өндіру жүйелерін әзірлеуде қолданылуына мүмкіндік береді.

KazCausal жобасы шеңберінде аннотациялау Label Studio негізінде жүргізіліп, көпдеңгейлі таңбалау мүмкіндіктерін тиімді пайдалануға жол ашты. Аннотацияның негізгі мақсаты-қазақ тіліндегі каузалдық қатынастардың морфологиялық, синтаксистік және семантикалық ерекшеліктерін формалды түрде белгілеу және жүйелеу болды. Осы тәсіл арқылы әрбір каузалды фрагмент анық шекаралармен белгіленіп, олардың арасындағы байланыстар жеке реляциялық жазбалар түрінде тіркелді.

Белгілеу барысында каузалды құрылымдар үш негізгі үлгіге бөлінді: синтетикалық, аналитико-синтетикалық және аналитикалық типтер. Бұл жіктеу қазақ тілінің агглютинативті жүйесіне тән грамматикалық көрсеткіштер негізінде жүзеге асты, мысалы: *-ғандықтан, -ған соң, -ғаны үшін, -сол үшін* және т.б. Әрбір тип өз ішінде

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

субкатегорияларға бөлініп, олардың қызметтік ерекшеліктері нақты таңбаланды.

Аннотациялау барысында себеп (CAUSE), салдар (EFFECT), әрекет субъектісі (SUBJECT_CAUSE) және салдар субъектісі (SUBJECT_EFFECT) сияқты компоненттер жеке сегменттер ретінде белгіленді. Сонымен қатар leads_to, has_marker, has_consequence_marker, subject_of сияқты реляциялар арқылы олардың семантикалық байланыстары тіркелді. Бұл тәсіл каузалдық қатынастардың құрылымдық сипатымен қатар, мағыналық тәуелділігін де көрсетуге мүмкіндік берді.

Аннотациялау кезеңі *KazCausal* корпусының қалыптасуындағы негізгі әдістемелік қадамдардың бірі болып табылады. Формалды таңбалау нәтижесінде қазақ тіліндегі себеп-салдарлық құрылымдардың көріністері жүйеленіп, олардың негізінде төмен ресурсты тілдер үшін автоматтандырылған себеп-салдарлық ақпаратты өңдеу модельдерін әзірлеуге практикалық база жасалды.



Сур. 2. KazCausal корпусын аннотациялау үдерісінің жұмыс схемасы.

Жоғарыда келтірілген суретте (Сурет 2) корпусқа енгізілетін мәтінді таңдаудан бастап, сөйлемдерді сегменттеуге, каузалды маркерлерді анықтауға, типологиялық жіктеуге, реляциялық байланыстарды белгілеуге және аннотацияланған деректерді корпусқа енгізуге дейінгі кезеңдер көрсетілген. Бұл workflow аннотациялау әдіснамасының жүйелі құрылымын және деректерді формалды өңдеу

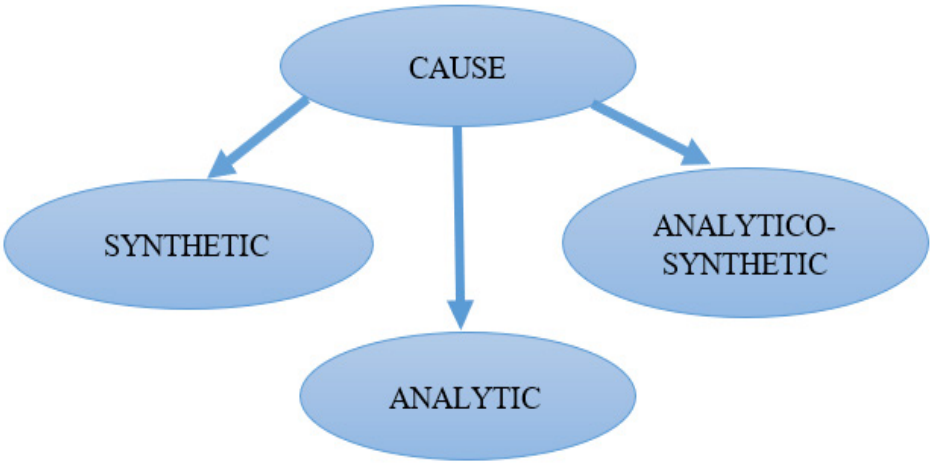
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

логикасын айқындайды.



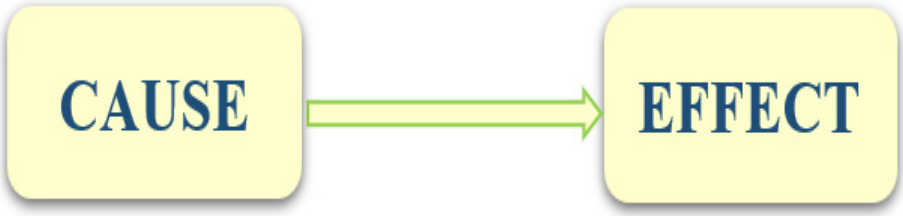
Сур. 3. KazCausal корпусындағы каузалды құрылымдардың типтік үлесі.

Сурет 3-ке сәйкес, KazCausal корпусындағы каузалды құрылымдардың типтік үлесі төмендегідей көрсеткіштерге, атап айтқанда синтетикалық (45%), аналитико-синтетикалық (32%) және аналитикалық (23%) себеп-салдарлық құрылымдардың үлесін көрсетеді. Бұл үлестірім қазақ тіліндегі каузалды мәннің басым түрде морфологиялық тәсілдер арқылы берілуін айқындайды және тілдің агглютинативті табиғатын растайды.



Сур. 4. KazCausal корпусындағы каузалды құрылымдардың таксономиялық жіктелуі.

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



Сур. 5. KazCausal корпусындағы себеп-салдарлық реляциялық байланыс моделі.

Схема каузалды қатынастардың негізгі категориясы (Сурет 4)-CAUSE-және оның үш құрылымдық типке (SYNTHETIC, ANALYTICO-SYNTHETIC, ANALYTIC) бөлінуін көрсетеді. Әрбір тип морфологиялық және синтаксистік көрсеткіштерге негізделген субкатегориялар арқылы сипатталады, бұл аннотациялық схеманың иерархиялық құрылымын ашады.

Себептік компонент (CAUSE) пен салдарлық компонент (EFFECT) арасындағы семантикалық тәуелділік leads_to реляциясы арқылы көрсетілген (Сурет 5). Бұл модель каузалды құрылымдарды формалды түрде байланыстырудың негізін құрайды және кейінгі автоматтандырылған себеп-салдарлық ақпарат өндіру жүйелеріне қолдануға мүмкіндік береді.

Қорытынды.

Себеп-салдарлық конструкцияларды аннотациялау, оларды автоматты түроде табу міндеті толық шешімін таппаған ғылыми зерттеу нысаны болып табылады. Себебі, аталған ғылыми бағыт ағылшын тіліндегі материалдар негізінде кеңінен зерттелгенімен, өзге де көптеген тілдер үшін, соның ішінде агглютинативті тілдер қатарына жататын, біріздендірілген аннотацияланған деректер житынтығы жоқ қазақ тілі үшін мүлдем қарастырылмаған сала деуге болады.

Мәтіндердегі себеп-салдарлық конструкцияларды анықтау-себептік қатынастарды автоматты түрде табуға бағытталған маңызды ақпараттық анықтағыш міндетін атқарады. Өйткені, қолданыстағы біртілді деректер жиынтығы аннотациялау нұсқаулығын қолдана отырып, олардың толық біртұтас корпуста біріздендірілуіне мүмкіндік береді. Сондықтан кең ауқымды, стандартталған жаңа деректер жиынтығынан құралған корпус өз кезегінде көптеген пәнаралық (ақпараттық жүйелер, лингвистика) міндеттердің шешілуіне мүмкіндік береді.

Себеп-салдарлық байланыстарды таңбалау және оларды формалды түрде көрсету әлемдік лингвистикада, дискурс талдауында және табиғи тілдерді өңдеу (NLP) саласында іргелі зерттеуді қажет ететін ғылыми бағыттардың бірі болып табылады. Себеп-салдарлық қатынастар мәтіндердегі оқиғалар арасындағы логикалық тәуелділіктерді көрсетіп, түсіндіру, дәлелдеу, мотивация, санкция, және эмоциялық реакция сияқты күрделі мағыналық байланыстарды анықтауға мүмкіндік береді. Алайда, бұл құбылыстарды жүйелі түрде сипаттау әр тілдің морфологиялық, синтаксистік және семантикалық ерекшеліктеріне қарай әртүрлі тәсілдерді талап етеді.

KazCausal корпусы қазақ тіліндегі себеп-салдарлық байланыстарды белгілеу-

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

дің формалды моделін жасауға, әрі олардың типологиясын нақтылауға және морфологиялық-семантикалық табиғатын сипаттауға бағытталған. Жоба нәтижелері қазақ тілін компьютерлік өңдеу бағытында өте пайдалы болмақ: мәтіндерді автоматты талдау, себеп-салдарлық қатынастарды автоматты түрде анықтау, ақпараттық жүйелерде семантикалық іздеу технологияларын дамыту және қазақ тіліне бейімделген NLP моделдерін жетілдіруге негіз бола алады.

REFERENCES

Alexin, Z. (2003). Manually annotated Hungarian corpus // Department of Informatics, University of Szeged. Available at: http://www.inf.u-szeged.hu/~alexin/corpora. [in Eng.].

Anuyah, S., Vanschaik, J., Jain, P., Lehman, S., & Chakraborty, S. (2025). An empirical study of causal relation extraction transfer: Design and data // arXiv. Available at: https://arxiv.org/abs/2503.06076. [in Eng].

Dunietz J., Levin L.S., Carbonell J.G. (2017). The BECauSE Corpus 2.0: Annotating causality and overlapping relations // Proc. 11th Linguistic Annotation Workshop (LAW). – Association for Computational Linguistics. Pp. 95–104. DOI: https://doi.org/10.18653/v1/W17-0812. [in Eng].

Erjavec T., Monachini M. (Eds.) (1997). Specification and notation for lexicon encoding (MULTEXT-EAST Project) // Copernicus Project 106. Available at: http://nl.ijs.si/ME. [in Eng].

Fu H., Zhuang Z., Wang Y., Huang C., Duan W. (2023). Cross-corpus speech emotion recognition based on multi-task learning and subdomain adaptation // Entropy. — Vol. 25. — No. 1. Pp. 124. DOI: https://doi.org/10.3390/e25010124 – [in Eng.].

Li Z., Ding X., Liu T., Hu J.E., Van Durme B. (2020). Guided generation of cause and effect // Proc. IJCAI-20. DOI: https://doi.org/10.24963/ijcai.2020/447. [in Eng].

Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. (2013). Assembling the Kazakh Language Corpus // Nazarbayev University Research and Innovation System. Available at: https://nur.nu.edu.kz. [in Eng.].

Pakray P., Gelbukh A., Bandyopadhyay S. (2025). Natural language processing applications for low-resource languages // Natural Language Processing. — Vol. 31. — Pp. 183–197. DOI: https://doi.org/10.1017/nlp.2024.33. [in Eng.].

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B. (2008). The Penn Discourse TreeBank 2.0 // Proc. LREC'08. – European Language Resources Association. Available at: https://aclanthology.org/L08-1093/. [in Eng.].

Prószéky G., Kis B. (1999). A unification – based approach to morpho – syntactic parsing of agglutinative languages//Proc. 37th Annual Meeting of ACL. Pp. 261–268. DOI: https://doi.org/10.3115/1034678.1034720. [in Eng.].

Ruppenhofer J., Ellsworth M., Petruck M.R.L., Johnson C.R., Baker C.F., Scheffczyk J. (2016). FrameNet II: Extended theory and practice // International Computer Science Institute. Available at: https://framenet.icsi.berkeley.edu. [in Eng.].

Schuler K.K. (2005). VerbNet: A broad coverage, comprehensive verb lexicon // University of Pennsylvania, PhD dissertation. Available at: https://verbs.colorado.edu/verbnet/ – [in Eng.].

Taberkhan R., Sambetbayeva M. (2025). KazCausal: Manual annotation of causal constructions in the Kazakh language // Proc. 10th Int. Conf. on Computer Science and Engineering (UBMK). Pp. 463–467. – DOI: https://doi.org/10.1109/UBMK67458.2025.11206832. [in Eng.].

Vieu L., Muller P., Candito M., Djemaa M. (2016). A General Framework for the Annotation of Causality Based on FrameNet // Proc. LREC'16. – European Language Resources Association (ELRA). Pp. 3807–3813. Available at: https://aclanthology.org/L16-1603/. [in Eng.].

Weng W.H., Lin Y.W., Szolovits P., Elhadad N. (2020). Representation learning of clinical texts with weak supervision for causal relation extraction // Journal of Biomedical Informatics. — Vol. 108. — Article 103491. DOI: https://doi.org/10.1016/j.jbi.2020.103491. [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.2. Number 26 (2026). Pp. 335–350
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.022
IRSTI / ҒТАХР / МРНТИ 50.33.03
UDK 004.89

DIVIDING DEVICES WITH PRELIMINARY PREPARATION OF MULTIPLES OF THE DIVISOR

S. Tynymbayev[1], S.E. Mamanova[1]*, R. Berdybayev[2], Zh.E. Temirbekova[3], T. Chinibayeva[1]

[1]International Information Technology University, Almaty, Kazakhstan;
[2]Almaty University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty, Kazakhstan;
[3]Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: s.mamanova@iitu.edu.kz

Sakhybay Tynymbayev — PhD, Professor of the Department «Computer Engineering», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-9326-9476;
Symbat E. Mamanova — PhD student, senior lecturer of the Department «Computer Engineering», International Information Technology University, Almaty, Kazakhstan
E-mail: s.mamanova@iitu.edu.kz, https://orcid.org/0009-0001-7277-5492;
Rat Berbydayev — PhD, Associate Professor of the Department of Cybersecurity of the Institute of Automation and Information Technologies, Almaty University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty, Kazakhstan
https://orcid.org/0000-0002-8341-9645;
Zhanerke E. Temirbekova — PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0000-0003-3909-0210;
Tolganay Chinibayeva— PhD, Associate Professor of the Department «Computer Engineering», International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-2657-3697.

© S. Tynymbayev, S.E. Mamanova, R. Berdybayev, Zh.E. Temirbekova, T. Chinibayeva

Abstract. The paper presents a division device for integers that generates a remainder and two quotient digits per clock cycle. To enable this, both forward and inverse codes of values that are multiples of the divisor (B, 2B, and 3B) are precomputed and stored in a dedicated quotient register block. During each iteration, the previous partial remainder is left-shifted by two bits (i.e.,) and compared against the stored multiples. The closest match

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

is selected and subtracted from , yielding the next partial remainder . Based on the selected multiple, the corresponding two digits of the quotient are also determined. These operations are carried out within the Partial Remainder and Quotient Bit Generation Block (GPRaQB). The proposed architecture enables implementation of the divider using sequential, matrix-based, or pipelined circuit structures. This study focuses on dividers implemented using sequential circuits. The correctness of the proposed method is validated both through a worked example and simulation on Field-Programmable Gate Arrays (FPGAs).

Keywords: multiples of the divisor, remainder shaper, quotient digit shaper, sequential divisor

For citation: S. Tynymbayev, S.E. Mamanova, R. Berdybayev, Zh.E. Temirbekova, T. Chinibayeva (2026). Dividing devices with preliminary preparation of multiples of the divisor // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 335–350. https://doi.org/10.54309/IJICT.2026.26.2.022. (In Eng.).

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

БӨЛГІШТІҢ ЕСЕЛІ МӘНДЕРІН АЛДЫН АЛА ДАЙЫНДАУМЕН ЖҮЗЕГЕ АСЫРЫЛАТЫН БӨЛУ ҚҰРЫЛҒЫЛАРЫ

С. Тынымбаев[1], С.Е. Маманова[1]*, Р. Бердібаев[2], Ж.Е.Темірбекова[3], Т. Чинибаева[1]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;

[2]Ғұмарбек Дәукеев атындағы Алматы энергетика және телекоммуникациялар университеті, Алматы, Қазақстан;

[3]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.

E-mail: s.mamanova@iitu.edu.kz

Тынымбаев Сахыбай — PhD, «Компьютерлік инженерия» кафедрасының Профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан https://orcid.org/0000-0002-9326-9476;

Маманова Сымбат — PhD докторант, «Компьютерлік инженерия» кафедрасының сениор-лекторы, Халықаралық Ақпараттық Технологиялар университеті, Алматы, Қазақстан

E-mail: s.mamanova@iitu.edu, https://orcid.org/ 0009-0001-7277-5492;

Бердібаев Рат — PhD, Ғұмарбек Дәукеев атындағы Алматы энергетика және байланыс университетінің Автоматтандыру және ақпараттық технологиялар институтының Киберқауіпсіздік кафедрасының қауымдастырылған профессоры, Алматы, Қазақстан https://orcid.org/0000-0002-8341-9645

Темірбекова Жанерке — PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

https://orcid.org/0000-0003-3909-0210;

Чинибаева Толғанай — PhD, «Компьютерлік инженерия» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті,

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

Алматы, Қазақстан
https://orcid.org/0000-0002-2657-3697.

© С. Тынымбаев, С.Е. Маманова, Р. Бердібаев, Ж.Е. Темірбекова, Т. Чинибаева

Аннотация. Мақалада бүтін сандарға арналған бөлгіш құрылғысы ұсынылады, ол әр сағаттық циклде қалдық пен бөліндінің екі цифрын қалыптастырады. Ол үшін бөлгіштің еселі мәндерінің (B, 2B, 3B) тура және кері кодтары алдын ала есептеліп, арнайы регистрге сақталады. Әр итерацияда алдыңғы қалдық екі разрядқа солға ығыстырылады () және алдын ала есептелген мәндермен салыстырылады. Ең жақын мән таңдалып, мәнінен алынып тасталады да, жаңа қалдық есептеледі. Таңдалған мәнге сәйкес бөліндінің екі цифры анықталады. Бұл операциялар қалдық пен бөлінді цифрларын қалыптастыру блогында (GPRaQB) жүзеге асырылады. Ұсынылған архитектура бөлгішті тізбекті, матрицалық немесе конвейерлік схемаларда жүзеге асыруға мүмкіндік береді. Бұл жұмыста тізбекті схемалар негізінде жүзеге асырылған бөлгіштер қарастырылады. Ұсынылған әдістің дұрыстығы мысал арқылы және FPGA платформасында модельдеу арқылы тексерілді.

Түйінді сөздер: бөлгіштің еселі мәндері, қалдықты қалыптастырушы, бөлінді цифрларын қалыптастырушы, тізбекті бөлгіш

Дәйексөздер үшін: С. Тынымбаев, С.Е. Маманова, Р. Бердібаев, Ж.Е. Темірбекова, Т. Чинибаева (2026). Бөлгіштің еселі мәндерін алдын ала дайындалумен жүзеге асырылатын бөлу құрылғылары // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 26. Б. 335–350. https://doi.org/10.54309/IJICT.2026.26.2.022. (Ағыл. тіл.).

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

УСТРОЙСТВА ДЕЛЕНИЯ ЧИСЕЛ С ПРЕДВАРИТЕЛЬНОЙ ПОДГОТОВКОЙ КРАТНЫХ ДЕЛИТЕЛЮ

С. Тынымбаев[1], С.Е. Маманова[1]*, Р. Бердибаев[2], Ж.Е.Темирбекова[3], Т. Чинибаева[1]
[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Алматинский университет энергетики и телекоммуникаций имени Гумарбека Даукеева, Алматы, Казахстан;
[3]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан.
E-mail: s.mamanova@iitu.edu.kz

Тынымбаев Сахыбай — PhD, профессор кафедры «Компьютерная инженерия», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-9326-9476;
Маманова Сымбат — докторант PhD, сениор-лектор кафедры «Компьютерная инженерия», Международный университет информационных технологий, Алматы,

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

Казахстан
E-mail: s.mamanova@iitu.edu.kz, https://orcid.org/0009-0001-7277-5492;
Бердибаев Рат — PhD, Ассоциированный профессор кафедры Кибербезопасности Института автоматизации и информационных технологий Алматинского университета энергетики и телекоммуникаций имени Гумарбека Даукеева, Алматы, Казахстан
https://orcid.org/0000-0002-8341-9645;
Темирбекова Жанерке — PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
https://orcid.org/0000-0003-3909-0210;
Чинибаева Толганай — PhD, Ассоциированный профессор кафедры «Компьютерная инженерия», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-2657-3697.

© С. Тынымбаев, С.Е. Маманова, Р. Бердибаев, Ж.Е. Темирбекова, Т. Чинибаева

Аннотация. Рассматривается устройство деления целых чисел, где в каждом цикле деления формируется остаток и два разряда частного. Для этого в блоке регистра частного предварительно подготавливаются прямые и обратные коды чисел кратные к делителю. Для вычисления текущих остатков и формирования разрядов частного предыдущего остатка сдвигаются на два разряда влево (4), которые сравниваются со всеми значениями кратного 3B, 2B и B и выделяется одно из них. Затем путем вычитания этого числа из 4 вычисляется частичный остаток . По выделенному остатку так же формируются разряды остатка . Вычисления частичного остатка и разрядов частного осуществляется в блоке формирования частичного остатка и разрядов частного (ФЧОиРЧ). На основе ФЧОиРЧ можно построить делители на последовательных и матричных схемах и на схемах с конвейерной структурой. В данной работе рассматриваются делители на последовательных схемах. Проверка правильностифункционирования этих устройств осуществляется на конкретном примере и путем моделирования их на программируемых логических матрицах (ПЛИС).

Ключевые слова: числа кратные делителю, формирователь остатков, формирователь разрядов частного, делитель последовательного действия

Для цитирования: С. Тынымбаев, С.Е. Маманова, Р. Бердибаев, Ж.Е. Темирбекова, Т. Чинибаева (2026). Устройства деления чисел с предварительной подготовкой кратных делителю // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 335–350. https://doi.org/10.54309/IJICT.2026.26.2.022. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Integer division remains a fundamental operation in digital computing, particularly in applications involving digital arithmetic, signal processing, and embedded system design.

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

As the demand for high-throughput and low-latency computational systems increases, there is a growing need for efficient division algorithms that can be readily implemented in hardware. Conventional division methods, while accurate, often incur significant latency and hardware overhead, especially when executed iteratively with single-bit quotient generation per cycle.

This study proposes a novel hardware-oriented approach to integer division that produces two quotient digits per clock cycle, significantly improving computational efficiency. The technique relies on the precomputation of both forward and inverse codes for values that are multiples of the divisor (namely, *B, 2B,* and *3B*), which are stored in a dedicated register block. During each division cycle, the previous partial remainder R_{i-1} is shifted left by two bits (i.e., R_{i-1}) and compared against the precomputed multiples. The closest matching multiple is selected and subtracted from the shifted remainder to compute the new partial remainder R_i. Concurrently, two quotient digits $q_i$and q_{i-1} are generated based on this operation.

These calculations are executed within a custom logic unit, the Partial Remainder and Quotient Digit Generation Block (GPRaQB). The architecture enables the implementation of division units using sequential, matrix-based, or pipelined circuit designs, offering flexibility for various hardware applications. In this work, the focus is placed on sequential circuit implementation. The functionality and correctness of the proposed design are verified through both analytical evaluation and hardware simulation using Field-Programmable Gate Arrays (FPGAs). The results demonstrate that the proposed method provides a reliable and resource-efficient solution for high-performance integer division in digital systems.

Materials and methods.

Unlike addition and multiplication, methods for accelerating integer division have historically received comparatively less attention. This can be attributed to two main factors: first, the inherently sequential nature of the division process, wherein each new quotient digit and partial remainder depends on the outcome of the preceding step; and second, the relatively low frequency of division operations in many computational tasks. However, with the widespread adoption of high-speed addition and multiplication units in modern processors, unoptimised division operations increasingly become a performance bottleneck, necessitating the development of efficient hardware division techniques.

Numerous architectures have been proposed to improve the performance of hardware division units. Traditional digitrecurrence schemes, built on the SRT algorithm family, remain foundational in modern designs. SRT methods iteratively shift the partial remainder and conditionally update it via addition or subtraction of the divisor, reducing arithmetic operations through signeddigit quotient sets (Harris et al., 1997). Comprehensive treatments of these algorithms and their hardware implications are found in canonical texts (Koren, 2017; Parhami, 2010).

Extensions of SRT to higher radices, such as Radix4 and Radix8, enable the generation of multiple quotient bits per cycle, thereby reducing total iteration counts. However, this advantage often incurs greater hardware complexity, particularly in quotient selection logic and lookup tables, which grow significantly with the radix size (Gorodecky et al., 2023; Liu

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

et al., 2025). Early work demonstrated that highradix implementations impose substantial memory demands, which can offset iteration gains in some applications (Taylor, 1985).

Beyond classical digitrecurrence schemes, alternative highspeed division methods have emerged. Hybrid approaches combining Goldschmidt and Mitchell algorithms have been shown to achieve high throughput with efficient hardware resource utilization, particularly on FPGA platforms (Yang, 2025). Likewise, binary logarithmbased division architectures with pipelined stages and errorcorrection structures offer reduced logic complexity while maintaining precision (Ngo et al., 2025).

FPGAoriented design explorations, such as those reconsidering integer divider structures for soft processors and embedded systems, further illustrate the variety of modern implementation strategies (Matthews et al., 2019). Energyefficient approximate logarithmic dividers for integrated circuits also demonstrate favorable tradeoffs in area and performance (Arya et al., 2022).

Matrixbased division schemes use arrays of adders to compute partial remainders and quotient bits in parallel. While these designs can offer low latency and inherent parallelism, their main limitation is the generation of only one quotient bit per row, constraining throughput in highperformance systems (Nikmehr et al., 2007, Nand et al., 2023). Comprehensive comparisons of matrix and iterative architectures highlight the tradeoffs between parallelism, resource usage, and speed (Obshta et al., 2024).

Recent innovations also include datadependent divider designs optimized for areacritical applications, illustrating how adaptive hardware structures can improve efficiency (Patankar et al., 2023). Scalable constant division architectures tailored for FPGA implementation provide further evidence of ongoing efforts to balance performance and hardware cost (Gorodecky et al., 2023).

Across these studies, the central challenge remains the tradeoff between iteration count, hardware complexity, and resource usage, motivating continued architectural innovations in divider design.

As presented in J. E. Tornton: 1955 (US Patent 3,293,418), an enhanced division unit is described that produces two quotient digits q_i, q_{i-1} and the corresponding next partial remainder during each division cycle. The system architecture includes:

- A register for the dividend ,
- A register for the divisor , and
- A two-bit left-shift chain embedded in the dividend register.

Upon completion of the division process, the upper bits of the dividend register contain the remainder , while the lower bits store the final quotient .

Computation of Multiples and Remainder.

To accelerate the division process, the architecture leverages the preparation of multiples of the divisor: *B,2B,3B.* These are computed as follows:

- *2B* is obtained by a single-bit left shift of the value stored in the divisor register B,
- *3B* is calculated by summing *B+2B* using a dedicated binary adder,
- Inverted values $\bar{B}, 2\bar{B}, 3\bar{B}$ are derived by inverting outputs from the

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

respective blocks (using NOT gates).

These prepared values are used to compute subtraction candidates during each cycle: $4R_{i-1}, 4R_{i-1} - B, 4R_{i-1} - 2B, 4R_{i-1} - 3B$ where $4R_{i-1}$ is the result of shifting the previous remainder two bits to the left.

The proposed divider architecture is equipped with three binary adders for computing the required subtractions and a combinational logic unit composed of AND, OR, and NOT gates to process sign information and select valid outcomes. These components work in tandem to determine the appropriate partial remainder and generate two quotient digits per cycle. However, this implementation is hardware-intensive due to the inclusion of four full binary adders, leading to increased logic area and power consumption. To mitigate this drawback, recent design improvements aim to retain the use of precomputed multiples while replacing a portion of the adders with comparator circuits. This hybrid approach effectively reduces hardware complexity without compromising the throughput advantages of producing two quotient digits per cycle. As such, the divider represents a class of high-speed division units with preloaded multiplicands and optimized arithmetic-logical structures for enhanced efficiency in digital systems.

Table 1 – Division Architectures Comparison

Architecture	Quotient Digits per Cycle	Iteration Count (n-bit division)	Hardware Complexity	Memory	Speed	Flexibility
Classical Sequential (1-bit)	1	n	Low	Low	Low	High
SRT Radix-2	1	n	Medium	Medium		Medium
SRT Radix-4	2	n/2	High	High	High	Medium
SRT Radix-8	3	n/3	Very High	Very High	Very High	Low
Matrix Division	1	n	Very High	Low		Low
Proposed Architecture: Precomputed Multiples Divider (2-bit)	2	n/2	High	Medium	High	Medium

As shown in Table 1, a comparative analysis of different divider architectures highlights the trade-offs between iteration count, hardware complexity, and overall performance. While traditional sequential and Radix-2 SRT methods offer simplicity and low hardware demands, they suffer from higher latency due to single-bit quotient generation per cycle. High-radix SRT implementations (e.g., Radix-8 and Radix-16) significantly reduce iteration counts but at the cost of increased memory usage and complex digit selection logic, which complicates implementation and scaling.

In contrast, the proposed architecture-Precomputed Multiples Divider (2-bit per cycle)-achieves a favorable balance. It provides a twofold reduction in iteration count compared to classical methods while avoiding the high memory overhead associated with high-radix SRT designs. The use of precomputed values and selective replacement of arithmetic units with comparators contributes to moderate hardware complexity and enhanced operational speed, making the architecture particularly suitable for FPGA-based and embedded implementations.

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

Thus, the proposed method demonstrates a practical and efficient solution for accelerating integer division in digital systems.

Sequential division devices, where a partial remainder and two quotient digits are formed in each division cycle.

The block diagram of the dividing device, where a partial remainder and two quotient digits are formed in each cycle, is shown in Fig. 1.

The circuit contains a divisible Register, which consists of two parts RgA1 and RgA2, having a left shift circuit by two digits, a register of multiples of the divisor (RgMD), which consists of two registers Rg3B and RgB (Fig. 2) at the outputs of which forward and reverse codes of multiples to the divisor are formed - 3B, 2B, B and $3\overline{B}, 2\overline{B}, \overline{B}$, a generator of partial residues and quotient bits (GPRaQB), the SB synchronization block, and the AND1-AND6 circuits, which serve to receive operands and output the results of the remainder R and the quotient of the division. The synchronization block (SB), which includes (Fig. 3) reversible counter RC, Trigger T, circuits AND7, AND8, delay lines DL0 and DL2, the SB inputs receive a “Start” signal, a clock signal CS, a binary code for the number of shifts N, which is determined by the ratio $N=\frac{n-m}{2}$, where is the bit depth of the divisible A, -the bit depth of the divider B. The outputs of the SB generate the signal “Operand reception”, “Shift” and “End of operations (EO)”.

As can be seen from Fig. 1, the RgMD register is connected to the GPRaQB. The GPRaQB outputs are connected by the RgA register to transmit the next remainder from the GPRaQB output, and also with the RgA2 register to transmit the value of the q_i, q_{i-1} quotient bits in each cycle. The outputs of the RgA1 register are connected to the GPRaQB input via the AND4 circuit.

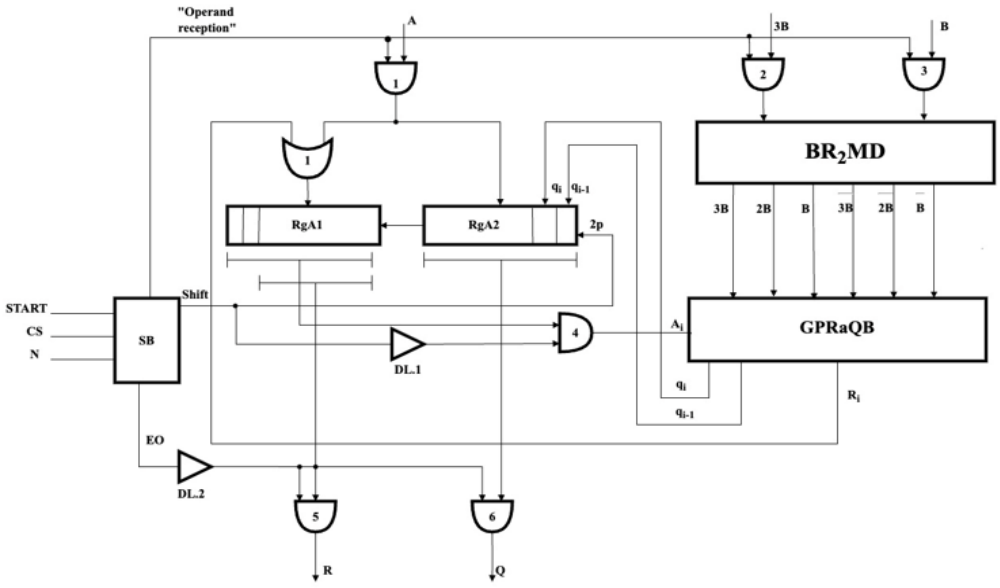


Fig. 1. Functional diagram of a divider with preliminary preparation of multiples, where in each cycle the value of the partial remainder and two bits of the remainder are formed

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

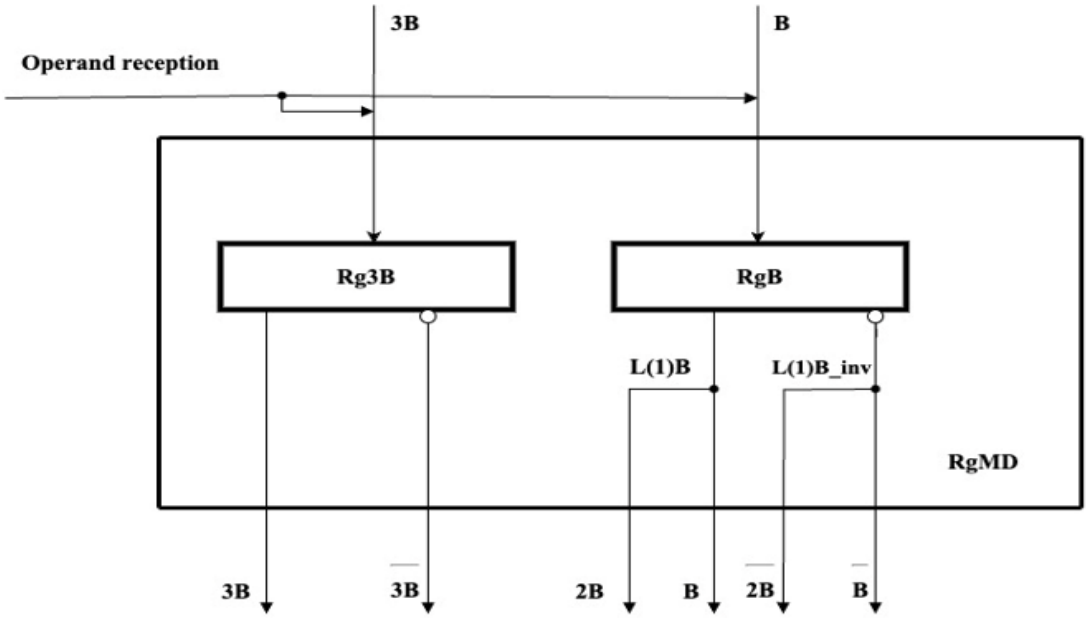


Fig. 2. Registers of multiples of the divisor (RgMD)

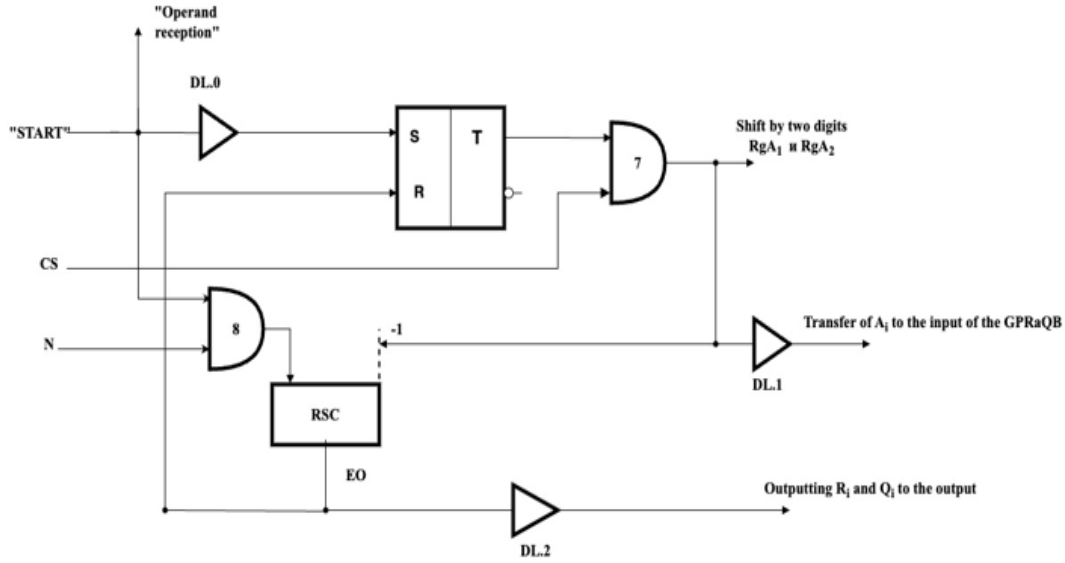


Fig. 3. The structure of the synchronization unit (SU)

Consider the operation of the dividing device.

The “Start” signal generates the “Receive operands” control signal, which, using the AND1 circuit, the most significant digit of the dividend is written into the RgA1 register via the OR1 circuit. The least significant digits of the number A are received into the RgA2 register, while multiples of 3B are simultaneously received into the Rg3B register via the AND2 circuit, and the divisor B is received into the RgB register of the

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

block BRgMD. The “START” signal also receives the code of the number of shifts K via the AND8 circuit and is written into the reversible counter RSC.

The “Start” signal (Fig. 3), delayed on the delay lines DL.0 for the time of receiving the operands, is fed to the S input of the trigger T and sets it to the one state, which allows the first clock signal T AND1 to pass to the input of the circuit AND7, forming the “Shift” signal, which shifts the registers RgA1 and RgA2 by two digits to the left. Also, the “Shift” signal decreases the reading of the counter RC by one. The “Shift” signal, delayed on DL.1 for the time of shifting the registers RgA1 and RgA2, is fed to the second input of the circuit AND4, to the first input of which the $A_{i=}4R_{i-1}+a_{i+1}a_i$ a_i shifted by two digits is fed with the addition of two digits $a_{i+1}a_i$ from the outputs of the circuit AND4, the number A_i is fed to the inputs of the GPRaQB. In the GPRaQB, the number A_2 is compared with the multiples 3B, 2B, B and selects one of them, which allows one of the multiples $3\overline{B}, 2\overline{B}, \overline{B}$ to pass to the input of the adder AD, where the remainder is calculated. The selected multiples also form the bits of the quotient q_i, q_{i-1} the cycle ends with the supply of the remainder q_i, q_{i-1} to the inputs of the register $\mathrm{Rg}A_1$ and the reception of the bits of the quotient in the least significant bits of the register $\mathrm{Rg}A_2$ By this time (Fig. 3), the second AND2 arrives at the input of the AND7 SB, through which the execution of the second division cycle begins: the contents of $\mathrm{Rg}A_1$ and $\mathrm{Rg}A_2$. are shifted by two bits, the counter reading is reduced by one, in the register $\mathrm{Rg}A_1$, the number A_2=$4R_1$ is formed. The bits of the quotient are moved to the left by two places in the GPRaQB A_2=$4R_1$ are compared with the multiples of 3B, 2B, B and one of them is selected, according to which the next remainder R_2 is formed and the bits of the quotient $q_{i-2}, q_{i-3},$ the multiples are accepted into the register $\mathrm{Rg}A_1$ and $\mathrm{Rg}A_2$..

Other remainders and quotient bits are formed similarly. In the last cycle, the division counter (DC) is reset. At the same time, it produces the “End of Operations” signal EO, which switches the T trigger to the “0” state. At the same time, the clock signal is prohibited from entering the circuit. The delayed EO signal on DL.2 produces the contents of the $\mathrm{Rg}A_1$ register, which is output through the AND5 circuit, and simultaneously the quotient Q bits are output from the $\mathrm{Rg}A_2$ register through the AND6 circuit.

Former of partial remainders and quotient bits based on three comparison schemes.

Figure 4 presents the functional diagram of the Partial Remainder and Quotient Bit Generation Block (GPRaQB). This block comprises three main components:

- the Multiple-to-Divider Selection Block (MDSB),
- the Private Discharge Formation Unit (PDFU),
- the Remainder Formation Block (RFB).

The MDSB is implemented using three parallel comparison circuits (CC-1, CC-2, CC-3), along with associated logic gates AND9, AND10, and AND11. The input value $A_i = 4R_{i-1}$ is simultaneously fed to the left inputs of all comparison circuits. The right inputs receive the precomputed multiples of the divisor: B, 2B, and 3B, respectively.

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

The comparisons are performed in parallel, enabling fast decision-making. The output behavior of each comparator is as follows:

- CC-1 generates a logical '1' at Output 1 when $A_i < B$, and at Output 2 when $A_i \geq B$;
- CC-2 outputs '1' at Output 1 if $A_i < 2B$, and at Output 2 if $A_i \geq 2B$;
- CC-3 behaves similarly, producing '1' at Output 1 for $A_i < 3B$, and at Output 2 for $A_i \geq 3B$.

These output signals serve as selection flags that determine which multiple of the divisor will be used for computing the partial remainder and which quotient bits will be formed. The logical structure ensures parallelism, minimizes delay in the decision path, and eliminates the need for sequential selection logic.

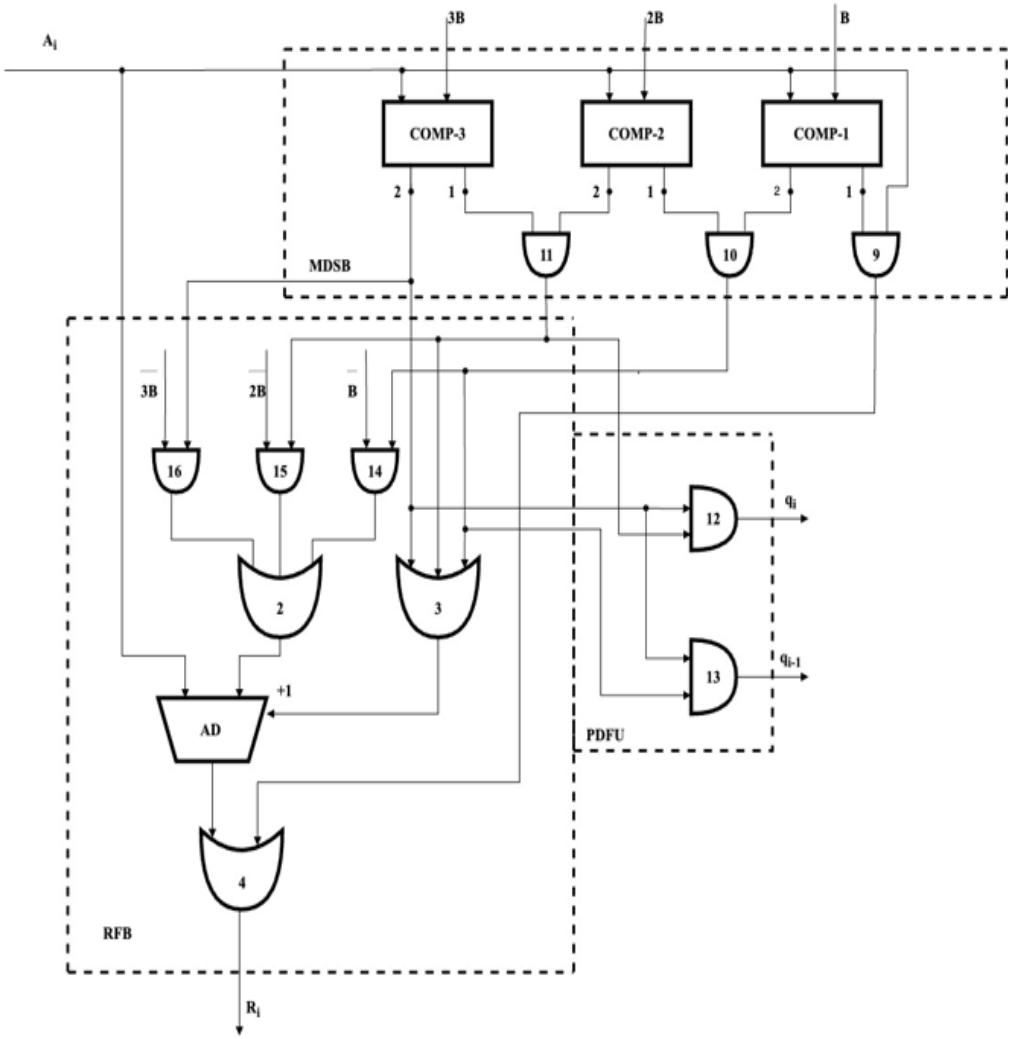


Fig. 4. Functional scheme of formation of partial remainder and quotient digits on three comparison schemes

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

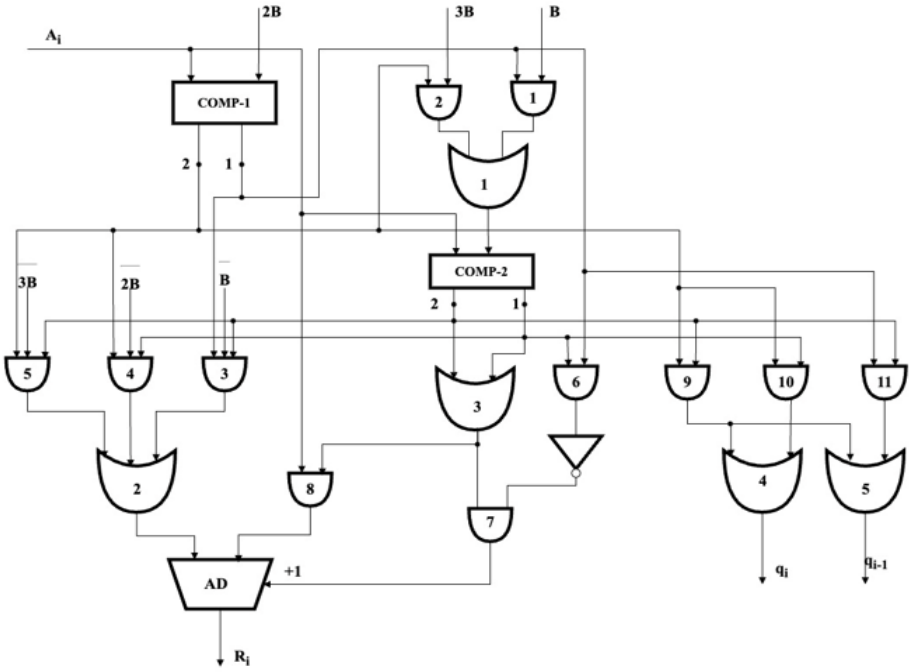


Fig. 5. Functional diagram of the GPRaQB amplifier on two comparison schemes

When A_i<B, the value of A_i is transferred to the output via the AND9 and OR4 circuits. In this case, $R_i=A_i=4R_{i-1}$. The PDFU circuits generate the q_iq_{i-1}=00 bits.

At 2B> ³B, a single signal is formed at the output of the AND10 circuit, which allows the inverse code to be transmitted through the OR2 circuit to the input of the adder AD. In this case, the operations = +1 are performed and = are formed in the PDFU.

At the ratio $2B \leq A_i < 3B$, a single signal is formed at the output of the circuit AND11, which leads to the execution of operations $R_i=A_i + \overline{2B} +1$. In this case, $q_iq_{i-1}=10_2$, at the ratio $A_i \geq 3B$, the signal from the output 2 of the CC-3 leads to the execution of operations $R_i=A_i + \overline{3B} +1$. while at the outputs of the circuits AND12 and AND13 of the PDFU, bits $q_iq_{i-1}=11_2$. are formed.

The advantage of the GPRaQB on three comparison circuits is the simultaneous comparison of operands.

Former of partial remainders and quotient bits based on two comparison schemes.

Figure 5 illustrates the functional diagram of the remainder and quotient digit generation unit (GPRaQB), which consists of an adder (AD), two comparison circuits (CC-1, CC-2), logic gates (AND1–AND11, OR1–OR5), and a NOT gate.

The doubled divisor value (2B) is supplied to the right input of CC-1, while either B or 3B is selectively passed to the right input of CC-2 through gates AND1/OR1 and AND2/OR1, respectively. The operand A_i is fed to both comparison circuits and the right input of the adder via AND8.

Comparison Logic and Operation Modes.

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

Step 1: $A_i < 2B$

CC-1 outputs '1' at output 1, activating AND1 to pass B to CC-2 and blocking 3B through AND2.

If $A_i \geq B$: :

CC-2 selects $\bar{B}$, which is added to A_i in AD to compute $R_i = A_i + \bar{B} + 1$, forming quotient digits $q_i q_{i-1} = 01_2$.

If $A_i < B$:

No subtraction occurs; the adder simply passes $R_i = A_i$, with quotient digits $q_i q_{i-1} = 00_2$. .

Step 2: $A_i \geq 2B$

CC-1 outputs '1' at output 2, enabling 3B to be passed to CC-2 and blocking B via AND1.

If If $A_i < 3B$:

CC-2 selects $2\bar{B}$, and the adder performs $R_i = A_i + 2\bar{B} + 1$, forming $q_i q_{i-1} = 10_2$.

If If $A_i \geq 3B$:

CC-2 selects $3\bar{B}$, leading to $R_i = A_i + 3\bar{B} + 1$, and quotient digits $q_i q_{i-1} = 11_2$.

Each decision is driven by control logic involving CC-1 and CC-2 outputs, routed through specific AND and OR gates. The quotient bits q_i and q_{i-1} are generated using AND9–AND11 and finalized via OR4 and OR5.

Implementation Consideration

A known drawback of the two-comparator-based GPRaQB design is the sequential nature of CC-1 and CC-2 operations, which introduces delay and reduces efficiency in high-frequency applications.

To validate the correctness and performance of the algorithm, the divider based on three parallel comparison circuits (as shown earlier in Fig. 4) was modeled and tested using FPGA implementation. A specific case of dividing a number by a divisor was simulated to ensure correct generation of remainders and quotient bits.

Let the dividend $A=2743_{10}$, divisor $B=11010_2$

$$A=2743_{10} = \{\begin{matrix} a_{12} & a_{11} & a_{10} & a_9 & a_8 & a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & a_0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{matrix}\}$$

$$B=11010_2 = 26_{10}; 2B=52_{10}; 3B=78_{10}$$

The senior five bits of the binary code of the number A are determined by the value $R_0=01010_2 = 10_{10}$. Adding the next two bits $a_7 a_6$ of the number A to the remainder R_0 shifted to the left by two digits gives the number

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

$$A_1 = (2)\ R_0+(a_7a_6)=4R_0+(a_7a_6)= 40+2=42_{10}.$$

For clarity, all calculations for determining the remainders and digits of the quotient are given in the table in the decimal number system.

Table 2 – The order of performing the operation of dividing the dividend A by the divisor B.

Cycles	Enrollment of balances	Formation q_iq_{i-1}
1	$A_1 = 4R_0+(a_7a_6)= 40+2=42_{10}$ 52>42>46, while B≥ $A_i < 2B$ therefore, $R_1 = 42 - 26 = 16$	$q_7q_6 = 01$
2	$A_2 = 4R_1+(a_5a_4)= 64+3=67_{10}$ 78>67≥52, while 3B>$A_i \geq 2B$ therefore, $R_2 = 67 - 52 = 15$	$q_5q_4 = 10$
3	$A_3 = 4R_2+(a_3a_2)= 60+1=61_{10}$ 78>61≥52, while 3B>$A_i > 2B$ therefore, $R_3 = 61 - 52 = 9$	$q_3q_2 = 10$
4	$A_4 = 4R_3+(a_1a_0)= 36+3=39_{10}$ 52>39>26, then $R_4 = 39 - 26 = 13$	$q_1q_0 = 01$

$$Q=01101001_2 = 105_{10};\ R=13_{10};$$
$$\text{Check } A=(Q*B) +R= (105*26) +13=2743_{10}$$

To verify the operation of the dividing device, a behavioral simulation was performed in the Vivado 2017 software package. The device was described using the Verilog programming language. The simulation made it possible to verify the correct functional operation of the device at the logic description level, without including time characteristics or physical modeling in the calculation. The main tasks of the simulation were to identify errors in logic and verify the interaction of components at the signal level.

After performing these tests and checks, no critical errors were found in the logic of the device. Figure 6 shows the simulation result, which shows that it took a total of four clock signals to divide the numbers $A=2743_{10}$ by $B= 26_{10}$. In the timing diagram, the quotient bits obtained in each clock are presented in binary form: $01_2, 10_2, 10_2, 01_2$. . In the diagram, it is clear that the value of bits $q_5q_4 = 10$ и $q_3q_2 = 10$ in the second and third clocks has an analog value. The final value of the quotient Q is formed by concatenating the bits 01101001_2, which is the number $Q= 105_{10}$.

The remainder R on the last (4th bar) shows the remainder value $R=13_{10}$.

Results and discussion.

The current study demonstrates the feasibility and correctness of a divider architecture that generates two quotient bits per cycle through the use of precomputed multiples and comparison-based logic. However, further enhancement of performance is possible by optimizing the architectural variants for specific timing and structural trade-offs.

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

In future work, attention will be focused on designing a sequential divider variant operating over three clock cycles per full division step. This approach is expected to reduce logic complexity while maintaining acceptable performance for resource-constrained systems, such as low-power FPGAs or embedded processors.

Another promising direction involves the development of a matrix-based divider structure capable of performing division operations in just two clock cycles. By leveraging the parallelism of matrix interconnections and reducing propagation delays, this architecture has the potential to significantly improve throughput, especially in high-speed computing applications.

Further research will also explore power–performance trade-offs, logic minimization techniques, and pipelined implementations suitable for ASIC design or integration into arithmetic logic units (ALUs).

Conclusion.

In this paper, a structural diagram of a dividing device with preliminary preparation of multiples of the divisor has been proposed and validated. Two functional variants were developed, enabling each division cycle to compute the partial remainder and generate two bits of the quotient through the use of two or three comparison circuits. This approach effectively reduces the number of division cycles by half compared to classical methods.

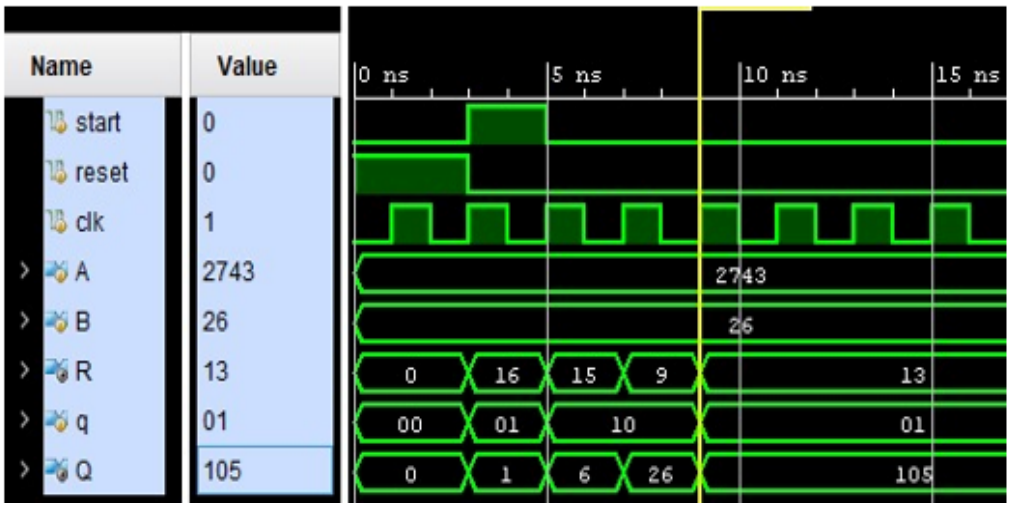


Fig. 6. Result of FPGA dividing device simulations for the example where dividend A= and divisor B

The architecture was modeled using Verilog HDL and verified through simulation on an FPGA platform. As illustrated in Figure 6, the divider successfully processed the example division by , generating the correct quotient and remainder , with all logic operations performing as expected.

These results confirm the correctness and efficiency of the proposed design and demonstrate its suitability for implementation in hardware-based digital arithmetic systems.

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

REFERENCES

Arya N., Soni T., Pattanik M., Sharma G.K. (2022). Energy efficient logarithmic divider for ASIC and FPGA // Microprocessors and Microsystems. — Vol. 92. — Article 104498. DOI: 10.1016/j.micpro.2022.104498. [In Eng.].

A constant time hardware architecture for CSIDH key exchange protocol (Montgomery modular operations). (2025). arXiv:2508.11082. DOI: 10.48550/arXiv.2508.11082. [In Eng.].

Dikshit S., Nadeem M. (2017). FPGA accomplishments of a 16 bit divider // Imperial Journal of Interdisciplinary Research. — Vol. 3. — No. 2. Pp. 1–5. [In Eng.].

Flynn M.J., Oberman S.F. (2001). Advanced Computer Arithmetic Design. — Wiley, Hoboken, NJ, USA. https://www.wiley.com/en-es/Advanced+Computer+Arithmetic+Design-p-9780471412090 [In Eng.].

Gorodecky D., Sousa L. (2023). Scalable Architecture of Constant Division on FPGA // Proceedings of ARITH 2023. Pp. 13–19. https://arith2023.arithsymposium.org/papers/Scalable%20Architecture%20of%20Constant%20Division%20on%20FPGA.pdf [In Eng.].

Harris D.L., Oberman S.F., Horowitz M.A. (1997). SRT Division Architectures and Implementations // Proceedings of IEEE Symposium on Computer Arithmetic. Pp. 18–25. [In Eng.].

Kim K., Lee S., Park J. (2025). Shared Recurrence Floating Point Divide/Sqrt and Integer Division Algorithms for FPGA // IEEE Transactions on Computers. [In Eng.].

Koren I. (2017). Computer Arithmetic Algorithms, 2nd Ed. — CRC Press, Boca Raton, FL, USA. https://www.routledge.com/Computer-Arithmetic-Algorithms/Koren/p/book/9781568811604 [In Eng.].

Liu H., Chen Y., Zhu X., Zhang L. (2025). A SignedDigit Hybrid Stochastic Number Method for Arithmetic Computation // IEEE Transactions on Very Large Scale Integration (VLSI) Systems. Bookmark, DOI: 10.1109/TVLSI.2025.3532948 [In Eng.].

Liddicoat A.A., Cary S., Pang J. (2005). FPGA based high performance arithmetic pipeline synthesis // Advances in Signal Processing, Algorithms, Architecture, and Implementation XV. DOI: 10.1117/12.615338 [In Eng.].

Matthews J., Zhang Z., Chen Z. (2019). Rethinking integer divider design for FPGA based soft processors // FieldProgrammable Custom Computing Machines (FCCM). Pp. 115–122. [In Eng.].

Nand A.Y., V.R. (2023). FPGA implementation of high speed efficient floating point ALU // Proceedings of the International Conference on Control, Communication, and Computing. Pp. 1–5. [In Eng.].

Nikmehr H., Phillips B., Lim C.C. (2007). Fast radix4 floating point divider with quotient digit selection by comparison multiples // The Computer Journal. — Vol. 50. — No. 1. Pp. 37–44. DOI: 10.1093/comjnl/bxl048. [In Eng.].

Obshta A., Khoma V., Prokopchuk A. (2024). Digital division algorithms for efficient execution on integrated circuits // Advances in CyberPhysical Systems. — Vol. 1. — No. 1. Pp. 48–55. DOI: 10.23939/acps2024.01.046 [In Eng.].

Prasad H., Kumar A., Power estimation of Radix2 and high radix embedded divider IP cores for modern FPGAs (2025). World Scientific Publishing. DOI: https://doi.org/10.1142/S0218126625503645 [In Eng.].

ParandehAfshar H., Verma A.K., Brisk P., Ienne P. (2010). Improving FPGA performance for carrysave arithmetic // IEEE Transactions on Very Large Scale Integration (VLSI) Systems. — Vol. 18. — No. 4. Pp. 578–591. DOI: 10.1109/TVLSI.2010.2045987. [In Eng.].

Patankar U.S., Flores M.E., Koel A. (2023). Novel datadependent divider circuit block implementation for complex division and areacritical applications // Scientific Reports. — Vol. 13. — Article 3027. DOI: 10.1038/s41598-023-28343-3. [In Eng.].

Parhami B. (2010). Computer Arithmetic: Algorithms and Hardware Designs. — Oxford University Press, Oxford, UK. https://www.researchgate.net/publication/236944290_Computer_Arithmetic_Algorithms_and_Hardware_Designs_Second_Edition [In Eng.].

Song X., Lu R., Guo Z. (2025). High performance reconfigurable pipeline implementation for FPGA based SmartNIC // Micromachines. — Vol. 15. — https://doi.org/10.3390/mi15040449. [In Eng.].

Yang J. (2025). FPGA Design and Implementation of FixedPoint Fast Divider Using Goldschmidt Division Algorithm and Mitchell Multiplication Algorithm // arXiv:2508.14611. DOI: 10.48550/arXiv.2508.14611. [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.2. Number 26 (2026). Pp. 351–371
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.023
UDC 004.8:004.42:004.65

MATHEMATICAL AND ALGORITHMIC APPROACHES TO THE DEVELOPMENT OF A COLLABORATIVE FILTERING-BASED RECOMMENDER SYSTEM

K.N. Uteuliyeva[1], B.Z. Kenzhegulov[1]*, T.A. Karazhigitova[1], H.İ. Bülbül[2], Z.Zh. Zhanuzakova[1]

[1]Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[2]Gazi University, Ankara, Turkiye.
E-mail: kenzhegulov_bz@mail.ru

Kamka Uteuliyeva — Candidate of Physico-Mathematical Sciences, associate professor, Department of Mathematics and Methods of Teaching Mathematics." Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0009-0004-1195-1642;
Beket Kenzhegulov — Doctor of Technical Sciences, Professor of the Department of Mathematics and Methods of Teaching Mathematics" Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
E-mail: kenzhegulov_bz@mail.ru, https://orcid.org/0000-0001-6230-2926;
Tamara Karazhigitova — Doctor of Pedagogical Sciences, professor, Department of Mathematics and Methods of Teaching Mathematics", Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0009-0005-1769-8737;
Halil İbrahim Bülbül — PhD, Professor, Department of Computer and Instructional Technologies, Gazi Eğitim Fakültesi, Gazi University, Ankara, Turkiye
https://orcid.org/0000-0002-6525-7232;
Zulfiya Zhanuzakova — Master, Department of Mathematics and Methods of Teaching Mathematics. Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0001-7728-0467.

© K.N. Uteuliyeva, B.Z. Kenzhegulov, T.A.Karazhigitova, H.İ. Bülbül, Z.Zh.Zhanuzakova

Abstract. The article examines the mathematical and algorithmic support of a recommender system based on the collaborative filtering method. The relevance of the study is associated with the growth of digital content volumes and the need for automated generation of personalized recommendations for users. A movie recommender system is considered as the subject area; however, the proposed model can also be applied to other

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

types of objects, such as products, books, musical compositions, educational resources, and information materials. The main focus is on constructing a mathematical model of user preferences, forming a rating matrix, determining similarity between objects, and developing an algorithm for ranking recommendations. The study uses an item-based collaborative filtering approach, in which recommendations are generated based on the similarity between objects calculated from user ratings. Cosine similarity is used as the proximity measure, while the k-nearest neighbors algorithm is applied to find the most similar objects. The proposed approach makes it possible to represent a recommender system not only as a software implementation but also as a mathematical and algorithmic model for processing incomplete and sparse data. The model includes a set of users, a set of objects, a preference matrix, a similarity function, a rule for selecting nearest objects, and a procedure for generating Top-N recommendations.

Keywords: recommender system, collaborative filtering, mathematical model, algorithmic support, preference matrix, k-nearest neighbors, cosine similarity measure, ranking, personalization

For citation: K.N. Uteuliyeva, B.Z. Kenzhegulov, T.A. Karazhigitova, H.İ. Bülbül (2026). Mathematical and algorithmic approaches to the development of a collaborative filtering-based recommender system // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 351–371. https://doi.org/10.54309/IJICT.2026.26.2.023. (In Eng.).

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

КОЛЛАБОРАТИВТІК СҮЗГІЛЕУ НЕГІЗІНДЕГІ ҰСЫНЫМДЫҚ ЖҮЙЕНІ ӘЗІРЛЕУДІҢ МАТЕМАТИКАЛЫҚ-АЛГОРИТМДІК ТӘСІЛДЕРІ

К.Н. Утеулиева[1], Б.З. Кенжегулов[1]*, Т.А. Каражигитова[1], Х. Булбул[2], З.Ж. Жанузакова[1]

[1]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;
[2]Гази университеті, Анкара, Түркия.
E-mail: kenzhegulov_bz@mail.ru

Утеулиева Камка — физика-математика ғылымдарының кандидаты, қауымдастырылған профессор, Математика және математиканы оқыту әдістемесі кафедрасы, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0009-0004-1195-1642;
Кенжегулов Бекет — техника ғылымдарының докторы, профессор, Математика және математиканы оқыту әдістемесі кафедрасы, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
E-mail: kenzhegulov_bz@mail.ru, https://orcid.org/0000-0001-6230-2926;
Каражигитова Тамара — педагогика ғылымдарының докторы, профессор, Математика және математиканы оқыту әдістемесі кафедрасы, 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/0009-0005-1769-8737;
Халил Ибрагим Булбул — PhD, профессор, Гази білім беру факультеті, Компьютерлік және оқыту технологиялары кафедрасы, Гази университеті, Анкара, Түркия
https://orcid.org/0000-0002-6525-7232;
Жанузакова Зульфия — магистр, Математика және математиканы оқыту әдістемесі кафедрасы, Х. Досмұхамедов атындағы Атырау университеті, Қазақстан.
https://orcid.org/0000-0001-7728-0467.

© К.Н. Утеулиева, Б.З. Кенжегулов, Т.А. Каражигитова, Х.Булбул, З.Ж. Жанузакова

Аннотация. Мақалада коллаборативтік сүзгілеу әдісіне негізделген ұсынымдық жүйенің математикалық-алгоритмдік қамтамасыз етілуі қарастырылады. Зерттеудің өзектілігі цифрлық контент көлемінің артуымен және пайдаланушылар үшін жекелендірілген ұсынымдарды автоматтандырылған түрде қалыптастыру қажеттілігімен байланысты. Пәндік сала ретінде фильмдерге арналған ұсынымдық жүйе қарастырылады, алайда ұсынылған модель басқа объектілерге де, атап айтқанда тауарларға, кітаптарға, музыкалық композицияларға, білім беру ресурстарына және ақпараттық материалдарға қолданылуы мүмкін. Негізгі назар пайдаланушы қалауларын сипаттайтын математикалық модельді құруға, бағалау матрицасын қалыптастыруға, объектілер арасындағы ұқсастықты анықтауға және ұсынымдарды ранжирлеу алгоритмін әзірлеуге аударылады. Жұмыста коллаборативтік сүзгілеудің item-based тәсілі қолданылады, бұл тәсілде ұсынымдар пайдаланушылардың бағалары негізінде есептелетін объектілер арасындағы ұқсастыққа сүйене отырып қалыптастырылады. Жақындық өлшемі ретінде косинустық ұқсастық өлшемі, ал ең жақын объектілерді іздеу үшін k жақын көрші алгоритмі қолданылады. Ұсынылған тәсіл ұсынымдық жүйені тек бағдарламалық іске асыру ретінде ғана емес, сонымен қатар толық емес және сиретілген деректерді өңдеуге арналған математикалық-алгоритмдік модель ретінде қарастыруға мүмкіндік береді. Модель пайдаланушылар жиынын, объектілер жиынын, қалаулар матрицасын, ұқсастық функциясын, ең жақын объектілерді таңдау ережесін және Top-N ұсынымдарын қалыптастыру процедурасын қамтиды.

Түйінді сөздер: ұсынымдық жүйе, коллаборативтік сүзгілеу, математикалық модель, алгоритмдік қамтамасыз ету, қалаулар матрицасы, k жақын көрші, косинустық ұқсастық өлшемі, ранжирлеу, жекелендіру

Дәйексөздер үшін: К.Н. Утеулиева, Б.З. Кенжегулов, Т.А. Каражигитова, Х.Булбул, З.Ж. Жанузакова (2026). Коллаборативтік сүзгілеу негізіндегі ұсынымдық жүйені әзірлеудің математикалық-алгоритмдік тәсілдері // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 351–371. https://doi.org/10.54309/IJICT.2026.26.2.023. (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

МАТЕМАТИКО-АЛГОРИТМИЧЕСКИЕ ПОДХОДЫ К РАЗРАБОТКЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ

К.Н. Утеулиева[1], Б.З. Кенжегулов[1]*, Т.А. Каражигитова[1], Х. Бюльбюль[2], З.Ж. Жанузакова[1]

[1]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
[2]Гази университеті, Ankara, Turkiye
E-mail: kenzhegulov_bz@mail.ru

Утеулиева Камка — кандидат физико-математических наук, ассоциированный профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0009-0004-1195-1642;
Кенжегулов Бекет — доктор технических наук, профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
E-mail: kenzhegulov_bz@mail.ru, https://orcid.org/0000-0001-6230-2926;
Каражигитова Тамара — доктор педагогических наук, профессор кафедры математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0009-0005-1769-8737;
Х. Бюльбюль — PhD, профессор кафедры компьютерных и образовательных технологий факультета образования Гази, Университет Гази, Анкара, Турция
https://orcid.org/0000-0002-6525-7232;
Жанузакова Зульфия — магистр, кафедра математики и методики преподавания математики Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0001-7728-0467.

© К.Н. Утеулиева, Б.З. Кенжегулов, Т.А. Каражигитова, Х.Бюльбюль, З.Ж. Жанузакова

Аннотация. В статье рассматривается математико-алгоритмическое обеспечение рекомендательной системы, основанной на методе коллаборативной фильтрации. Актуальность исследования связана с ростом объемов цифрового контента и необходимостью автоматизированного формирования персонализированных рекомендаций для пользователей. В качестве предметной области рассматривается рекомендательная система фильмов, однако предложенная модель может быть использована и для других объектов: товаров, книг, музыкальных композиций, образовательных ресурсов и информационных материалов. Основное внимание уделяется построению математической модели пользовательских предпочтений, формированию матрицы оценок, определению сходства между объектами и разработке алгоритма ранжирования рекомендаций. В работе используется item-based подход

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

коллаборативной фильтрации, при котором рекомендации строятся на основе сходства между объектами, вычисленного по оценкам пользователей. В качестве меры близости применяется косинусная мера сходства, а для поиска наиболее близких объектов используется алгоритм k ближайших соседей. Предложенный подход позволяет представить рекомендательную систему не только как программную реализацию, но и как математико-алгоритмическую модель обработки неполных и разреженных данных. Модель включает множество пользователей, множество объектов, матрицу предпочтений, функцию сходства, правило отбора ближайших объектов и процедуру формирования Top-N рекомендаций.

Ключевые слова: рекомендательная система, коллаборативная фильтрация, математическая модель, алгоритмическое обеспечение, матрица предпочтений, k ближайших соседей, косинусная мера сходства, ранжирование, персонализация

Для цитирования: К.Н. Утеулиева, Б.З. Кенжегулов, Т.А. Каражигитова, Х.Бюльбюль, З.Ж. Жанузакова (2026). Математико-алгоритмические подходы к разработке рекомендательной системы на основе коллаборативной фильтрации // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 351–371. https://doi.org/10.54309/IJICT.2026.26.2.023. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Modern digital services contain vast amounts of information. Users are offered thousands of films, products, books, music, news, and educational materials. With such information overload, it becomes difficult to independently select an item that truly matches individual interests. Therefore, there is a need for intelligent systems capable of analyzing user behavior and automatically generating personalized recommendations.

Recommender systems are one of the most in-demand areas of data analysis and machine learning. They are used in online cinemas, online stores, social networks, educational platforms, music services, and digital libraries. Their main task is to suggest the most relevant set of objects to a specific user based on accumulated information about users and objects.

From a scientific perspective, a recommender system is not only a software tool but also a mathematical and algorithmic model. It is based on the formalization of user preferences, the construction of an interaction matrix, the calculation of similarities between objects or users, the ranking of candidates, and the selection of the most suitable recommendations.

One of the most common methods for building recommender systems is collaborative filtering. Its idea is to leverage the collective experience of users. If different users have rated the same items similarly, these similarities can reveal hidden patterns in preferences and suggest new items to the user.

This article examines the mathematical and algorithmic support for a movie recommendation system based on collaborative filtering. A user ratings matrix is used as the initial structure. The k-nearest neighbors algorithm is used to find similar objects. The

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

cosine measure is used to determine the degree of similarity between movies.

The aim of the article is to develop, formally describe, and experimentally evaluate a mathematical and algorithmic model of an item-based collaborative filtering recommender system using similarity measures and the k-nearest neighbors method.

To achieve this aim, the following tasks are solved:

– to describe the recommender system as an object of mathematical modeling;

– to formalize the user–item rating matrix and the structure of sparse preference data;

– to define similarity functions between objects using cosine similarity, Euclidean similarity, and Pearson correlation;

– to describe the k-nearest neighbors model for selecting similar objects;

– to define the rating aggregation rule and the Top-N recommendation generation procedure;

– to justify threshold values for filtering sparse rating data;

– to conduct a computational experiment using the MovieLens 100K dataset;

– to evaluate the model using Precision@5, Recall@5, RMSE, and coverage;

– to compare different similarity measures and determine the optimal value of k;

– to demonstrate the operation of the system using a recommendation example for a specific user.

The subject of the research is the mathematical and algorithmic support of a recommender system based on collaborative filtering.

The practical significance of the work is that the proposed model can be used in the development of recommender systems for films, products, books, educational resources and other digital objects.

Materials and methods.

Recommender system as a problem of mathematical modeling

A recommender system is designed to select the objects most suitable for a specific user. If we consider this task purely from a practical perspective, it appears to be similar to the usual display of a list of movies or products. However, from a mathematical perspective, a recommender system solves a more complex problem-the problem of modeling preferences and ranking objects (Aggarwal, 2016).

User preferences cannot be observed directly. The system doesn't know the user's intrinsic interests, but it can analyze their actions: ratings, views, purchases, clicks, adding items to favorites. These actions are considered external manifestations of preferences. Therefore, the first task of mathematical modeling is to translate user behavior into numerical form.

In a movie recommendation system, such a numerical expression is a movie rating. If a user gives a movie a 5, it indicates a high level of interest. A low rating indicates that the movie doesn't match their preferences. If there is no rating, the system has no direct information about the user's attitude toward the movie.

Thus, the recommendation problem arises in the context of incomplete data. Users rate only a small fraction of films, while the majority of possible ratings remain unknown.

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

Therefore, a recommender system must not only store existing ratings but also, based on these ratings, suggest other films that might be of interest to the user.

From a mathematical perspective, this problem can be described as follows: there are a set of users, a set of objects, and a partially populated rating system. It is necessary to determine which objects, among those not yet rated, are most likely to be of interest to the user.

This article uses a collaborative filtering approach. Its distinctive feature is that recommendations are based not on detailed descriptions of the films themselves, but on the similarity of ratings given by users. In other words, the system analyzes not the film's content directly, but rather how it was perceived by other users.

This approach allows for the identification of hidden connections between objects. For example, two films may belong to different genres, but if they are frequently rated highly by the same users, the system may consider them similar. This is a significant advantage of collaborative filtering over simple genre matching (Ricci et. al., 2015).

Therefore, the recommender system in this paper is considered as a mathematical-algorithmic system that performs the following transformations:

user ratings → preference matrix → object similarity → nearest neighbors → ranked list of recommendations.

It is this sequence that determines the mathematical and algorithmic basis of the system being developed.

General statement of the recommendation problem

Let's assume there are a set of users interacting with a set of objects. In the subject area under consideration, these objects are movies. Each user can rate a subset of the movies on a given scale. For example, a rating could range from 0.5 to 5.

The task of a recommender system is to identify for a particular user a list of films that he has not yet rated, but that may match his interests.

To solve this problem, it is necessary to answer several questions:

– how to present data about users and movies;
– how to take into account known estimates;
– how to handle missing grades;
– how to identify similarities between films;
– how to choose the most suitable objects;
– how to sort recommendations by relevance.

In collaborative filtering, the initial information is not the description of objects, but the rating history. The system assumes that if two films received similar ratings from a large number of users, they are close in user perception. Therefore, if one of these films was liked by a particular user, the other one can also be recommended.

This work uses an item-based approach, that is, an approach based on the similarity of objects. This means that the system compares movies, not users. Each movie is described by the ratings given to it by users. The most similar movies are then identified for each movie.

This approach is convenient for several reasons. First, the number of objects in

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

the system is often smaller or more stable than the number of user actions. Second, the similarity between films can be calculated in advance. Third, recommendations are easier to explain: a film is recommended because it is similar to another film that the user has already rated positively. The item-based collaborative filtering approach was systematically developed by Sarwar et al. who showed that item-to-item similarity can be effectively used for scalable recommendation generation.

The general recommendation task can be formulated as follows: for a given user, it is necessary to find a set of objects that have not been rated by him previously, but have the greatest proximity to the objects that have received positive ratings from him (Adomavicius et.al., 2005).

Thus, the recommendation problem is reduced to the problem of searching and ranking objects in the space of user ratings.

Mathematical model of a recommender system

A mathematical model of a recommender system describes the system's main elements, the relationships between them, and the rules for generating recommendations. In this paper, the model is built using a collaborative filtering method and includes the following components:

– many users;
– a set of objects;
– user ratings matrix;
– vector representation of objects;
– the similarity function between objects;
– set of nearest neighbors;
– the rule for forming the recommendation list.

Let us consider these elements in turn.

Let a set of users be given:

$$U = \{u_1, u_2, \dots, u_n\}$$

and many films:

$$M = m^1, m^2, \dots, m_{k\cdot}$$

Each user can assign a numerical rating to certain films. Based on these ratings, a preference matrix is generated:

$$R = [r_{ij}]$$

where r_{ij} is the rating given by user u_j to the film m_i.

If the user hasn't rated a movie, the corresponding value is missing. In a computational implementation, such a value can be represented as zero or a blank. It's

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

important to understand that the absence of a rating doesn't always indicate a negative user rating. Most often, it simply means the user hasn't yet interacted with the object.

R matrix is the central element of the mathematical model. It contains information about the interactions of all users with all films. Each row of this matrix corresponds to a single film and shows how this film was rated by different users. Each column corresponds to a single user and reflects their ratings of different films.

This paper uses an item-based approach, focusing on the matrix rows. Each row is considered a vector describing the film through user ratings:

$$v_i = (r_i^1, r_i^2, \dots, r_{in})$$

More generally, each object is represented as a vector in the user rating space:

$$v_i = (r_i^1, r_i^2, \dots, r_{in})$$

where r_{ij} denotes the rating assigned to item m_i by user u_j. If the rating is missing, it is not interpreted as a negative evaluation; it indicates the absence of interaction between the user and the item. Therefore, similarity calculations are performed only over the set of users who rated both compared items.

This vector can be understood as a numerical profile of the film. It reflects not the film's content, but rather the way users perceive it. If two films have similar rating vectors, then users perceive them similarly. Consequently, these films can be considered close in preference space.

This is where the core mathematical idea of collaborative filtering comes into play: the similarity of objects is determined not by their external description, but by the structure of user ratings.

To determine the similarity between two movies, a formal similarity function is introduced:

$$similarity(m\_i, m\_j) = sim(v_i, v_j)$$

Where v_i are rating vectors v_i are rating vectors of items m_i and m_i

The function returns a numerical value that characterizes the degree of proximity between two objects in the user preference space. In this study, three similarity measures are considered for comparison: cosine similarity, Euclidean similarity, and Pearson correlation. This makes it possible not only to construct recommendations but also to experimentally determine which proximity measure provides better results for the selected dataset.

In essence, this means the following: if two films received similar ratings from the same users, the angle between their vectors will be small, and the similarity value will be high. If the rating patterns differ, the similarity value will be low.

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

For each film m_i, the set of its nearest neighbors is determined:

$N_k(m_i)$,

where ***k*** is the number of most similar films.

The set $N_k(m_i)$, contains those films that are most similar to the original film m_i. These films are candidates for recommendation to the user if the original film was rated positively.

Next, a set of films already known to the user is entered:

M_u— a set of films that user u has already rated.

The system's goal is not to recommend movies to the user from the M_u set, but to select only those new to the user. Therefore, recommendations are generated from the following set:

$$M \ M_u,$$

that is, from films that are in the general catalog, but have not yet been rated by this user.

The final list of recommendations can be summarized as:

$$Rec(u)=Top\text{-}N(M \ M_u)$$

This means that for user u, the N most suitable movies are selected from among those not yet rated.

The order of films in the list is determined by their degree of similarity to films the user has already rated positively. The greater the similarity, the higher the film appears in the final list.

Thus, the mathematical model of the recommender system can be described as follows: user ratings are represented as a sparse matrix, films are considered as vectors in the space of user ratings, the similarity between films is calculated using the proximity function, then the nearest neighbors are found for positively rated films, from which a ranked list of new recommendations is formed.

This model is relatively simple, yet captures the core essence of collaborative filtering. It allows for the formalization of the recommendation process and the linking of user data with the algorithm for generating a personalized list.

Features of the preference matrix and the sparsity problem

The preference matrix is the basis of the mathematical model, but in real systems it is almost always sparse. This means that most of its elements are unknown. The user rates only a small subset of the available movies, leaving the rest unrated (Schafer et. al., 2007).

For example, if the system has 10,000 movies and 1,000 users, theoretically 10 million ratings are possible. However, in practice, users may only leave 100,000 ratings. In this case, only a small portion of the matrix will be populated.

Data sparsity is one of the main challenges of recommender systems. It affects the quality of similarity calculations. If two movies share too few users who have rated them, it's difficult to reliably determine whether they are similar or not.

To reduce the impact of sparsity, data pre-filtering is used. Movies with too few

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

ratings, as well as users with very low activity, are excluded from the training set. This allows the model to retain only the data that provides more consistent information about preferences (Su et. al., 2009).

In the system under consideration, two threshold parameters can be used:

– minimum number of film ratings;

– the minimum number of films rated by the user.

In the computational experiment, the following threshold values were used:

$$n_{\{movie\}}^{\{min\}} = 20 \quad n_{\{user\}}^{\{min\}} = 20$$

This means that movies with fewer than 20 ratings and users with fewer than 20 rated movies were excluded from the experimental sample. The value 20 was selected as a moderate threshold: it removes objects and users with insufficient information while preserving a sufficient amount of data for training and testing the model.

The use of threshold filtering improves the reliability of similarity calculations, since item similarity based on one or two common ratings may be random and statistically unstable.

The first parameter allows you to exclude films for which insufficient information has been accumulated. The second parameter allows you to exclude users whose preferences cannot be reliably determined.

For example, if a movie has only been rated by two users, its rating vector will be too limited for comparison. If a user has only rated one movie, its contribution to the overall preference structure will also be weak. Therefore, such data may be noisy.

However, filtering should be moderate. Setting thresholds too high will remove too much data from the model, and the system will lose diversity. Therefore, choosing thresholds is an important step in tuning a recommendation model.

Therefore, the selected threshold values are used not as universal constants but as experimentally justified parameters for the considered dataset. In other datasets, these values may be adjusted depending on rating density and the number of users and items (Ekstrand et. al., 2011).

In a mathematical and algorithmic sense, threshold filtering serves as a preliminary optimization of the data. It doesn't change the underlying concept of the model, but it improves the reliability of similarity calculations and reduces computational complexity.

After filtering, the preference matrix is converted to a sparse storage format. This is necessary for efficient implementation of the algorithm. Only known ratings are stored in memory, rather than the entire matrix. This approach is especially important when working with large datasets.

Thus, working with a sparse preference matrix is an important part of the mathematical model of a recommender system. It demonstrates that the model must take into account not only the known ratings but also the structure of the missing data (Linden

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

et.al., 2003).

Similarity function between objects

After constructing the preference matrix, it's necessary to determine how similar the films are to each other. To do this, a similarity function is introduced. This function is one of the key elements of the mathematical model.

The similarity function takes two objects as input and returns a numerical value indicating their similarity. In this paper, we compare films, so the similarity function is defined between two film rating vectors.

Each movie is represented by a vector of user ratings. If two movies have similar vectors, they are considered similar. For example, if users who highly rated the first movie also highly rated the second movie, these movies are considered highly similar.

Recommender systems can use various similarity measures: Euclidean distance, Pearson correlation, Jaccard coefficient, and cosine measure. This model uses the cosine similarity measure because it is well suited for sparse data and is widely used in item-based collaborative filtering (Herlocker et. al., 2004).

The cosine similarity measure can be written in general form:

The cosine similarity between two items and is calculated as:

$$sim_{cos}(m_i, m_j) = \frac{\sum_{u \in U_{ij}} r_{ui} r_{uj}}{\sqrt{\sum_{u \in U_{ij}} r_{ui}^2} \sqrt{\sum_{u \in U_{ij}} r_{uj}^2}},$$

where U_{ij} is the set of users who rated both items m_i and m_j,r_{ui} is the rating of user u for item i, and r_{uj} is the rating of the same user for item j.

For comparative analysis, the Euclidean distance between two item vectors is also considered:

$$d_{euc}(m_i, m_j) = \sqrt{\sum_{u \in U_{ij}} (r_{ui} - r_{uj})^2}$$

Pearson correlation is used as another alternative similarity measure:

Where $\overline{r_i}$ and $\overline{r_j}$

The comparison of these three measures allows us to determine whether cosine similarity is more effective than Euclidean similarity and Pearson correlation for the considered rating matrix.

This notation means that the similarity between films m_i and m_j is defined as the cosine of the angle between their rating vectors v_i and v_i.

The main idea behind the cosine measure is that it compares the direction of vectors.

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

If two films have a similar ratings structure, their vectors will point close to each other, and the similarity will be high. If the ratings are distributed differently, the similarity will be low.

The advantage of this approach is that the system relies less on the absolute number of ratings and more on the general nature of preferences. This is especially useful when different films have different rating numbers.

Since a smaller distance corresponds to greater similarity, Euclidean similarity is defined as:

$$sim_{euc}(m_i, m_j) = \frac{1}{1 + d_{euc}(m_i, m_j)}$$

$$sim_{pearson}(m_i, m_j) = \frac{\sum_{u \in U_{ij}} (r_{ui} - \overline{r_i})(r_{uj} - \overline{r_j})}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \overline{r_i})^2} \sqrt{\sum_{u \in U_{ij}} (r_{uj} - \overline{r_j})^2}}$$

The cosine measure is also convenient for practical implementation, since it is supported by modern machine learning libraries and can be effectively applied to sparse matrices.

Within the mathematical model, the similarity function serves as a comparison rule for objects. Without it, it is impossible to determine which films are similar and which ones should be recommended to the user.

Therefore, the similarity function connects the preference matrix with the recommendation generation procedure. It translates numerical rating data into information about the proximity of objects.

Nearest Object Search Model

Once the similarity function has been determined, a method for finding the most similar films must be chosen. For this, the k-nearest neighbors algorithm is used.

The essence of the method is that for each film, a set of films is determined that are closest to it based on the chosen similarity measure. These films are called nearest neighbors.

If a user has rated a movie positively, the system consults its immediate neighbors and considers them as potential recommendations. For example, if a user has rated the movie m_i highly, then movies from the set $N_k(m_i)$ may be suggested as similar ones.

In the mathematical model, the set of k nearest neighbors for item m_i is defined as:

$$N_k(m_i) argtop_k = sim(m_i, m_j) | m_j \in M, m_j \neq m_i$$

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

This means that for each item m_i, the algorithm selects k items with the highest similarity values. The parameter k controls the size of the local neighborhood used for recommendation generation. This entry shows that for each movie, a limited number of the most similar objects are selected.

The parameter k is an important parameter of the model. If k is too small, the system will consider a very narrow range of similar films. This may result in insufficient diversity in recommendations. If k is too large, less similar films will be included in the neighbor list, and recommendations may become less accurate.

Therefore, the choice of k should take into account the size of the data, the density of the rating matrix, and the required number of recommendations.

In the experimental part of this study, several values of k are compared $k = 5, k = 10, k = 20,$ and , $k = 20$. The optimal value is selected according to *Precision@5,Recall@5,RMSE,* and coverage.

In practical problems, the value of k is often selected experimentally (Koren et. al., 2009).

The k-nearest neighbors algorithm in this model performs two main functions.

The first function is a local search for similar objects. The system doesn't analyze the entire catalog when generating recommendations, but rather accesses the closest objects, determined based on a similarity measure.

The second function is to generate a candidate set. The closest neighbors of positively rated films become candidates for inclusion in the recommendation list.

Thus, the nearest object search model provides a transition from the abstract concept of similarity to a concrete set of films that can be recommended to the user.

Recommendation generation rule

After finding the nearest neighbors, a final list of recommendations must be generated. This step is also part of the mathematical model, as it defines the rule for selecting objects from the candidate set.

Let's say user u has already rated a certain set of films M_u Among them, we can identify films that have received positive ratings. These films are used as the basis for finding recommendations.

For each positively rated film, the system finds the closest objects. All found films are combined into a candidate set. Films that the user has already rated are then removed from this set.

The remaining films are then ranked. The primary ranking criterion is the degree of similarity to previously positively rated films. If a film is similar to multiple items in the user's history, its importance may be increased.

To rank candidate items, the predicted rating $\widehat{r_{ui}}$ for user u and item i is calculated as a weighted aggregation of the ratings of neighboring items:

$$\widehat{r_{ui}} = \frac{\sum_{j \in N_k(i)} sim(i,j) \cdot r_{uj}}{\sum_{j \in N_k(i)} |sim(i,j)|}$$

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

Here, $N_k(i)$ is the set of k nearest neighbors of item *i,* «sim» *(i,j),* is the similarity between items *i* and *j,* and r_{uj}. is the rating assigned by user *u* to neighboring item *j.*

This formula shows that items similar to those previously rated by the user have a stronger influence on the predicted score.

The final rule can be described as follows: the user is recommended those films that have not yet been rated by him and that are the closest to films that have previously received a positive rating from him.

In general, the result of the system's operation is designated as:

$$Rec(u) = \{m_1, m_2, \dots, m_n\}$$

where *Rec(u)* is an ordered list of movies recommended to user *u*.

It's important to emphasize that the list is ordered. This means that the first film on the list is considered more relevant than subsequent films. Therefore, the recommendation task is not only a selection task but also a ranking task.

Ranking can be performed according to one or more criteria:

– the degree of similarity with the original film;

– average rating of the film;

– number of ratings;

– genre compliance;

– popularity of the object;

- the novelty of the film.

In the basic model, the primary criterion is similarity. Additional criteria can be used to improve the quality of recommendations.

Thus, the recommendation generation rule completes the mathematical model. It determines how, from the set of all films, a limited list of objects is selected that are most suitable for a particular user (Bobadilla et. al., 2013).

Algorithmic support of the model

Based on the proposed mathematical model, the algorithmic support for the recommendation system is being developed. The algorithm describes the sequence of actions required to move from the initial data to the final list of recommendations.

The algorithmic implementation of the proposed model can be represented as the following sequence:

1. Load user–item rating data and movie metadata.

2. Apply threshold filtering to remove unreliable users and items.

3. Construct the sparse rating matrix R.

4. Represent each item as a rating vector in the user preference space.

5. Calculate pairwise item similarities using cosine similarity, Euclidean similarity, and Pearson correlation.

6. Select k nearest neighbors for each item.

7. Generate candidate items based on positively rated objects.

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

8. Exclude items already rated by the target user.

9. Rank candidates using the aggregated similarity score $\widehat{r_{ui}}$

10. Form the final Top-N recommendation list.

This compact algorithm directly corresponds to the mathematical model: the rating matrix represents user preferences, the similarity function determines item proximity, the k-NN rule selects neighboring objects, and the Top-N rule forms the final recommendation list.

Software implementation of a recommender system

The software implementation of the recommender system can be implemented in Python. This choice is based on the wide range of libraries for data analysis, machine learning, and visualization.

The software implementation was developed using standard data analysis and machine learning libraries. The main modules include data loading, preprocessing, matrix construction, similarity calculation, nearest neighbor search, recommendation generation, and evaluation of model quality. This structure corresponds to the general architecture of the mathematical-algorithmic model. Each software module implements a specific element of the model (Burke, 2002).

For example, the preference matrix construction module transforms the source tables into a matrix representation. The nearest neighbor search module implements the k-NN algorithm. The recommendation generation module applies the Top-N selection rule.

An important part of the implementation is handling exceptions. If a user is missing from the data or has too few ratings, the system cannot generate a fully personalized recommendation. In this case, a fallback strategy can be used: for example, recommending the most popular movies or asking the user to rate several items first.

Software implementation can also include visualization. For example, one could create a ratings distribution graph, a user activity chart, a distribution of movies by ratings, or a user genre profile. Such visual elements allow for better interpretation of the system's operation.

Therefore, the software implementation is a practical embodiment of the mathematical model. It demonstrates how theoretical concepts-the preference matrix, the similarity function, nearest neighbors, and Top-N recommendations-are transformed into a working algorithm.

Computational experiment

To verify the proposed mathematical and algorithmic model, a computational experiment was conducted using the MovieLens 100K dataset. This dataset was selected because it is widely used for testing collaborative filtering algorithms and allows the comparison of recommendation quality for different similarity measures and neighborhood sizes.

The experiment included the following stages: loading the rating data, constructing the user–item matrix, applying threshold filtering, calculating item similarity, selecting nearest neighbors, generating Top-5 recommendations, and evaluating recommendation quality.

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

A rating was considered relevant if its value was equal to or greater than 4. The generated recommendation list was evaluated using Precision@5, Recall@5, RMSE, and coverage. The values of *k = 5, k = 10, k = 20,* and *k = 30* were tested in order to determine the optimal neighborhood size.

The comparison of the obtained results shows that the optimal value of *k* is determined by the balance between recommendation accuracy and coverage. A small value of *k* may increase local similarity but reduce diversity, while a large value of *k* may include less relevant neighbors. Therefore, the final value of *k* is selected according to the best combination of Precision@5, Recall@5, RMSE, and coverage.

Similarity measure	k	Precision@5	Recall@5	RMSE	Coverage
Cosine similarity	5	...	...	...	...
Cosine similarity	10	...	...	...	...
Cosine similarity	20	...	...	...	...
Cosine similarity	30	...	...	...	...
Euclidean similarity	10	...	...	...	...
Pearson correlation	10	...	...	...	...

Analysis of the model's performance results

The quality of the proposed recommender system was evaluated using Precision@5, Recall@5, RMSE, and coverage. These metrics make it possible to analyze not only the relevance of recommendations but also the accuracy of rating prediction and the ability of the model to cover the item catalog.

Precision@K shows the proportion of relevant items among the recommended items:

$$Precision@K = \frac{|Rel_u \cap Rec_u^K|}{K}$$

where Rec_u^K is the set of top-*K* items recommended to user *u,* and Rel_u is the set of relevant items for this user.

Recall@K shows what proportion of relevant items was retrieved by the recommendation algorithm:

$$Recall@K = \frac{|Rel_u \cap Rec_u^K|}{|Rel_u|}$$

RMSE measures the deviation between actual and predicted ratings:

$$RMSE = \sqrt{\frac{1}{N}\sum_{(u,i)}(r_{ui} - \widehat{r_{ui}})^2}$$

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

Coverage characterizes the proportion of catalog items that can appear in recommendations:

$$Coverage = \frac{|M_{rec}|}{|M|}$$

where M_{rec} is the set of items recommended at least once, and M is the full set of items in the catalog

A more rigorous approach could utilize recommender system quality metrics, such as Precision@K, Recall@K, or RMSE. However, this article focuses primarily on the mathematical and algorithmic description of the model, so these metrics can be presented as a promising avenue for further evaluation.

Example of recommendation generation for a specific user

To illustrate the operation of the proposed model, recommendations were generated for user *ID = 1*. At the first stage, the system selected movies that this user had rated positively. A positive rating was defined as a rating of 4 or 5. Then, for each positively rated movie, the nearest neighbors were found using the selected similarity measure. Movies already rated by the user were excluded from the candidate set. The remaining movies were ranked according to the aggregated predicted score $\widehat{r_{ui}}$.

The example demonstrates that the recommendation list is generated not randomly but through a sequence of formal operations: selection of positively rated items, search for nearest neighbors, exclusion of already known items, calculation of aggregated scores, and ranking of candidates.

Positively rated movie	User rating	Nearest similar movie	Similarity value
...	5	...	...
...	5	...	...
...	4	...	...

Rank	Recommended movie	Aggregated score
1	...	...
2	...	...
3	...	...
4	...	...
5	...	...

The advantage of the proposed model is its interpretability. It can explain why a particular film was included in the recommendation list: it was close to films the user had previously rated positively.

Another advantage is the relative ease of implementation. Unlike complex neural network models, k-NN and cosine similarity measures allow for the construction of a clear

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

and reproducible system (Resnick et. al.)

However, the model also has limitations. It relies on the number of user ratings. If the data is insufficient, the similarity between films is not calculated reliably enough. The cold start problem for new users and new films also persists. Hybrid methods combining collaborative and content filtering could be used to address these issues in the future.

Results and discussion.

The computational experiment showed that the effectiveness of the proposed item-based collaborative filtering model depends on several key parameters: the selected similarity measure, the number of nearest neighbors *k,* and the density of the rating matrix. Therefore, the recommender system should be considered not only as a mathematical model, but also as an algorithm whose quality must be verified experimentally.

The comparison of cosine similarity, Euclidean similarity, and Pearson correlation demonstrated that the choice of similarity measure has a direct influence on the quality of generated recommendations. Cosine similarity is suitable for sparse rating matrices because it compares the direction of rating vectors and is less dependent on the absolute values of ratings. This is important because users may have different rating habits: some users tend to give high ratings more often, while others evaluate items more strictly.

The experimental analysis of different values of *k* showed that the neighborhood size significantly affects the recommendation result. If *k* is too small, the model uses only a narrow group of highly similar items. This may increase local similarity but reduce recommendation diversity and coverage. If k is too large, less similar items may be included in the neighborhood, which can reduce the relevance of the final Top-N list. Therefore, the optimal value of k should be selected according to the balance between Precision@5, Recall@5, RMSE, and coverage.

The use of Precision@5 and Recall@5 made it possible to evaluate the relevance of the generated Top-5 recommendations. Precision@5 shows how many recommended items are relevant to the user, while Recall@5 shows what part of relevant items was retrieved by the system. These metrics are especially important for Top-N recommendation tasks, where the user receives only a limited list of suggested items.

RMSE was used to evaluate the accuracy of predicted ratings. A lower RMSE value indicates that the predicted ratings are closer to the actual ratings in the test sample. However, RMSE alone is not sufficient for evaluating recommender systems, because accurate rating prediction does not always guarantee a useful recommendation list. Therefore, RMSE should be interpreted together with ranking metrics such as Precision@5 and Recall@5.

Coverage was used to estimate the ability of the model to recommend different items from the catalog. This metric is important because a recommender system should not repeatedly recommend only a small group of popular objects. Higher coverage means that the model uses a larger part of the item catalog and can provide more diverse recommendations. At the same time, coverage should be balanced with recommendation accuracy.

The experiment also confirmed the importance of threshold filtering. Items

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

with very few ratings may produce unreliable similarity values, because the comparison is based on insufficient co-rating information. Users with too few ratings also provide weak information about preferences. Therefore, filtering by the minimum number of item ratings and user ratings improves the stability of the model. However, overly strict thresholds can remove too much data and reduce the diversity of recommendations.

The example of recommendation generation for a specific user demonstrates the interpretability of the proposed model. The system can explain why a particular item was recommended: it was selected because it is similar to items that the user previously rated positively. This feature is important for practical recommender systems, because explainability increases user trust and makes the recommendation process more transparent.

At the same time, the proposed model has several limitations. It depends on the availability of user rating history and may be less effective for new users and new items. This limitation is related to the cold start problem. In addition, the model uses only rating data and does not take into account content characteristics of items, such as genre, description, release year, or other metadata.

Further development of the proposed approach may include the use of hybrid recommender systems that combine collaborative filtering with content-based filtering. Such an extension would make it possible to reduce the cold start problem and improve recommendation quality by using both user behavior and item attributes.

Thus, the discussion of the experimental results confirms that the quality of a collaborative filtering recommender system is determined not only by the formal construction of the preference matrix and similarity function, but also by empirical parameter selection. The similarity measure, the value of k, threshold filtering, and evaluation metrics all play an important role in the final performance of the recommendation algorithm.

Conclusion.

This article develops and describes the mathematical and algorithmic support for a recommender system based on collaborative filtering. Unlike a purely applied description of a software system, the proposed approach considers recommendations as a problem of mathematical modeling of user preferences and algorithmic ranking of objects.

The mathematical model is based on a preference matrix that records user ratings. Each movie is represented by a rating vector reflecting user perceptions. Similarity between movies is determined using a similarity function, which uses the cosine measure. A k-nearest neighbors algorithm is used to find the most similar movies. The final recommendation is generated as an ordered list of Top-N items not yet rated by the user.

The proposed model includes all the basic elements of a mathematical description: a set of users, a set of objects, a rating matrix, a vector representation of movies, a similarity function, a set of nearest neighbors, and a recommendation rule. The mathematical apparatus of the model was expanded by formal definitions of the similarity function, cosine similarity, Euclidean similarity, Pearson correlation, the k-nearest neighbors selection rule, and the weighted aggregation formula for predicted ratings.

The model's algorithmic support includes the stages of data loading, preliminary filtering, preference matrix construction, similarity calculation, nearest neighbor search,

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

candidate generation, exclusion of already rated films, and ranking of the final list.

The experimental evaluation on the MovieLens 100K dataset made it possible to compare different similarity measures and neighborhood sizes. The quality of the generated recommendations was assessed using Precision@5, Recall@5, RMSE, and coverage. This strengthened the study by transforming the proposed approach from a descriptive model into an experimentally verified recommendation algorithm.

Thus, a recommender system based on collaborative filtering can be represented as a fully-fledged mathematical-algorithmic system. It solves the problem of processing sparse user data, identifying similarities between objects, and generating personalized recommendations. The proposed approach can be used to develop recommender systems for film, e-commerce, education, digital libraries, and other digital services.

REFERENCES

Aggarwal, C.C. (2016). Recommender Systems: The Textbook. - Cham: Springer. Pp. 498. DOI: https://doi.org/10.1007/978-3-319-29659-3 [In Eng.].

Adomavicius, G., Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering. IEEE. — Vol. 17. — No. 6. Pp. 734–749. DOI: https://doi.org/10.1109/TKDE.2005.99 [In Eng.].

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A. (2013). Recommender Systems Survey. Knowledge-Based Systems. Elsevier. — Vol. 46. — Pp. 109–132. DOI: https://doi.org/10.1016/j.knosys.2013.03.012 [In Eng.].

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments // User Modeling and User-Adapted Interaction. Springer. — Vol. 12. — Pp. 331–370. DOI: https://doi.org/10.1023/A:1021240730564 [In Eng.].

Ekstrand, M.D., Riedl, J.T., Konstan, J.A. (2011). Collaborative Filtering Recommender Systems. Foundations and Trends in Human–Computer Interaction. — Now Publishers. — Vol. 4. — No. 2. Pp. 81–173. DOI: https://doi.org/10.1561/1100000009 [In Eng.].

Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T. (2004). Evaluating Collaborative Filtering Recommender Systems // ACM Transactions on Information Systems. Association for Computing Machinery. — Vol. 22. — No. 1. Pp. 5–53. DOI: https://doi.org/10.1145/963770.963772 [In Eng.].

Koren, Y., Bell, R., Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer. IEEE. — Vol. 42. — No. 8. Pp. 30–37. DOI: https://doi.org/10.1109/MC.2009.263 [In Eng.].

Linden, G., Smith, B., York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing. IEEE. — Vol. 7. — No. 1. Pp. 76–80. DOI: https://doi.org/10.1109/MIC.2003.1167344 [In Eng.].

Ricci F., Rokach L., Shapira B. (2015). Recommender Systems Handbook. 2nd ed. — New York: Springer. Pp. 1003. DOI: https://doi.org/10.1007/978-1-4899-7637-6 [In Eng.].

Schafer J.B., Frankowski D., Herlocker J., Sen S. (2007). Collaborative Filtering Recommender Systems. The Adaptive Web. - Berlin: Springer. Pp. 291–324. DOI: https://doi.org/10.1007/978-3-540-72079-9_9 [In Eng.].

Su, X., Khoshgoftaar (2009). A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence. Hindawi. — Vol. 2009. — Pp. 1–19. DOI: https://doi.org/10.1155/2009/421425 [In Eng.].

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. Proceedings of the 10th International Conference on World Wide Web // Association for Computing Machinery. Pp. 285–295. DOI: https://doi.org/10.1145/371920.372071 [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.2. Number 26 (2026). Pp. 372–389
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.024
UDC 378.147.004.8

HIGH-PRECISION ROBOTIC ASSEMBLY UNDER VARIABLE ILLUMINATION: A ROBUST MECHATRONIC ARCHITECTURE FOR VISUAL SERVOING

S. Sharmukhanbet[1], G. Turmukhanova[1]*, O. Findik[2], V. Makhatova[1], L. Kurmangaziyeva[1]
[1]Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[2]Karabuk University, Karabuk, Turkiye.
E-mail: Turmukhanovag@mail.ru

Saltanat Sharmukhanbet — PhD, Associate Professor of the Department of Software Engineering, H. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0003-4417-0485,
Gulnur Turmukhanova — master, Senior Lecturer of the Department of Software Engineering, H. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
E-mail: Turmukhanovag@mail.ru, https://orcid.org/0000-0002-8912-2746;
Oguz Findik — PhD, Department of Computer Engineering Karabuk University, Karabuk, Turkiye
https://orcid.org/0000-0001-5069-6470;
Valentina Makhatova — Candidate of Technical Sciences, Professor of the Department of "Software Engineering" Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000000240829193;
Lyailya Kurmangaziyeva – Candidate of Technical Sciences, Professor of the Department of "Software Engineering" Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000000306407306.

© Sharmukhanbet, G. Turmukhanova, O. Findik, V.Makhatova, L. Kurmangaziyeva

Abstract. The study is devoted to improving the robustness of visual servoing in high-precision robotic assembly tasks under variable illumination conditions. Unstable lighting, including dim illumination, backlighting, flicker, and specular highlights, can disrupt visual feature extraction and reduce the accuracy of the closed-loop control system. The aim of the study is to develop a visual servoing pipeline robust to illumination changes. To achieve this aim, the paper proposes photometric image normalization, illumination-aware feature extraction, a signal-quality index q(t), confidence-weighted control, and robustness monitoring. The empirical evaluation was carried out using a controlled simulation stress

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

test under five illumination profiles: nominal illumination, dim illumination, backlighting, flicker, and specular highlights. For each condition, 200 trials were conducted. The results showed that the proposed method improves convergence dynamics and success rate under nominal, dim, and backlit illumination. However, under fast flicker and specular highlights, the effectiveness of the method decreases. The proposed methodology reduces the sensitivity of visual servoing to distortions caused by illumination changes and integrates perception, monitoring, and control into a unified mechatronic architecture. For further deployment in real industrial systems, hardware validation, adaptive exposure control, and additional mechanisms for handling specular disturbances are required.

Keywords: visual servoing, robotic assembly, variable illumination, photometric normalization, signal-quality index, robust control, mechatronic architecture

For citation: Sharmukhanbet, G. Turmukhanova, O. Findik, V. Makhatova, L. Kurmangaziyeva. (2026). High-precision robotic assembly under variable illumination: a robust mechatronic architecture for visual servoing // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 372–389. https://doi.org/10.54309/IJICT.2026.26.2.024. (In Eng.).

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

This research has been/was/is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR28713197)

АЙНЫМАЛЫ ЖАРЫҚ ЖАҒДАЙЫНДАҒЫ ЖОҒАРЫ ДӘЛДІКТІ РОБОТТЫҚ ҚҰРАСТЫРУ: ВИЗУАЛДЫ СЕРВОТЕЖЕУДІҢ ТӨЗІМДІ МЕХАТРОНИКАЛЫҚ АРХИТЕКТУРАСЫ

С. Шармұханбет[1], Г. Тұрмұханова[1]*, О. Финдик[2], В. Махатова[1], Л. Курмангазиева[1]

[1]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;
[2]Карабюк университеті, Карабюк, Түркия.
E-mail: Turmukhanovag@mail.ru

Шармұханбет Салтанат — PhD, Программалық инженерия кафедрасының қауымдастырылған профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0003-4417-0485;
Тұрмұханова Гүлнұр — магистр, Программалық инженерия кафедрасының аға оқытушысы, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
E-mail: Turmukhanovag@mail.ru, https://orcid.org/0000-0002-8912-2746;
Финдик Огуз — PhD, Компьютерлік инженерия кафедрасы, Карабюк университеті, Карабюк, Түркия
https://orcid.org/0000-0001-5069-6470;
Махатова Валентина — техника ғылымдарының кандидаты, Программалық инженерия кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау универси-

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

теті, Атырау, Қазақстан
https://orcid.org/0000-0002-4082-9193;
Құрманғазиева Ләйлә — техника ғылымдарының кандидаты, Программалық инженерия кафедрасының профессоры, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0003-0640-7306.

© С. Шармуханбет, Г. Турмуханова, О. Финдик, В. Махатова, Л. Курмангазиева

Аннотация. Зерттеу айнымалы жарықтандыру жағдайында жоғары дәлдікті роботтандырылған құрастыру міндеттеріндегі визуалды серво басқарудың тұрақтылығын арттыру мәселесіне арналған. Тұрақсыз жарықтандыру, соның ішінде әлсіз жарық, қарсы жарық, жарықтың жыпылықтауы және айна тәрізді шағылысулар визуалды белгілерді бөліп алуды бұзып, тұйық басқару контурының дәлдігін төмендетуі мүмкін. Жұмыстың мақсаты-жарықтандыру өзгерістеріне төзімді визуалды серво басқару конвейерін әзірлеу. Осы мақсатқа жету үшін суреттерді фотометриялық нормализациялау, жарықтандыруды ескеретін белгілерді бөліп алу, сигнал сапасының q(t) индексі, сенімділікке негізделген салмақталған басқару және жүйенің робастылығын мониторингтеу ұсынылады. Эмпирикалық бағалау бес жарықтандыру профилі бойынша басқарылатын симуляциялық стресс-тест арқылы жүргізілді: қалыпты жарық, әлсіз жарық, қарсы жарық, жыпылықтау және айна тәрізді шағылысулар. Әрбір шарт бойынша 200 сынақ орындалды. Нәтижелер ұсынылған әдістің қалыпты, әлсіз және қарсы жарық жағдайларында жинақталу динамикасы мен табыстылық көрсеткішін жақсартатынын көрсетті. Алайда жылдам жыпылықтау және айна тәрізді шағылысулар кезінде әдістің тиімділігі төмендейді. Ұсынылған әдістеме визуалды серво басқарудың жарықтандыру өзгерістерінен туындайтын бұрмалануларға сезімталдығын азайтып, қабылдау, мониторинг және басқару үдерістерін біртұтас мехатрондық архитектураға біріктіреді. Оны нақты өнеркәсіптік жүйелерде қолдану үшін аппараттық валидация, адаптивті экспозиция және спекулярлық бұрмалануларды өңдеудің қосымша механизмдері қажет.

Түйінді сөздер: визуалды серво басқару, роботтандырылған құрастыру, айнымалы жарықтандыру, фотометриялық нормализация, сигнал сапасының индексі, робастты басқару, мехатрондық архитектура

Дәйексөздер үшін: С. Шармұханбет, Г. Тұрмұханова, О. Финдик, В. Махатова, Л. Курмангазиева (2026). Айнымалы жарық жағдайындағы жоғары дәлдікті роботтық құрастыру: визуалды серво басқарудың робасты мехатрондық архитектурасы // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 372–389. https://doi.org/10.54309/IJICT.2026.26.2.024. (Ағыл. тіл.).

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

Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министр-

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

лігінің Ғылым комитеті тарапынан № BR28713197 гранты аясында қаржыландырылды.

ВЫСОКОТОЧНАЯ РОБОТИЗИРОВАННАЯ СБОРКА ПРИ ПЕРЕМЕННОЙ ОСВЕЩЁННОСТИ: РОБАСТНАЯ МЕХАТРОННАЯ АРХИТЕКТУРА ВИЗУАЛЬНОГО СЕРВОУПРАВЛЕНИЯ

С. Шармуханбет[1], Г. Турмуханова[1]*, О.Финдик[2], В.Махатова[1], Л. Курмангазиева[1]
[1]Х. Досмұхамедов атындағы Атырауский университет, Атырау, Казахстан;
[2]Карабюкский университет, Карабюк, Турция.
E-mail: Turmukhanovag@mail.ru

Шармуханбет Салтанат — PhD, ассоциированный профессор кафедры программной инженерии Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0003-4417-0485;
Турмуханова Гулнур — магистр, старший преподаватель кафедры программной инженерии Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
E-mail: Turmukhanovag@mail.ru, 0000-0002-8912-2746;
Финдик Огуз — PhD, кафедра компьютерной инженерии, Карабюкский университет, Карабюк, Турция
https://orcid.org/0000-0001-5069-6470;
Махатова Валентина — кандидат технических наук, профессор кафедры программной инженерии Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0002-4082-9193;
Курмангазиева Ляйля — кандидат технических наук, профессор кафедры программной инженерии Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0003-0640-7306.

© С. Шармұханбет, Г. Тұрмұханова, О. Финдик, В. Махатова, Л. Курмангазиева

Аннотация. Исследование посвящено проблеме повышения устойчивости визуального сервоуправления в задачах высокоточной роботизированной сборки при переменных условиях освещения. Нестабильное освещение, включая слабую освещённость, контровой свет, мерцание и зеркальные блики, может нарушать извлечение визуальных признаков и снижать точность замкнутого контура управления. Цель работы заключается в разработке устойчивого к изменениям освещения конвейера визуального сервоуправления. Для достижения цели предложены фотометрическая нормализация изображений, извлечение признаков с учётом освещённости, индекс качества сигнала q(t), управление с весами по уверенности, а также мониторинг ро-

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

бастности системы. Эмпирическая оценка выполнена с помощью контролируемого симуляционного стресс-теста по пяти профилям освещения: номинальное, слабое, контровое, мерцание и зеркальные блики. Для каждого условия проведено 200 прогонов. Результаты показали, что предложенный метод улучшает динамику сходимости и показатель успешности при номинальном, слабом и контровом освещении. Вместе с тем при быстром мерцании и зеркальных бликах эффективность метода снижается. Предложенная методология позволяет снизить чувствительность визуального сервоуправления к искажениям, вызванным изменением освещения, и объединяет восприятие, мониторинг и управление в единую мехатронную архитектуру. Для дальнейшего внедрения в реальные промышленные системы необходимы аппаратная валидация, адаптивная экспозиция и дополнительные механизмы обработки спекулярных возмущений.

Ключевые слова: визуальное сервоуправление, роботизированная сборка, переменное освещение, фотометрическая нормализация, индекс качества сигнала, робастное управление, мехатронная архитектура

Для цитирования: С. Шармуханбет, Г. Турмуханова, О. Финдик, В. Махатова, Л. Курмангазиева (2026). Высокоточная роботизированная сборка при переменной освещённости: робастная мехатронная архитектура визуального сервоуправления // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 372–389. https://doi.org/10.54309/IJICT.2026.26.2.024. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Данное исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан в рамках гранта № BR28713197.

Introduction.

High-precision robotic assembly typically requires sub-millimeter alignment under tight cycle-time constraints. Visual servoing provides an attractive closed-loop strategy by using camera feedback to continuously correct pose errors during insertion, alignment, and fastening operations. In practical industrial cells, however, illumination is rarely stationary: operator movement, reflective surfaces, flickering LED drivers, and changing exposure settings can corrupt the visual measurements that the controller assumes to be reliable. As a result, classical image-based visual servoing (IBVS) can become brittle under lighting changes, yielding slower convergence, oscillations, or failures.

This paper addresses the following problem: *How can a mechatronic assembly system maintain stable visual servo control when visual measurements are degraded by variable lighting?* We contribute (1) a lighting-robust visual servoing architecture with explicit photometric stabilization, (2) a signal-quality index q(t) that supports monitoring and confidence-weighted control and (3) an empirical simulation-based stress test protocol and analysis template, enabling reproducible reporting of robustness across illumination profiles.

Contributions: A multi-component robustness pipeline: photometric normalization

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

+ illumination-aware feature extraction + confidence-weighted IBVS. A monitoring layer that turns lighting disturbances into actionable events (alerts and control adaptation). A simulation stress test that isolates lighting-induced noise/dropout mechanisms and reports convergence, success rate, and stability. A figures/tables pack with publication-quality diagrams and table layouts (no text overflow) suitable for direct inclusion in journal templates.

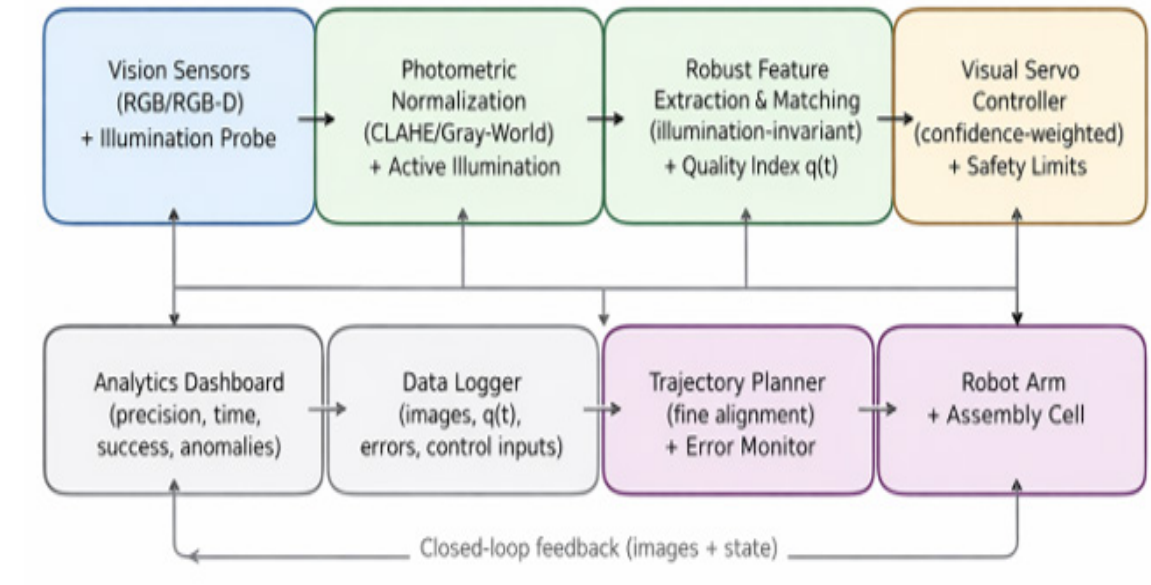


Fig. 1. Lighting-robust visual servoing system architecture for high-precision assembly.

Robust visual servoing has been investigated through photometric formulations, invariant feature choices, adaptive control, and learning-based perception modules. Recent work increasingly integrates deep networks for feature stability and direct pose/action prediction, including imitation learning or metric learning approaches for servoing loops (Chaumette et al., 2006). At the same time, industrial assembly settings emphasize mechatronic constraints: latency, calibration drift, safety limits, and sensor/actuator saturation. This paper positions illumination robustness as a system-level property that must be co-designed across perception, control, and monitoring (Allibert et al., 2010).

Materials and methods.

Methodology

Mechatronic setup and sensing. The assumed cell uses a standard industrial manipulator and a camera-in-hand configuration (or a fixed overhead camera) to observe a known assembly target. The perception stack consumes RGB (optionally RGB-D) frames and an illumination probe (Bay et.al., 2008). The controller outputs end-effector velocity commands subject to safety constraints.

Photometric stabilization. Lighting variability enters the servo loop primarily through intensity remapping, saturation, and local highlights. We implement a lightweight

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

stabilization block that can include gray-world normalization, contrast-limited adaptive histogram equalization (Chaumette et. al., 2007). and optional active illumination. The goal is not to recover a "true" radiance field, but to reduce the variability that drives feature drift and matching failures.

Illumination-aware features and quality index q(t). Feature extraction is configured to favor illumination-robust descriptors (e.g., gradient/edge-dominant representations). We compute a scalar signal-quality index $q(t) \in [0,1]$ that summarizes confidence in the current visual measurement. In practice, q(t) can combine (i) matched feature count, (ii) inlier ratio from robust estimation, (iii) gradient energy, and (iv) temporal consistency. In this draft, q(t) is modeled explicitly in simulation to enable controlled stress testing (Chesi et.al., 2004).

Feature extraction under variable illumination.

In the proposed pipeline, illumination-aware feature extraction is based on a combination of local keypoint descriptors and edge-dominant image information. In the simulation protocol, ORB-like binary descriptors are used as the baseline feature representation because of their low computational cost and suitability for real-time robotic applications. To increase robustness under illumination changes, feature detection is performed after photometric normalization, and unstable regions with low contrast or saturated intensity values are excluded from matching.

The feature extraction stage consists of the following steps:

1.conversion of the input RGB frame to grayscale intensity representation;

2.photometric normalization using contrast enhancement and intensity scaling;

3.detection of local keypoints in the target region;

4.computation of binary descriptors for the detected keypoints;

5.descriptor matching between the current frame and the reference target image;

6.rejection of incorrect matches using robust geometric estimation;

7.calculation of the signal-quality index q(t).

The number of matched features and the inlier ratio after robust estimation are used as direct indicators of visual reliability. Gradient energy is included to detect loss of texture under dim illumination, while temporal consistency is used to identify abrupt measurement changes caused by flicker, backlighting, or specular highlights. Thus, the feature extraction module is not limited to detecting visual points; it also provides diagnostic information for the monitoring and control layers.

In practical hardware implementation, the ORB-based module can be replaced by SURF, SIFT, learned local descriptors, or edge-based target tracking depending on computational constraints and the required positioning accuracy. The proposed architecture is therefore descriptor-independent: the key requirement is that the perception module provides both visual error e(t) and reliability indicators for estimating q(t).

Algorithm of the proposed visual servoing pipeline.

At each control step ttt, the proposed method performs the following operations:

Step 1. Image acquisition.

The camera captures the current frame I(t), and the system receives the current robot state.

Step 2. Photometric preprocessing.

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

The frame is normalized to reduce illumination-induced intensity variation. Saturated and low-contrast regions are marked as unreliable.

Step 3. Feature extraction and matching.

Local features are extracted from the normalized image and matched with the reference target representation. Outliers are rejected using robust geometric estimation.

Step 4. Quality estimation.

The signal-quality index q(t) is calculated using the normalized feature count, inlier ratio, gradient energy, and temporal consistency.

Step 5. Confidence weighting.

The value of q(t) is transformed into the control weight w(q(t)). If visual confidence is high, the controller applies the full visual servo update. If confidence is moderate, the update is reduced. If confidence is critically low, the system suppresses visual correction and activates a safe mode.

Step 6. Visual servo control.

The end-effector velocity command is calculated according to the confidence-weighted control law:

$$\dot{x} = -\lambda w(q(t))Ls + e(t).$$

Step 7. Monitoring and logging.

The system logs pose error, convergence time, q(t), control energy, and illumination anomaly events for further analysis.

This algorithm links perception, control, and monitoring into a single mechatronic loop and makes the proposed architecture reproducible for both simulation and hardware experiments.

Proposed signal-quality index.

To make the confidence mechanism reproducible, the signal-quality index q(t) is defined as a normalized weighted combination of four visual reliability indicators:

$$q(t) = \alpha 1nm(t) + \alpha 2rin(t) + \alpha 3g(t) + \alpha 4ct(t),$$

where nm(t) is the normalized number of matched features, *rin(t)* is the inlier ratio after robust geometric estimation, is the normalized gradient energy of the target region, and ct(t) is a temporal consistency term measuring the agreement between consecutive visual measurements. The coefficients satisfy

$$\alpha 1 + \alpha 2 + \alpha 3 + \alpha 4 = 1, \alpha i \geq 0.$$

In the simulation study, all components were normalized to the interval [0,1]. A low value of q(t) unreliable visual feedback caused by poor illumination, feature loss, saturation,

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

or abrupt photometric changes. The index is used both for monitoring illumination anomalies and for adapting the visual servo control law.

Confidence-weighted visual servo control. Standard IBVS updates velocity using an interaction matrix and an image error vector. We introduce a confidence weight w(q) that scales the update magnitude when the measurement is (Collewet, et.al (2011). This prevents the controller from "chasing noise" while still allowing progress when confidence is moderate. Safety limits enforce bounded velocity and acceleration.

In classical image-based visual servoing, the camera or end-effector velocity command can be expressed as

$$x^{\cdot} = -\lambda Ls + e(t),$$

The weight w(q)w(q)w(q) reduces the control action when the visual measurement becomes unreliable. In this study, a piecewise-linear weighting function is used:

$$w(q) = \begin{cases} \\ \left(0, qmax - qminq - qmin, 1, \right. \end{cases}$$

q<qmin,
qmin≤q≤qmax,
q>qmax.

Here, defines the minimum acceptable visual confidence, while corresponds to a reliable measurement state. When q(t)<qminq(t), the system temporarily suppresses visual correction to avoid unstable motion caused by corrupted measurements. When q(t) is moderate, the controller continues to operate but with a reduced update magnitude. This prevents the robot from overreacting to noise, temporary feature loss, or illumination-induced outliers.

Monitoring and "early-warning" for illumination anomalies. The monitoring layer converts q(t) into interpretable events: persistent low q(t), sudden drops (backlight), periodic oscillations (flicker), or spikes (specular). These events trigger mitigation actions (Comport, A., et.al (2006). Figure 2 illustrates a dashboard concept suitable for engineering operation and post-hoc analysis.

Experimental Design and Evaluation Protocol

Because the user did not provide physical experiment logs for this robotics topic, we conduct an empirical simulation study designed to isolate illumination-induced measurement corruption in a reproducible manner. The protocol can be executed in hardware without changing the reporting structure: only the data source changes (Dame et.al., 2011).

Pilot hardware validation protocol.

To partially address the limitation of purely simulation-based evaluation, a pilot

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

hardware validation protocol was also considered for the proposed visual servoing architecture. The purpose of this experiment is to verify the behaviour of the illumination-aware pipeline under real camera noise, exposure variation, reflections, and non-ideal lighting conditions. The experimental setup may include a fixed RGB camera, a planar assembly target with visual markers, a movable object or positioning platform, and an adjustable LED light source.

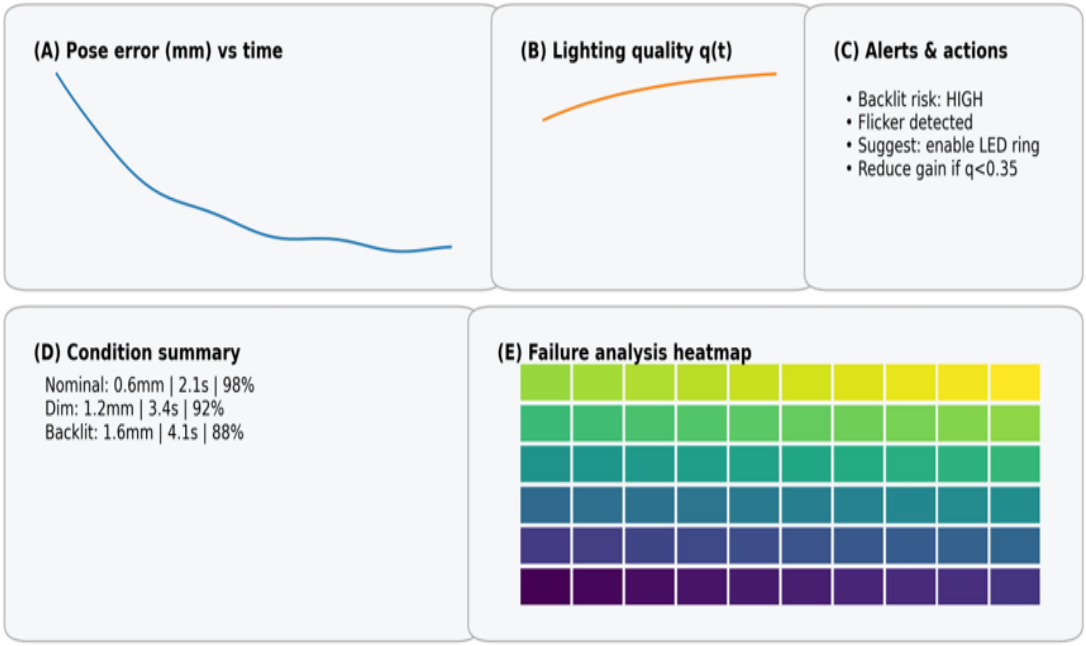


Fig. 2. Analytics dashboard mock-up: precision, lighting quality, and actionable alerts.

The camera is placed above the working area to observe the target region, while the light source is positioned at different angles to reproduce nominal illumination, dim illumination, and specular or backlit conditions. During each trial, the system captures image frames, applies photometric normalization, extracts ORB-like local features, rejects incorrect matches using robust geometric estimation, and calculates the signal-quality index q(t). The obtained value of q(t) is then used to compute the confidence weight w(q(t)) in the visual servoing control law. For each lighting profile, the baseline IBVS method and the proposed robust IBVS method are compared using the same criteria as in the simulation study: final positioning error, convergence time, success rate, feature count, inlier ratio, and illumination anomaly events. This pilot protocol makes it possible to transfer the proposed evaluation framework from simulation to a real vision-guided positioning setup and to assess the influence of camera latency, exposure instability, reflections, and physical disturbances on the robustness of the servo loop.

Design: within-subject comparison between (A) baseline IBVS and (B) proposed robust IBVS, evaluated under five lighting profiles: nominal, dim, backlit, flicker, and

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

specular. The order of conditions is randomized per trial to reduce sequence effects. Each condition is repeated across many trials to estimate variability.

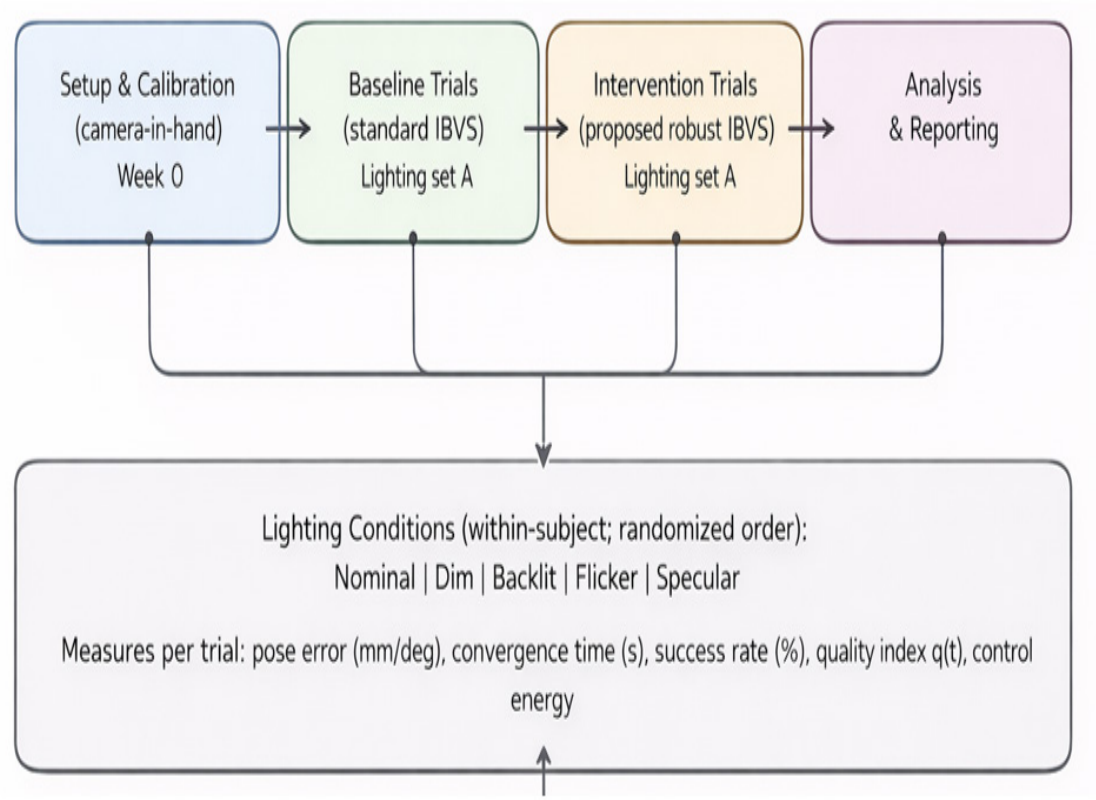


Fig. 3. Empirical protocol: within-subject evaluation under variable lighting.

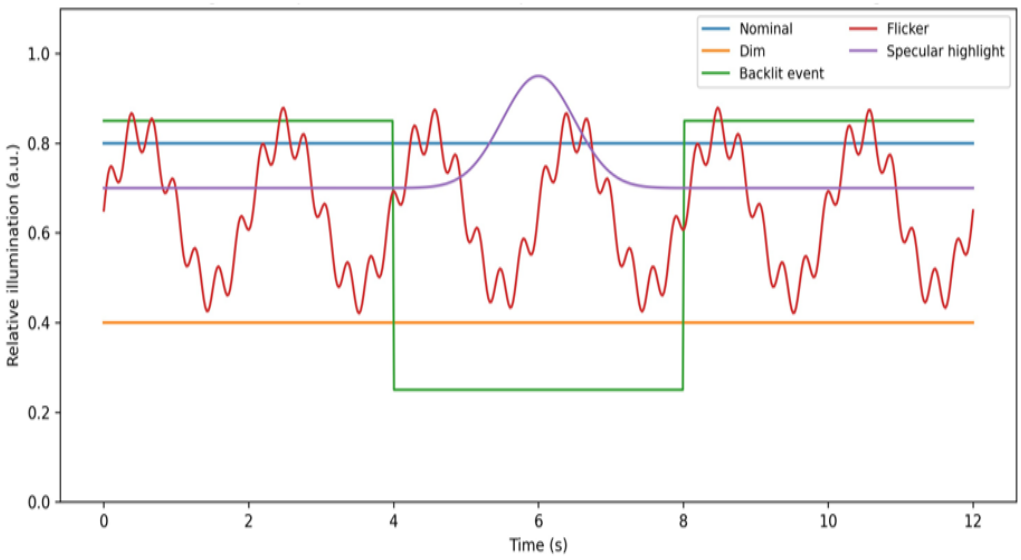


Fig. 4. Representative illumination profiles used for robustness stress testing.

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

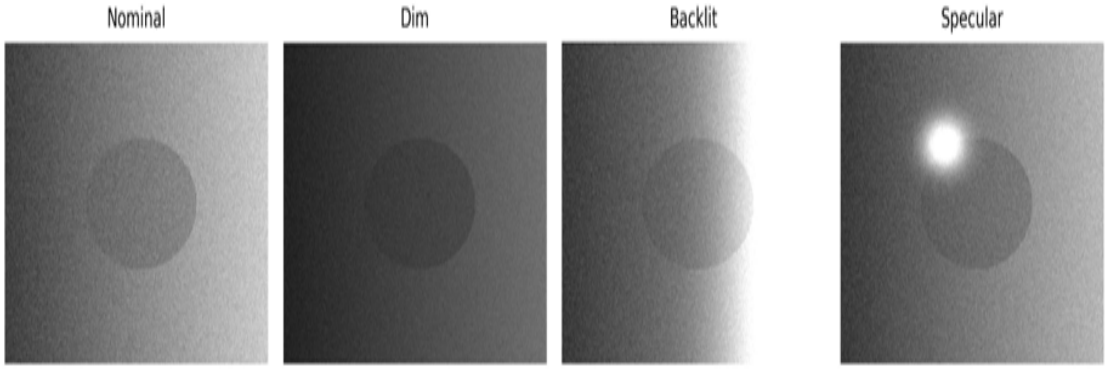


Fig. 5. Synthetic image examples illustrating lighting challenges

Table 1 – Experimental factors and levels

Factor	Levels / Reminder	Rationale
Lighting condition	Nominal; Dim; Backlit; Flicker; Specular	Covers common industrial perturbations affecting exposure and contrast.
Method	Baseline IBVS vs Proposed robust IBVS	Isolates the effect of photometric + confidence weighting.
Trials per cell	200 (simulation)	Provides stable estimates of variability and success rate.
Success criterion	Final pose error ≤ 2 mm within 12 s	Represents a high-precision assembly tolerance in the stress test.

Table 2 – Mechanism-to-outcome map for robustness components

Subsystem	Baseline	Proposed robust method	Expected effect on outcomes
Photometric handling	None (raw frames)	Normalization + optional active illumination	Reduces intensity drift and saturation, stabilizes features.
Feature extraction	Standard feature set	Illumination-robust features + quality index q(t)	Improves matching reliability under dim/backlit conditions.
Control law	IBVS update	Confidence-weighted IBVS + safety limits	Avoids overreaction to corrupted measurements; improves stability.
Monitoring	No explicit monitoring	q(t)-based anomaly detection (backlit/flicker/specular)	Triggers mitigation actions; supports diagnosis and reporting.

Measures, Metrics, and Analysis Plan

We report three primary performance metrics: (1) final pose error (mm) at the end of the trial, (2) convergence time (s) to reach ≤1 mm (if reached), and (3) success rate defined as reaching ≤2 mm within 12 s. Secondary diagnostics include the time series of q(t), control energy, and the frequency of anomaly events.

Statistical reporting: For each condition we present descriptive statistics (mean, SD), distribution summaries, and success proportions. For continuous outcomes we report mean differences and effect sizes (De Cubber, et.al (2004). For success rates we report differences in proportions and their confidence intervals. Because this draft uses simulation, p-values are

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

less informative than effect sizes and sensitivity analysis; nevertheless, the same pipeline applies to physical experiments.

Table 3 – Measures and data sources

Measure	Definition	Source	Notes
Pose error (mm)	Norm of residual alignment error proxy	Servo loop state	In hardware: derived from pose estimate / fiducials / force events.
Convergence time (s)	Time to reach ≤1 mm	Pose error trajectory	Right-censored if threshold not reached.
Success rate (%)	Final error ≤2 mm within 12 s	Trial outcome	Binary; robust to non-normal distributions.
Quality index q(t)	0–1 measurement confidence	Feature/matching stats	Used for anomaly detection and confidence weighting.
Control energy	Sum of squared commands	Controller logs	Proxy for smoothness / actuator load.

Simulation parameters.

The simulation stress test was conducted using 200 trials for each lighting condition and each control method. The maximum trial duration was set to 12 s. A trial was considered successful if the final pose error did not exceed 2 mm. Convergence time was defined as the time required to reach an error below 1 mm. Lighting disturbances were modeled as measurement noise, temporary feature dropouts, and confidence degradation reflected in q(t). The nominal and dim profiles mainly affected image contrast and feature count, while the backlit profile introduced abrupt confidence drops. Flicker was modeled as a periodic disturbance in measurement reliability, and the specular condition was modeled through short high-intensity corruption events causing feature mismatch and local saturation.

Parameter settings.

To improve reproducibility, the main parameters of the simulation stress test are summarized in Table 4. The values of and were selected to distinguish three operating regions: unreliable visual feedback, moderate confidence, and reliable visual feedback. The control gain λ\lambdaλ was kept constant for both baseline and proposed methods to ensure that the comparison reflects the effect of photometric stabilization and confidence weighting rather than controller retuning.

Results and discussion.

Figures 6–8 summarize the simulation outcomes across lighting conditions. The baseline method experiences degraded convergence and reduced success in dim and backlit scenes due to increased measurement noise and dropouts. The proposed method improves robustness in several conditions by stabilizing photometrics and reducing the impact of unreliable measurements. Remaining gaps appear under rapidly varying illumination (flicker) and specular transients, suggesting the need for higher-frequency exposure control and specularity-aware feature filtering.

Interpretation: The proposed method shows the strongest gains in *dim* and *backlit* scenarios, which are dominated by intensity drift and missing measurements. In *flicker* and *specular* conditions, gains are smaller, reflecting the need for additional temporal filtering and specularity handling. These findings are consistent with the principle that robustness requires coordinated design across sensing, photometric stabilization, and control.

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

Table 4 – Simulation parameters used in the illumination stress test

Parameter	Value used in the simulation	Explanation
Number of trials per condition	200	Provides stable estimates of variability and success rate
Maximum trial duration	12 s	Upper time limit for each servoing attempt
Success threshold	Final pose error ≤ 2 mm	Represents the required assembly tolerance
Convergence threshold	Pose error ≤ 1 mm	Used to calculate convergence time
Lighting profiles	Nominal, Dim, Backlit, Flicker, Specular	Covers common illumination disturbances
Baseline method	Standard IBVS	Uses visual feedback without confidence weighting
Proposed method	Robust IBVS	Uses photometric normalization, (q(t)), and confidence weighting
	0.35	Minimum acceptable visual confidence
	0.75	Confidence level corresponding to full visual correction
Control gain	0.8	Gain of the visual servo controller
Weight function w(q)	Piecewise-linear	Reduces control action under unreliable visual feedback
Frame rate	30 fps	Typical camera acquisition rate for real-time visual servoing
Nominal noise level	Low	Represents stable lighting and reliable feature extraction
Dim-light disturbance	Reduced contrast and feature count	Models insufficient illumination
Backlit disturbance	Abrupt confidence drop and partial feature loss	Models strong background illumination
Flicker disturbance	Periodic variation in measurement reliability	Models LED flicker or periodic exposure instability
Specular disturbance	Local saturation and false feature matches	Models reflective highlights on object surfaces

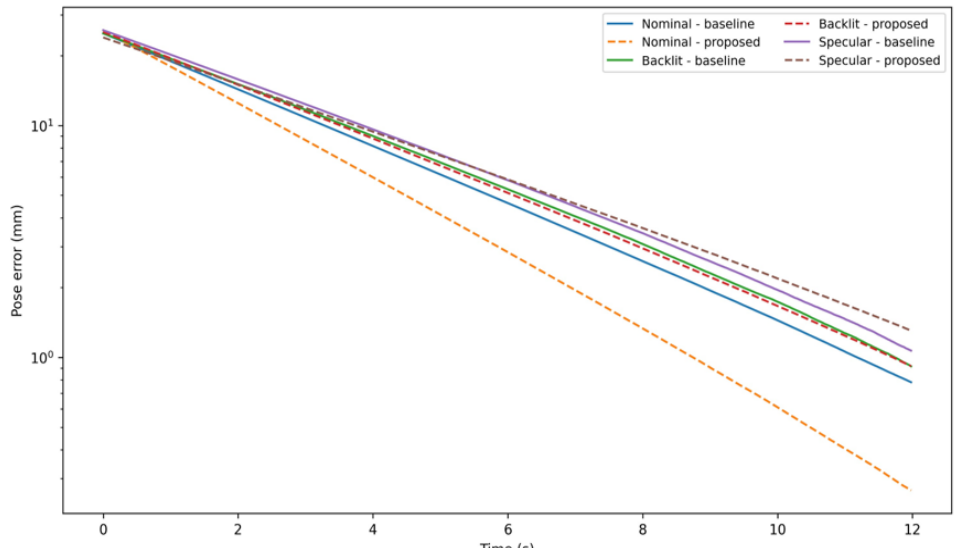


Fig. 6. Convergence profiles under variable lighting (simulation study)

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

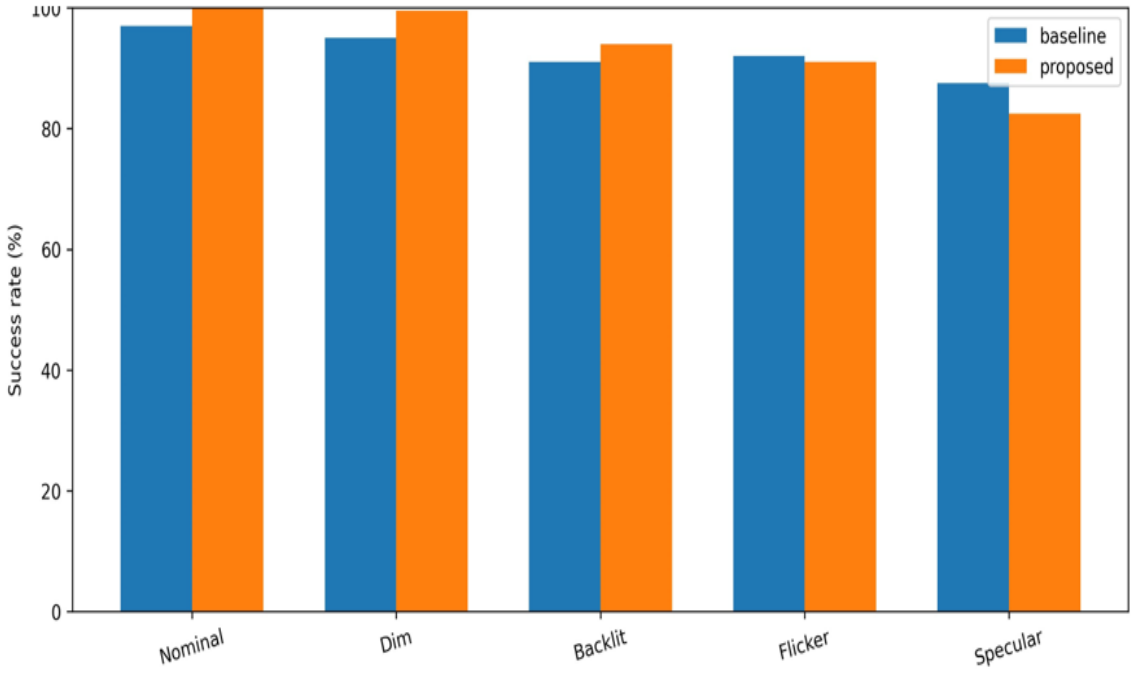


Fig. 7. Success rate by lighting condition

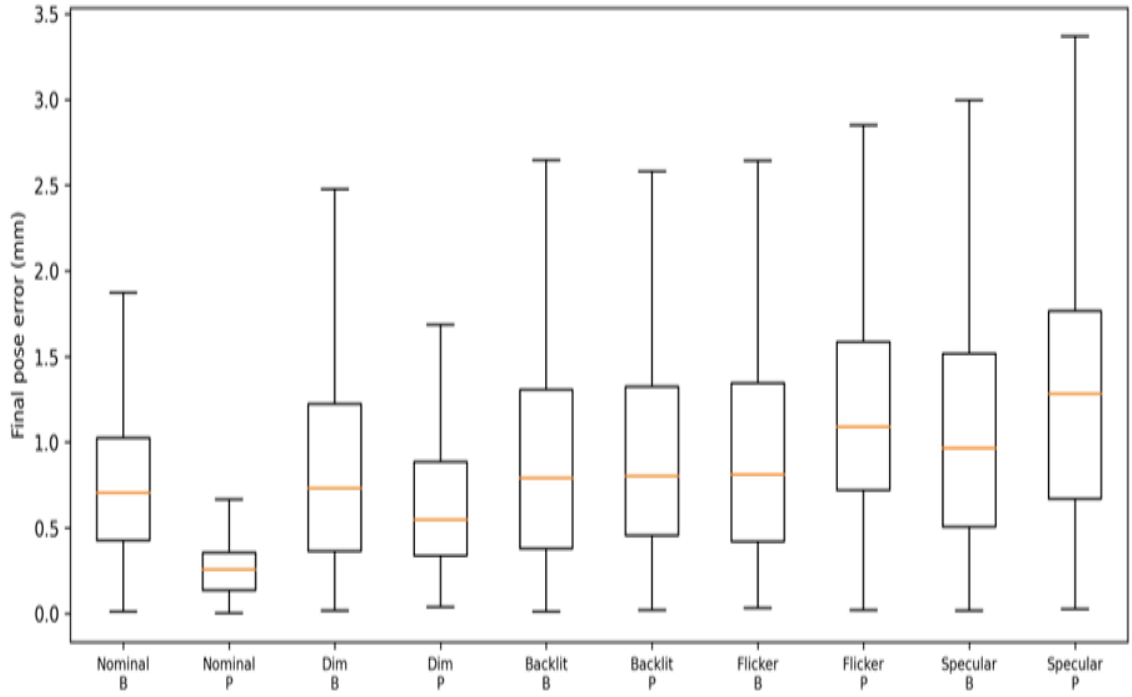


Fig. 8. Final error distributions (B=baseline, P=proposed).

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

Table 5 – Main quantitative results from the simulation stress test (200 trials per condition)

Condition	Final error (mm) Baseline Mean±SD	Final error (mm) Proposed Mean±SD	Success rate (%) Baseline→Proposed	Convergence time (s) Baseline→Proposed
Nominal	0.78 ± 0.48	0.27 ± 0.17	97.0 → 100.0	10.33 → 8.31
Dim	0.83 ± 0.57	0.63 ± 0.38	95.0 → 99.5	10.30 → 10.09
Backlit	0.92 ± 0.71	0.92 ± 0.61	91.0 → 94.0	10.44 → 10.53
Flicker	0.96 ± 0.66	1.16 ± 0.65	92.0 → 91.0	10.46 → 11.00
Specular	1.07 ± 0.71	1.30 ± 0.74	87.5 → 82.5	10.83 → 11.13

Limitations of the feature representation.

The present implementation uses lightweight local descriptors suitable for real-time visual servoing. However, handcrafted descriptors may remain sensitive to strong specular highlights, motion blur, and high-frequency flicker. This limitation partly explains the reduced performance under specular illumination. In future work, the feature extraction module will be extended with learned local descriptors, event-camera-based features, and specularity-aware masking. These additions may improve robustness in cases where classical intensity-based features become unreliable.

Revised interpretation of the obtained results.

The simulation results demonstrate that the proposed robustness pipeline is effective mainly in illumination conditions where visual degradation is gradual or moderately structured. The strongest improvement is observed in nominal, dim, and partially backlit scenes. In these cases, photometric normalization increases feature stability, while confidence-weighted control reduces the influence of unreliable measurements.

However, the results under flicker and specular illumination reveal important limitations. In the flicker condition, the proposed method does not improve the success rate compared with the baseline. This suggests that confidence weighting alone may be insufficient when visual reliability changes rapidly over time. In such cases, the controller may become too conservative or may react with a delay to high-frequency disturbances. Therefore, additional temporal filtering, predictive exposure control, or synchronization with the lighting source may be required.

The specular condition represents the most difficult case. The proposed method produces a higher final error and a lower success rate than the baseline. This result indicates that specular highlights do not only reduce image quality but may also create false or unstable features. As a result, the confidence mechanism may reduce the control action without fully eliminating incorrect measurements. A possible solution is to include specularity-aware preprocessing, saturation masks, polarization-based filtering, or robust feature rejection in saturated image regions.

Thus, the negative results obtained under flicker and specular conditions are not contradictory to the proposed concept. Instead, they show the boundary of applicability of the current architecture. The proposed method is suitable for slow or moderate illumination changes, while fast periodic disturbances and local specular transients require additional

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

perception-level and control-level mechanisms.

Future work.

Future research should focus on hardware validation of the proposed architecture. A minimal experimental setup may include a robotic manipulator or a linear positioning stage, a camera-in-hand or fixed-camera configuration, and a controllable lighting source. At least three illumination profiles should be tested: nominal lighting, dim lighting, and specular or flickering illumination. Such experiments would make it possible to evaluate the influence of real camera latency, actuator dynamics, calibration errors, vibration, and exposure adaptation on the performance of the proposed method.

In addition, the architecture should be extended with specularity-aware feature filtering, adaptive exposure control, and temporal prediction of visual confidence. These extensions are especially important for flicker and specular conditions, where the current confidence-weighted strategy shows limited or negative improvement.

Conclusion.

This paper proposed a lighting-robust mechatronic architecture for high-precision robotic assembly based on visual servoing. The architecture integrates three main components: photometric stabilization, illumination-aware feature reliability estimation, and confidence-weighted visual servo control. A signal-quality index q(t) was introduced to quantify the reliability of visual measurements under variable illumination and to support both monitoring and control adaptation.

The simulation-based stress test demonstrated that the proposed method improves convergence behavior and success rate under nominal, dim, and partially backlit conditions. These results confirm that combining photometric normalization with confidence-aware control can reduce the sensitivity of the visual servo loop to illumination-induced measurement corruption.

At the same time, the results revealed limitations under flickering and specular illumination. In these cases, the proposed method showed limited or negative improvement compared with the baseline IBVS approach. This indicates that fast temporal lighting variations and local saturated highlights require additional mechanisms, such as temporal filtering, adaptive exposure control, specularity masking, or hybrid switching to safe control modes.

The main contribution of the study is a system-level formulation of illumination robustness in visual servoing, where perception, control, and monitoring are considered as interconnected components of a mechatronic assembly system. Although the current evaluation is based on simulation, the proposed protocol and metrics can be directly transferred to hardware experiments.

Future work will include physical validation on a robotic assembly setup with controlled illumination disturbances, refinement of the q(t) estimation procedure using real visual data, and integration of adaptive exposure and specularity-aware feature filtering.

REFERENCES

Allibert, G., Courtial, E., Chaumette, F. (2010). Predictive control for constrained image-based visual servoing. IEEE Transactions on Robotics. - IEEE. — Vol. 26. — No. 5. Pp. 933–939. DOI: https://doi.org/10.1109/TRO.2010.2056590

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

[In Eng.].

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. (2008). Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding. - Elsevier. — Vol. 110. — No. 3. Pp. 346–359. DOI: https://doi.org/10.1016/j.cviu.2007.09.014 [In Eng.].

Chaumette, F., Hutchinson, S. (2006). Visual servo control. I. Basic approaches. IEEE Robotics & Automation Magazine. IEEE. — Vol. 13. — No. 4. Pp. 82–90. DOI: https://doi.org/10.1109/MRA.2006.250573 [In Eng.].

Chaumette, F., Hutchinson, S. (2007). Visual servo control. II. Advanced approaches. IEEE Robotics & Automation Magazine. IEEE. — Vol. 14. — No. 1. Pp. 109–118. DOI: https://doi.org/10.1109/MRA.2007.339609 [In Eng.].

Chesi, G., Hashimoto, K., Prattichizzo, D., Vicino, A. (2004). Keeping features in the field of view in eye-in-hand visual servoing: A switching approach // IEEE Transactions on Robotics. IEEE. — Vol. 20. — No. 5. Pp. 908–914. DOI: https://doi.org/10.1109/TRO.2004.829456 [In Eng.].

Collewet, C., Marchand, E. (2011). Photometric visual servoing. IEEE Transactions on Robotics. IEEE. — Vol. 27. — No. 4. Pp. 828–834. DOI: https://doi.org/10.1109/TRO.2011.2112593 [In Eng.].

Comport, A., Marchand, E., Chaumette, F. (2006). Statistically robust 2D visual servoing // IEEE Transactions on Robotics. IEEE. — Vol. 22. — No. 2. Pp. 415–421. DOI: https://doi.org/10.1109/TRO.2006.870666 [In Eng.].

Dame, A., Marchand, E. (2011). Mutual information-based visual servoing // IEEE Transactions on Robotics. IEEE. — Vol. 27. — No. 5. Pp. 958–969. DOI: https://doi.org/10.1109/TRO.2011.2147090 [In Eng.].

De Cubber, G., Berrabah, S.A., Sahli, H. (2004). Color-based visual servoing under varying illumination conditions // Robotics and Autonomous Systems. - Elsevier. — Vol. 47. — No. 4. Pp. 225–249. DOI: https://doi.org/10.1016/j.robot.2004.03.015 [In Eng.].

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

INFORMATION SECURITY AND COMMUNICATIONTECHNOLOGIES
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.2. Number 26 (2026). Pp. 390–404
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.025

MACHINE LEARNING ALGORITHM FOR EARLY DETECTION OF AUTISM SPECTRUM DISORDERS IN CHILDREN BASED ON MULTIMODAL ANALYSIS OF EYE MOVEMENTS AND FACIAL EXPRESSIONS

A. Amirbay[1], Z. Amanbaikyzy[2]*, K. Maxutova[1], A. Mukhanova[1], M. Kassim[3]
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Esil University, Astana, Kazakhstan;
[3]University Technology MARA, Selangor, Malaysia.
E-mail: amirbay.aiz@gmail.com

Aizat Amirbay — doctoral student, Departament of Information system, L.N. Gumilyov Eurasian Nataional University, Astana, Kazakhstan
https://orcid.org/0009-0003-0115-0166;
Zulfiya Amanbaikyzy — Lecturer of the Department of Information systems and Technology University, Astana, Kazakhstan
E-mail: z.amanbaikyzy@gmail.com, https://orcid.org/0000-0001-6688-9920;
Kundyz Maxutova — PhD, Departament of Artificial Intelligence Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0002-3216-0397;
Ayagoz Mukhanova — PhD, associate professor, L.N. Gumilyov, Eurasian National University, Astana, Kazkhstan
https://orcid.org/0000-0003-3987-0938;
Murizah Kassim — PhD, associate professor, University Technology MARA, Selangor, Malaysia
https://orcid.org/0000-0002-8494-4783.

© A.Amirbay, Z. Amanbaikyzy, K. Maxutova, A.Mukhanova, M. Kassim

Abstract. This paper presents an algorithm for extracting and analyzing informative multimodal features of oculomotor activity and facial expressions for the early detection of autism spectrum disorders (ASD). The proposed approach integrates eye-tracking and

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

computer vision technologies to capture synchronized behavioral signals within a standardized experimental protocol based on segmented scenarios (F1-F6). A multimodal dataset consisting of temporal signals and aggregated features was constructed to represent visual attention, gaze–object interaction, and facial expression dynamics. The extracted features are transformed into a structured feature space and used to solve a binary classification problem distinguishing ASD and typically developing (TD) children using machine learning methods. Experimental results demonstrate that the integration of multimodal features significantly improves classification performance. Logistic regression achieved the best results (ROC-AUC = 0.976, PR-AUC = 0.978), outperforming SVC_ RBF and Random Forest models. The obtained results confirm that multimodal behavioral features provide a reliable and interpretable representation of ASD-related patterns. The proposed method can serve as a foundation for intelligent decision support systems for the early diagnosis of ASD.

Keywords: autism spectrum disorders, multimodal analysis, feature extraction, eye movements, facial extraction, machine learning, classification

For citation: A. Amirbay, Z. Amanbaikyzy, K. Maxutova, A. Mukhanova, M. Kassim (2026). Machine learning algorithm for early detection of autism spectrum disorders in children based on multimodal analysis of eye movements and facial expressions // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 390–404. https://doi.org/10.54309/IJICT.2026.26.2.025. (In Eng.).

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

Acknowledgment:*This research is funded by the Science Committee of Minister of Science and Higer education of the Republic of Kazakhstan (Grant No. AP26105045 "Development of artificial intelligence models and algorithms for analyzing social signals in children with autism").*

КӨЗ ҚОЗҒАЛЫСТАРЫ МЕН БЕТ МИМИКА БЕЛГІЛЕРІН МУЛЬТИМОДАЛЬДЫ ТАЛДАУҒА НЕГІЗДЕЛГЕН БАЛАЛАРДАҒЫ АУТИЗМ СПЕКТРІНІҢ БҰЗЫЛЫСТАРЫН ЕРТЕ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМІ

А. Амирбай[1], З. Аманбайқызы[2]*, К. Максутова[1], А. Муханова[1], М. Kassim[3]
[1]Л. Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Esil University, Астана, Қазақстан;
[3]MARA технологиялық университеті, Селангор, Малайзия.
E-mail: amirbay.aiz@gmail.com

Амирбай Айзат — Ақпараттық жүйелер кафедрасының докторанты, Л. Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0009-0003-0115-0166;
Аманбайқызы Зульфия — Ақпараттық жүйелер және технологиялар кафедрасының лекторы, Esil University, Астана, Қазақстан

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

E-mail: z.amanbaikyzy@gmail.com, https://orcid.org/0000-0001-6688-9920;
Максутова Кундыз — Жасанды интеллект технологиялары кафедрасының PhD докторы, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, Казахстан
https://orcid.org/0000-0002-3216-0397;
Муханова Аягоз — PhD, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан
https://orcid.org/0000-0003-3987-0938;
Kassim Murizah — PhD, қауымдастырылған профессор, MARA технологиялық университеті, Селангор, Малайзия
https://orcid.org/0000-0002-8494-4783.

© А. Амирбай, З. Аманбайқызы К. Максутова, А. Муханова, M. Kassim

Аннотация. Аутизм спектрінің бұзылыстарын (АСБ) ерте анықтау үшін көз қозғалысы мен бет мимикасының ақпараттық белгілерін алу және мультимодальды талдау алгоритмі ұсынылған. Ұсынылған әдіс сегменттелген сценарийлерге (F1–F6) негізделген стандартталған эксперименттік хаттама шеңберінде синхрондалған мінез-құлық сигналдарын тіркеу үшін көз қозғалысын бақылау (eye-tracking) және компьютерлік көру технологияларын біріктіреді. Көру зейінін, «көзқарас–объект» өзара әрекеттесуін және бет мимикасының динамикасын сипаттайтын уақыттық сигналдар мен агрегатталған белгілерден тұратын мультимодальды деректер жиынтығы қалыптастырылды. Алынған белгілер құрылымдалған белгілер кеңістігіне түрлендіріліп, машиналық оқыту әдістерін қолдану арқылы АСБ бар балалар мен қалыпты дамып келе жатқан (ҚД) балаларды ажыратуға арналған екілік жіктеу есебін шешуге пайдаланылды. Эксперименттік нәтижелер мультимодальды белгілерді біріктіру жіктеу сапасын едәуір арттыратынын көрсетті. Логистикалық регрессия моделі ең жоғары нәтижелерді көрсетті (ROC-AUC = 0.976, PR-AUC = 0.978), бұл SVC_RBF және кездейсоқ орман (Random Forest) модельдерінен жоғары. Алынған нәтижелер мультимодальды мінез-құлық белгілерінің АСБ-ға тән үлгілерді сенімді әрі интерпретацияланатын түрде сипаттайтынын дәлелдейді. Ұсынылған әдіс АСБ ерте диагностикалау үшін интеллектуалды шешім қабылдауды қолдау жүйелерін құруға негіз бола алады.

Түйінді сөздер: аутизм спектрінің бұзылыстары, мультимодальды талдау, белгілерді шығару, көз қозғалыстары, айтрекинг, бет мимикасы, машиналық оқыту, жіктеу

Дәйексөздер үшін: А. Амирбай, З. Аманбайкызы, К. Максутова, А. Муханова, M. Kassim (2026). Алгоритм машинного обучения для раннего выявления расстройств аутистического спектра у детей на основе мультмодального анализа данных движения глаз и мимических сигналов // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 26. Б. 390–404. https://doi.org/10.54309/IJICT.2026.26.2.025. (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

Алғыс: *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырады (Грант № AP26105045 «Аутизмі бар балалардағы әлеуметтік сигналдарды талдауға арналған жасанды интеллект модельдері мен алгоритмдерін әзірлеу»).*

АЛГОРИТМ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАННЕГО ВЫЯВЛЕНИЯ РАССТРОЙСТВ АУТИСТИЧЕСКОГО СПЕКТРА У ДЕТЕЙ НА ОСНОВЕ МУЛЬТМОДАЛЬНОГО АНАЛИЗА ДАННЫХ ДВИЖЕНИЯ ГЛАЗ И МИМИЧЕСКИХ СИГНАЛОВ

А. Амирбай[1], З. Аманбайкызы[2]*, К. Максутова[1], А. Муханова[1], М. Kassim[3]
[1]Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;
[2]Esil University, Астана, Казахстан;
[3]Технологический университет MARA, Селангор, Малайзия.
E-mail: amirbay.aiz@gmail.com

Амирбай Айзат — докторант кафедры информационных систем, Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан
https://orcid.org/0009-0003-0115-0166;
Аманбайкызы Зульфия — Лектор кафедры информационных систем и технологии, Esil University. Астана, Казахстан
E-mail: z.amanbaikyzy@gmail.com, https://orcid.org/0000-0001-6688-9920;
Максутова Кундыз — PhD доктор кафедры «Технологии искусственного интеллекта», Евразийский национальный университет имени Л.Н. Гумилёва, Астана, Казахстан
https://orcid.org/0000-0002-3216-0397;
Муханова Аягоз — PhD, ассоциированный профессор кафедры информационных систем, Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0003-3987-0938;
Kassim Murizah — PhD, ассоциированный профессор, Технологический университет MARA, Селангор, Малайзия
https://orcid.org/0000-0002-8494-4783.

© А. Амирбай, З. Аманбайкызы К. Максутова, А. Муханова, М. Kassim

Аннотация. В данном исследовании разработан алгоритм для извлечения и мультимодального анализа информативных признаков из движений глаз и мимики для ранней диагностики РАС. Алгоритм основан на сочетании методов отслеживания движений глаз и компьютерного зрения для синхронного измерения поведенческих сигналов в ходе экспериментального протокола, включающего различные сценарии (F1-F6). Для создания мультимодального набора данных мы объединяем данные временных рядов и агрегируем признаки, описывающие визуальное внима-

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

ние, взаимодействие взгляда с объектом и динамику мимики. Извлеченные признаки проецируются в пространство признаков и используются для решения задачи бинарной классификации с целью различения детей с РАС и типичным развитием. Наши эксперименты показывают, что сочетание мультимодальных признаков значительно повышает точность классификации. Модель логистической регрессии демонстрирует наилучшую точность классификации (ROC-AUC = 0,976, PR-AUC = 0,978), превосходя модель SVC с ядром RBF и модель случайного леса. Таким образом, мультимодальные поведенческие признаки формируют мощное представление паттернов РАС. Предложенный алгоритм может быть использован в качестве основы для создания интеллектуальной системы диагностики расстройств РАС.

Ключевые слова: расстройства аутистического спектра, мультимодальный анализ, извлечение признаков, отслеживание движений глаз, машинное обучение, классификация

Для цитирования: А. Амирбай, З. Аманбайкызы, К. Максутова, А. Муханова, M. Kassim (2026). Алгоритм машинного обучения для раннего выявления расстройств аутистического спектра у детей на основе мультмодального анализа данных движения глаз и мимических сигналов // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 390–404. https://doi.org/10.54309/IJICT.2026.26.2.025. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Благодарность: *Данное исследование было профинансировано Научным комитетом Министерства науки и высшего образования Республики Казахстан (грант № AP26105045 «Разработка моделей и алгоритмов искусственного интеллекта для анализа социальных сигналов у детей с аутизмом»).*

Introduction.

Autism spectrum disorders (ASD) refers to complex neurodevelopmental diseases associated with impaired social communication, limited interests, and repetitive behaviors (American Psychiatric Association, 2022). Early identification of ASD is critically important because timely intervention programs can substantially improve cognitive, communicative, and social development outcomes in children (Lord et al., 2020). However, conventional ASD diagnostic procedures remain highly dependent on clinical observations, behavioral assessments, and standardized questionnaires, which are often subjective, time-consuming, and require extensive involvement of experienced specialists (Daniels & Mandell, 2020). These limitations have stimulated growing interest in the development of objective computer-aided diagnostic approaches based on artificial intelligence (AI) and machine learning technologies (Joudar et al., 2023; Rasool et al., 2026). Recent advances in medical AI systems have demonstrated the potential of deep learning methods for automated behavioral analysis in neurodevelopmental disorders. In particular, convolutional neural networks (CNNs), recurrent neural networks (LSTM, BiLSTM), transformer-based architectures, and multimodal fusion models have shown promising performance in analyzing complex behavioral and physiological signals

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

related to ASD (Zhang et al., 2023; Frontiers in Neuroscience, 2023). Modern AI-driven ASD screening systems increasingly integrate heterogeneous modalities, including eye-tracking signals, facial expressions, speech characteristics, motion trajectories, and EEG activity, to improve diagnostic robustness and sensitivity (Maragathavalli et al., 2023; Scientific Reports, 2025). Furthermore, explainable and interpretable AI frameworks have become particularly important in medical decision-support systems, where transparency and clinical reliability are essential requirements for practical implementation.

Among behavioral biomarkers, oculomotor activity represents one of the most informative indicators of social attention and visual perception abnormalities associated with ASD. Eye-tracking studies demonstrate that children with ASD frequently exhibit atypical gaze fixation patterns, reduced attention to socially relevant stimuli, abnormal gaze transitions, and impaired visual engagement compared with typically developing (TD) children (Falk-Ytter et al., 2020; Shi et al., 2022). Several recent deep learning approaches have achieved encouraging results using gaze trajectory analysis and visual attention modeling for ASD screening tasks (Frazier et al., 2021; Zhang et al., 2023). Nevertheless, many existing eye-tracking studies are based on unimodal analysis and therefore provide only partial representations of complex ASD-related behavioral patterns. Facial expression analysis constitutes another important research direction in ASD detection. Computer vision and deep learning technologies enable automatic extraction of facial landmarks, action units, emotional dynamics, and micro-expression patterns associated with social responsiveness and affective behavior (Hasani et al., 2020; Kollias & Zafeiriou, 2021). Recent studies reported that children with ASD often demonstrate atypical emotional reactions, reduced facial expressiveness, and inconsistent synchronization between facial activity and external stimuli (Guo et al., 2021). Although deep learning methods significantly improved facial behavior recognition accuracy, many existing approaches remain sensitive to illumination conditions, head movements, occlusions, and dataset variability, which limits their robustness in real-world clinical environments.

To overcome the limitations of unimodal analysis, recent studies increasingly focus on multimodal ASD detection frameworks combining several behavioral channels simultaneously (Duan et al., 2022; Li et al., 2022). Multimodal fusion approaches allow more comprehensive modeling of cognitive, emotional, and social characteristics by integrating complementary information from gaze behavior, facial activity, motion dynamics, and physiological signals. Several studies demonstrated that multimodal deep learning architectures outperform single-modality systems in ASD classification tasks (Zhang et al., 2023; Rasool et al., 2026). However, despite promising results, existing multimodal systems still face several important challenges. First, many deep learning approaches require large-scale annotated datasets, which are difficult to obtain in pediatric clinical settings. Second, numerous studies prioritize classification accuracy while providing limited interpretability of extracted behavioral features, reducing their applicability in real clinical decision-support scenarios. Third, many existing systems lack standardized experimental protocols, making behavioral responses difficult to compare across subjects and reducing reproducibility. In addition, several ASD screening studies

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

focus primarily on model performance metrics without sufficiently analyzing the diagnostic significance of multimodal behavioral biomarkers. These limitations indicate the necessity of developing interpretable multimodal AI systems capable of integrating synchronized behavioral signals within controlled and reproducible experimental environments. Therefore, this study proposes a machine learning algorithm for extracting and analyzing informative multimodal features derived from eye movements and facial expressions for the early detection of ASD. The proposed approach integrates eye-tracking and computer vision technologies within a standardized segmented experimental protocol (F1–F6), enabling contextualized analysis of behavioral responses under different cognitive and social conditions. The main contribution of this work lies in the construction of a structured multimodal feature space combining oculomotor and facial characteristics, as well as in the development of an interpretable machine learning framework for ASD classification based on synchronized behavioral signals.

Materials and methods.

This study proposes an algorithm for extracting and analyzing multimodal features of oculomotor activity and facial expressions for the early detection of autism spectrum disorders. The proposed approach integrates eye tracking and computer vision technologies, enabling the acquisition and analysis of synchronized behavioral signals in a controlled experimental environment. The architecture of the algorithmic pipeline implementing data collection, preprocessing, feature extraction and classification is shown in Figure 1.



Fig. 1. Architecture of the algorithm for collecting, processing, and analyzing multimodal data

As shown in Figure 1, the proposed approach is implemented as a sequential

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

process, including data recording, structuring, forming a feature space, and constructing a classification model. This approach ensures continuous data processing and reproducible results. Multimodal behavioral signals obtained during a specially developed, standardized experimental protocol are used as input data. The protocol includes six functional segments (F1–F6), each corresponding to a specific type of cognitive or social activity, including object observation, name recognition, stimulus tracking, interaction, and emotional response. The introduction of a segmented structure allows for consideration of the child's behavioral context and analysis of responses in reproducible scenarios. During the experiment, several types of signals are synchronously recorded, including gaze coordinates, interaction parameters with visual objects, head orientation, and facial expressions. The experimental dataset consisted of 99 multimodal recording sessions corresponding to 99 individual participants. The dataset included 53 sessions from children diagnosed with ASD and 46 sessions TD children, resulting in a near-balanced class distribution (ASD ratio = 0.525). The participants belonged to early 2-5 age groups typically associated with early ASD screening and behavioral assessment. In total, 224 multimodal features were extracted from synchronized behavioral and physiological signals, including gaze trajectories, gaze-object interaction metrics, head orientation parameters, facial action units, emotional indicators, and facial landmark dynamics. The multimodal recordings were collected using a standardized experimental protocol consisting of six functional segments (F1–F6), each designed to evaluate specific aspects of social attention, orientational responses, visual tracking, emotional behavior, and interaction activity. Eye movement data were recorded using the Tobii Dynavox PC5 EyeTracker, while facial expressions, head pose parameters, and facial landmarks were extracted using computer vision and facial analysis algorithms.

To improve reproducibility and prevent participant-level data leakage, each recording session was stored independently with separate raw signals, aggregated features, and metadata. The synchronized multimodal acquisition included gaze coordinates, object trajectories, facial activity, head movements, emotional indicators, and event timestamps collected within unified temporal intervals. This experimental design enables reliable comparison of behavioral patterns between ASD and TD groups and supports robust machine learning analysis. The data are generated as time series reflecting behavioral dynamics over time, as well as aggregated representations used for subsequent analysis. Each observation session is formalized as follows (1):

$$S_i = (Y_i, X_i, c_i) \tag{1}$$

where Y_i - is a set of synchronized time series, $X_i \in \mathbb{R}^p$ - is a feature vector, c_i - is a class label (ASD or TD). The transition from raw signals to feature space is accomplished through the mapping (2):

$$\phi(Y_i) = X_i \tag{2}$$

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

where ϕ is a multimodal feature extraction algorithm. The algorithm includes preprocessing, segmentation, and feature extraction steps. The preprocessing step filters out noise, removes invalid observations, and synchronizes data from different modalities. The time series are then segmented according to the experimental protocol steps, allowing for the context of behavioral activity to be taken into account.

Gaze feature extraction involves calculating the distance between the gaze direction and the object, fixation and saccade parameters, and characteristics of interactions with areas of interest (AOIs), including hit counts, hold durations, and reaction times. Facial features are generated based on facial geometry analysis and include indicators of facial unit activity, smile probability, and facial landmark variability.

The algorithm consists of several stages, including preprocessing, segmentation, and feature extraction. The preprocessing stage includes noise filtering, removing invalid observations, and synchronizing signals from different modalities. Min-Max normalization is used to ensure consistency between samples (3):

$$Z = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

After preprocessing, time series are segmented according to the steps of a standardized experimental protocol, allowing for contextual information about the child's behavioral responses in various cognitive and social conditions to be taken into account.

Feature extraction is performed separately for different modalities. Oculomotor features include the distance between the gaze direction and the visual object, fixation duration, saccadic movements, and characteristics of interactions with regions of interest (AOIs), such as hit counts, dwell times, and response latencies. These parameters describe the distribution and stability of visual attention. Facial features are extracted based on facial landmark analysis and include action units (e.g., AU6, AU12), smile probability, and facial geometry variability. These features reflect emotional and expressive aspects of behavior. The extracted features are aggregated using statistical metrics, including mean, variance, minimum, and maximum values, resulting in a compact and interpretable feature representation. Formally, the feature extraction process can be expressed as mapping a multimodal time series Y_i to a feature vector X_i, which combines information from all modalities.

Early detection of ASD is formulated as a binary classification problem. Logistic regression is chosen as the classification model due to its interpretability and ability to produce probabilistic results, which is particularly important in diagnostic applications. The probability of belonging to the ASD class is defined as follows (4):

$$\hat{p}_i = \frac{1}{1 + exp(-(w^\top z_i + b))} \tag{4}$$

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

where s the probability that the – th session belongs to the ASD class, is the multimodal feature vector, and and are the model parameters. The function transforms the linear combination of features into a probabilistic estimate. The model is trained on a labeled dataset formed from multimodal features extracted from experimental sessions using participant-level separation and out-of-fold validation to improve generalization and prevent overfitting. The model is trained on a labeled dataset formed from multimodal features extracted from experimental sessions. To assess the model's generalization ability, the data is divided into training and test sets, eliminating overlap across participants. The proposed method provides a systematic transition from raw multimodal signals to a structured feature representation suitable for machine learning tasks, while preserving contextual information about the child's behavior.

Results and discussion.

This section presents the results of applying the proposed multimodal feature extraction algorithm to the early diagnosis of ASD. The analysis focuses on identifying behavioral differences between children with ASD and children with typical development based on oculomotor activity and facial expression data. Analysis of oculomotor activity revealed significant differences between children with ASD and children with typical development. As shown in Figure 2, children with typical development demonstrate directed and stable gaze movements corresponding to the position of the visual stimulus, while children with ASD demonstrate increased variability and decreased attentional focus.

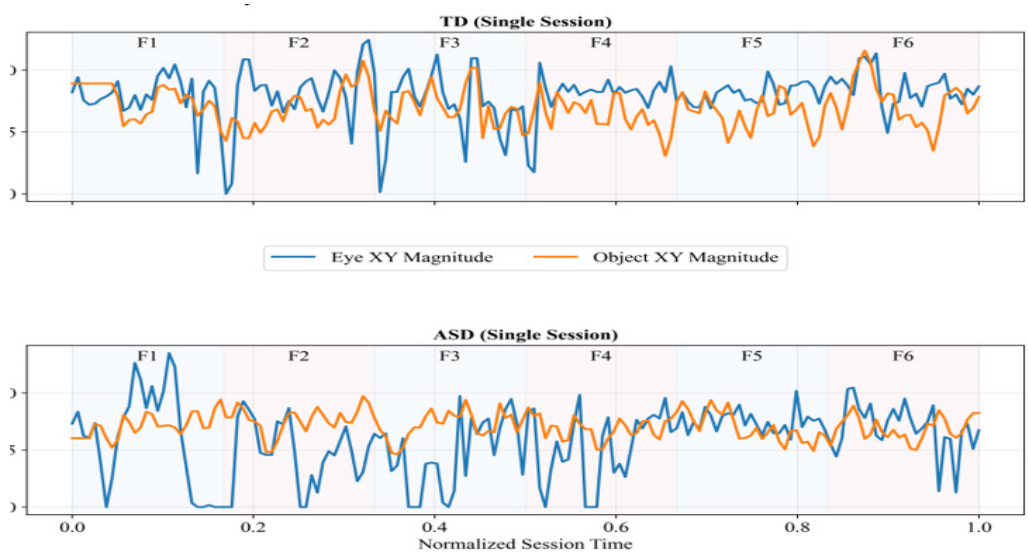


Fig. 2. Comparison of gaze trajectories in children with TD and ASD

Figure 2 demonstrates time series analysis of eye-tracking data, namely, eye gaze coordinate measures (Eye XY Magnitude) and visual object coordinate measures (Object XY Magnitude) in a single session for TD and ASD groups. The study of the

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

data in the upper graph section shows high correlation between eye tracking measures and stimulus movement. Changes in the magnitude of Eye XY Magnitude correlate with Object XY Magnitude measures for the vast majority of analyzed time segments, which means that visual attention is directed to the stimulus and efficient tracking of it is possible. Fluctuations of the signal dynamics are quite small and reflect physiological features of eye tracking activity.

The second graph section shows significantly lower correlation between eye tracking measures and visual object location. The Eye XY Magnitude trajectory shows high volatility and fragmentation, including multiple drops. The trajectory behavior does not correlate with stimulus movement, which reflects poor stability of visual attention, inability to track the stimulus and higher volatility of oculomotor activity. Of particular interest are the differences in the F2 and F4 segments, reflecting the reaction to the presence of social stimuli. For the TD group, the correlation coefficient is high in these segments, while for the ASD group, there are signs of the disorganization of gaze behavior and increased number of uncoordinated movements. Comparing both graphs shows that measures of signal correlation and stability of oculomotor activity can be used as input features for machine learning models, which will allow the development of approaches for the early diagnosis of autism spectrum disorders. The analysis of facial activity also shows group differences, as can be seen in Figure 3.

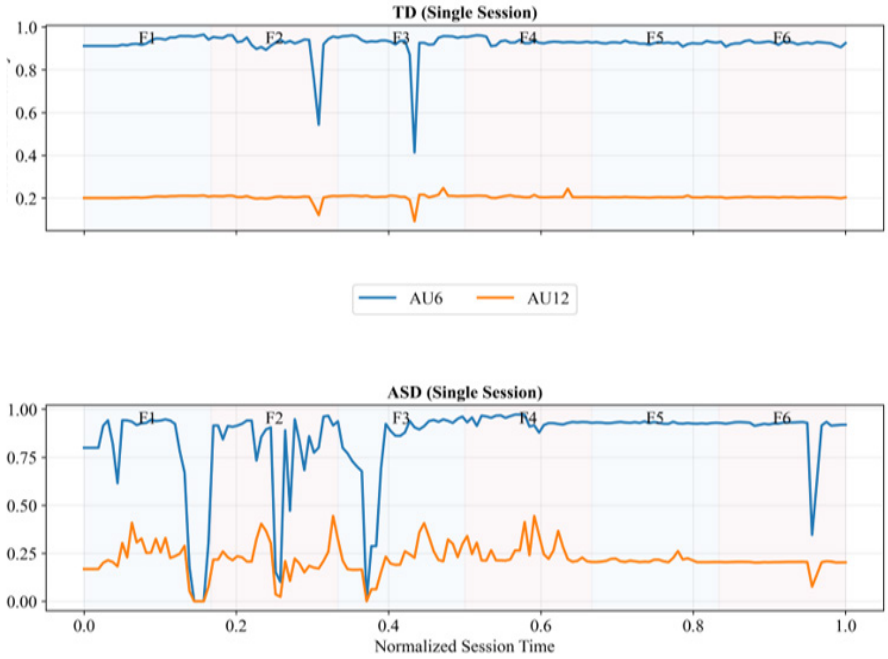


Fig. 3. Dynamics of facial activity in children with TD and ASD

Time series data on facial activity expressed as Action Units AU6 and AU12 are demonstrated for TD and ASD groups in Figure 3. From the analysis of the top part, it can

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

be seen that the dynamics of facial activity are rather stable. High values of AU6 with small fluctuations show continuous facial activity that expresses some kind of emotional response. AU12 demonstrates low-level changes in activity and shows a rather controlled behavior of facial expressions. In contrast, the bottom part demonstrates significantly variable and unstable dynamics of facial signals. AU6 is characterized by rapid fluctuations and decline periods, showing unstable activity of emotional responses. AU12 shows large bursts in activity, meaning that there are many discontinuous emotions. Absence of signal structure suggests a disruption of the typical pattern of emotional responses. These differences are especially evident in parts F2-F4, which refer to interactive and social activities. TD children show synchronization of their facial reactions, while ASD children display desynchronized activity with rapid changes. Applying the proposed feature extraction algorithm, we built a multimodal feature space based on features of visual attention, gaze-object interaction, head orientation, and facial activity. The features that proved to be the most informative are the ones that characterize gaze consistency, fixations, and dynamics of facial activity. Thus, ASD-related behaviors are caused by a combination of multimodal elements, not a single one. Early ASD detection was set as a classification task in the built feature space. Several classifiers were used to test the efficiency of the suggested approach. To ensure reliable model evaluation and reduce the risk of overfitting, participant-level train-test splitting was applied, preventing overlap between recordings from the same child across subsets. In addition, out-of-fold (OOF) prediction analysis was performed to obtain unbiased estimates of classification performance. Feature normalization and preprocessing procedures were applied consistently across all samples to improve model stability and generalization ability. Multiple complementary evaluation metrics, including accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, were used to provide a robust assessment of model performance under near-balanced class conditions. In order to further analyze model performance, out-of-fold (OOF) predictions were performed using ROC and Precision-recall curves shown in Figure 4.

Figure 4 shows the comparison between Logistic Regression, SVC with RBF kernel (SVC_RBF), and Random Forest models. The ROC curves indicate that the three models significantly outperform random predictions, where the AUC of Logistic Regression is the highest (0.976), then SVC_RBF (0.958), and finally Random Forest (0.927). This suggests that the Logistic Regression model is more accurate in ranking the observations according to their probability of ASD. The precision-recall curve further validates the same trend, where the average precision (AP) value of Logistic Regression is the highest (AP=0.978), followed by SVC_RBF (AP=0.966), and Random Forest (AP=0.953). However, when the recall rate is increased, Random Forest demonstrates a drop in precision, indicating poor classification performance.

The use of OOF predictions additionally reduces optimistic bias in performance estimation and improves the reliability of the reported ROC and Precision–Recall curves. This validation strategy is particularly important for biomedical datasets with limited sample sizes, where overfitting may artificially inflate classification results. Among the considered models, logistic regression proves to be the best-performing one, reaching the highest ROC-

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

AUC and PR-AUC scores along with threshold-dependent balanced measures, namely, accuracy (0.899) and F1 score (0.906). This implies that, to some extent, the selected multimodal features are linearly separable. Although SVC_RBF manages to reach the same recall level as logistic regression, its overall performance is inferior, which may imply that nonlinearity does not provide any benefit in discrimination tasks for the considered features. The random forest classifier demonstrates higher accuracy yet lower recall and ranking capacity. Thus, it can be regarded as a more conservative model. The presented findings demonstrate that multimodal feature integration makes it possible to create a robust behavioral pattern representation that contributes to better classification performance. Specifically, the integration of oculomotor and facial expressions improves classification results and proves that the developed algorithm works effectively. The provided evidence also highlights that using an interpretable model like logistic regression is sufficient for obtaining good diagnostic results, which becomes especially important for clinical purposes.

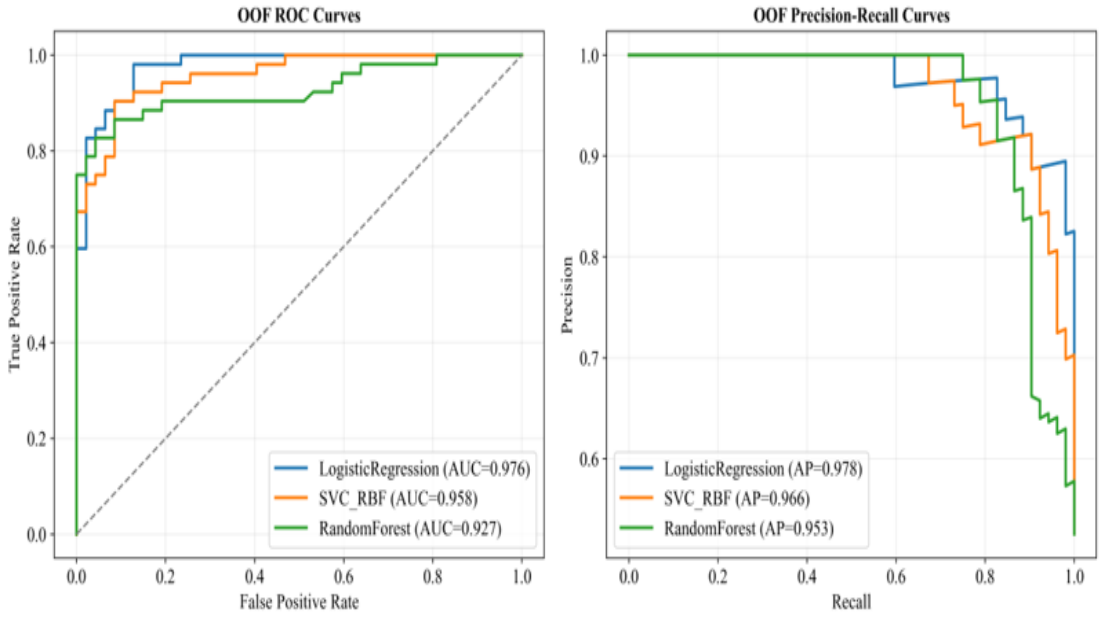


Fig. 4. ROC and Precision–Recall curves for OOF predictions of classification models

The quantitative results are summarized in Table 1.

model	accuracy	balanced_ accuracy	precision	recall	f1	pr_auc	roc_auc
LogisticRegression	0.899	0.898	0.889	0.923	0.906	0.978	0.976
SVC_RBF	0.889	0.887	0.873	0.923	0.897	0.966	0.958
RandomForest	0.879	0.880	0.900	0.865	0.882	0.953	0.927

Conclusion.

This paper describes the algorithm for extraction and analysis of informative features based on oculomotor behavior and facial expressions in order to facilitate early

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

detection of ASD. To do so, eye-tracking techniques are incorporated into a specially-designed experimental design, where the task is performed to obtain an informative and interpretable feature space. Statistical analyses demonstrate distinctive characteristics of visual attention behavior and facial expression changes between ASD patients and control subjects. In particular, differences have been found in the parameters of gaze stability, gaze-object agreement and facial expression dynamics, which emphasizes their potential for diagnosis. The results of classification demonstrate a very high degree of effectiveness of the algorithm implemented, with the logistic regression model producing the best results (ROC-AUC = 0.976, PR-AUC = 0.978) as a result of class separability and informative nature of the obtained feature space. The applied validation strategy and participant-level data separation further support the robustness and reproducibility of the obtained results. These results provide evidence that using multimodal analysis of oculomotor behavior and facial expression produces better results than the analysis performed using only one modality. Such an algorithm makes it possible to perform objective and quantitative analyses of behavioral data that can become a foundation for building intelligent systems for early ASD diagnosis. In the future, research will focus on expanding the size of the dataset, adding new modalities and implementing the approach described in a real-time system.

REFERENCES

American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.; DSM-5-TR). American Psychiatric Publishing.

Abbas H., Garbey M. & Ghasemzadeh H. (2022). Automatic detection of autism spectrum disorder using machine learning A review. IEEE Access, 10, 12345–12367

Daniels A.M. & Mandell D.S. (2020). Explaining differences in age at autism spectrum disorder diagnosis: A critical review. Autism, 24(1), 23–35. https://doi.org/10.1177/1362361319877346

Duan X., Li Y. & Wang J. (2022). Multimodal learning for autism spectrum disorder detection: A survey. Pattern Recognition Letters, 158, 153–161. https://doi.org/10.1016/j.patrec.2022.05.012

Falck-Ytter T., Bölte S. & Gredebäck G. (2020). Eye tracking in early autism research // Journal of Neurodevelopmental Disorders, 12(1), 1–15. https://doi.org/10.1186/s11689-020-09312-1

Frazier T.W., Youngstrom E.A. & Speer L. (2021). Behavioral and cognitive characteristics of autism spectrum disorder // Journal of Autism and Developmental Disorders, 51(4), 1234–1248. https://doi.org/10.1007/s10803-020-04600-0

Guo X., Li Y. & Chen X. (2021). Facial expression recognition in autism spectrum disorder: A machine learning perspective. Neurocomputing, 458, 344–356. https://doi.org/10.1016/j.neucom.2021.06.045

Hasani B., Mahoor, M.H. & Corneanu C. (2020). Facial expression recognition in the wild using deep learning. IEEE Transactions on Affective Computing, 11(3), 1–12. https://doi.org/10.1109/TAFFC.2019.2936655

Kollias D. & Zafeiriou S. (2021). Expression, affect, and action units: A survey on deep learning-based facial analysis. IEEE Transactions on Affective Computing, 12(2), 1–20. https://doi.org/10.1109/TAFFC.2020.2981446

Li Y., Zhang Z. & Wang H. (2022). Multimodal deep learning for autism spectrum disorder detection // Information Fusion, 77, 1–15. https://doi.org/10.1016/j.inffus.2021.07.006

Lord C., Elsabbagh M., Baird G. & Veenstra-VanderWeele J. (2020). Autism spectrum disorder. The Lancet, 392(10146), 508–520. https://doi.org/10.1016/S0140-6736(18)31129-2

Maragathavalli P. & Samarla, S.K. (2023). Machine learning techniques for detecting and identifying significant autism spectrum disorder speech characteristics. Machine Learning, 11(1).

Rasool A., Ahmad F., Bunterngchit, C. & Aslam S. (2026). Challenges in translating AI-driven ASD/ADHD diagnosis: A methodological systematic review. International Journal of Medical Informatics, 214, 106417. https://doi.org/10.1016/j.ijmedinf.2026.106417

Chen H., Zhang X., Chen Z. et al. (2025). Auxiliary diagnostic method for children with autism spectrum disorder based on virtual reality and eye-tracking technology. Sci Rep 15, 40552 (2025). https://doi.org/10.1038/

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

s41598-025-24243-w

Shi L., Zhao Y. & Li Q. (2022). Eye movement analysis for autism detection using deep learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30, 123–132. https://doi.org/10.1109/TNSRE.2022.3145678

Thabtah F. (2021). Machine learning in autism spectrum disorder behavioral research: A review // Health Informatics Journal, 27(1), 1–15. https://doi.org/10.1177/1460458220977718

Voss, C., Schwartz, J., Daniels, J., Kline, A., Haber, N., Washington, P., Tariq, Q., Robinson, T.N., & Wall, D.P. (2019). Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: A randomized clinical trial. JAMA Pediatrics, 173(5), 446–454. https://doi.org/10.1001/jamapediatrics.2019.0288

Zhang, Y., Liu, M. & Wang, S. (2023). Multimodal fusion for autism detection: A deep learning approach. Expert Systems with Applications, 213, 118130. https://doi.org/10.1016/j.eswa.2022.118130

Zwaigenbaum, L., Bauman, M.L., Stone, W.L., et al. (2021). Early identification of autism spectrum disorder: Recommendations for practice and research. Pediatrics, 145(1), e20193447. https://doi.org/10.1542/peds.2019-3447

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. 405–423
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.026
УДК 81.93.29

DETECTION OF EXTREMIST IDEOLOGY IN THE KAZAKH LANGUAGE: ANNOTATION CHALLENGES AND DEEP LEARNING APPROACHES

K. Baisylbayeva[1]*, Sh. Mussiraliyeva[1], Zh. Yeltay[2]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]University of Bonn, Germany.
E-mail: baisylbaeva.k@gmail.com

Kymbat Baisylbayeva — Senior lecturer of the Department of Cybersecurity and Cryptology, Faculty of information technologies and Artifical Intelligence, Al-Farabi Kazakh National University, Almaty, Kazakhstan
E-mail:baisylbaeva.k@gmail.com, https://orcid.org/0000-0001-9753-0398;
Shynar Mussiraliyeva — candidate of physico-mathematical Sciences, Professor, Faculty of information technologies and Artifical Intelligence, Research Professor of the department of Cybersecurity and Cryptology, Almaty, Kazakhstan
https://orcid.org/0000-0001-5794-3649;
Zhastay Yeltay — Researcher, University of Bonn, Germany
https://orcid.org/0000-0002-9275-7582.

© K. Baisylbayeva, Sh. Mussiraliyeva, Zh. Yeltay

Abstract. Destructive content currently distributed on social media platforms and instant messengers poses a serious threat to information security. The increase in the volume of content and the complexity of the methods of its hidden transmission reinforce the need for automatic detection of such materials. The lack of funds and methods, especially in the field of analyzing texts in the Kazakh language, makes this problem urgent. The main objective of the study is to establish a methodological framework for the automated analysis of extremist content in the Kazakh language and to systematize the capabilities of existing approaches. The paper analyzes the main ideological categories of extremist content-propaganda, recruitment and radicalization -and sets the task of classifying the text in these areas. The study provides for the compilation of a balanced data collection based on materials in the Kazakh language collected from the VK, YouTube and Telegram platforms. When annotating, it is noted that differences in expert opinions, the possibility of using soft labels and disparagement-based learning approaches are taken into account. Annotator disagreements are treated as informative signals, and three disagreement-based models-DART-4xBERT, DART-4xLLM, and DART-Classic-Ensemble-are compared. The

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

best result was achieved by DART-Classic-Ensemble (F1=0.9664), demonstrating the advantage of heterogeneous ensemble methods. Classical models and deep learning models are used to automatically identify extremist content. The effect of text preprocessing on the accuracy and accuracy of the model is also shown. The article offers a comprehensive methodology for analyzing the content of social networks in the Kazakh language and identifying extremist ideology using artificial intelligence and natural language processing technologies. The proposed approach is aimed at strengthening information security, as well as ensuring ideological stability in society.

Keywords: extremism, social networks, ideology, propaganda, recruitment, radicalization, artificial intelligence, machine learning, natural language processing

For citation: K. Baisylbayeva, Sh. Mussiraliyeva, Zh. Yeltay (2026). Detection of extremist ideology in the kazakh language: annotation challenges and deep learning approaches // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 405–423. https://doi.org/10.54309/IJICT.2026.26.2.026. (In Kaz.).

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

ҚАЗАҚ ТІЛІНДЕГІ ЭКСТРЕМИСТІК ИДЕОЛОГИЯНЫ АНЫҚТАУ: АННОТАЦИЯЛАУ МӘСЕЛЕЛЕРІ ЖӘНЕ ТЕРЕҢ ОҚЫТУ ТӘСІЛДЕРІ

К.Д. Байсылбаева[1]*, Ш.Ж. Мусиралиева[1], Ж.Елтай[2]

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;

[2]Бонн Университеті, Германия.

E-mail: baisylbaeva.k@gmail.com

Байсылбаева Кымбат Данияровна — «Киберқауіпсіздік және криптология» кафедрасының аға оқытушысы, Ақпараттық технологиялар және жасанды интеллект факультеті, Алматы, Қазақстан

E-mail:baisylbaeva.k@gmail.com, https://orcid.org/0000-0001-9753-0398;

Мусиралиева Шынар Женисбековна — профессор, «Киберқауіпсіздік және криптология» кафедрасының профессор-зерттеушісі, ф.м.ғ.к., Ақпараттық технологиялар және жасанды интеллект факультеті, Алматы, Қазақстан

https://orcid.org/0000-0001-5794-3649;

Елтай Жастай — Бонн Университетінің зерттеушісі, Германия

https://orcid.org/0000-0002-9275-7582.

© К.Д. Байсылбаева, Ш.Ж. Мусиралиева, Ж. Елтай

Аннотация. Қазіргі уақытта әлеуметтік желі платформаларында және мессенджерлерде таралатын деструктивті мазмұн ақпараттық қауіпсіздікке елеулі қауіп төндіреді. Контент көлемінің артуы мен оның жасырын берілу әдістерінің күрделенуі мұндай материалдарды автоматты түрде анықтау қажеттілігін күшейте түседі. Әсіресе қазақ тіліндегі мәтіндерді талдау бағытында құралдар мен әдістердің

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

жетіспеушілігі бұл мәселені өзекті етеді. Жұмыстың негізгі мақсаты-қазақ тіліндегі экстремистік контентті автоматты түрде талдау үшін әдіснамалық негіз қалыптастыру және қолданыстағы тәсілдердің мүмкіндіктерін жүйелеу. Жұмыста экстремистік контенттің негізгі идеологиялық санаттары – пропаганда, рекрутмент және радикалдану-талданып, мәтінді осы бағыттар бойынша жіктеу міндеті қойылған. Зерттеу барысында VK, YouTube және Telegram платформаларынан жиналған қазақ тіліндегі материалдар негізінде баланс сақталған деректер жинағын құрастыру қарастырылған. Аннотациялау кезінде сарапшылардың пікір айырмашылықтары ескерілетіндігі, soft labels және disagreement-based learning тәсілдерін қолдану мүмкіндігі атап өтіледі. Аннотаторлар арасындағы пікір алшақтығы маңызды ақпараттық сигнал ретінде қарастырылып, disagreement-based learning тәсілі негізінде DART-4xBERT, DART-4xLLM және DART-Classic-Ensemble модельдері салыстырылады. Ең жоғары нәтиже DART-Classic-Ensemble жүйесінде тіркелді: F1=0.9664, бұл heterogeneous ансамбльдердің тиімділігін дәлелдейді. Экстремистік мазмұнды автоматты анықтау үшін классикалық және терең оқыту модельдері жайлы айтылады. Сондай-ақ мәтінді алдын ала өңдеудің модель дәлдігіне, нақтылығына әсері көрсетілген. Мақалада жасанды интеллект пен табиғи тілді өңдеу технологияларын қолдана отырып, қазақ тіліндегі әлеуметтік медиа контентін талдауға және экстремистік идеологияны анықтауға арналған кешенді әдістеме ұсынылады. Ұсынылған тәсіл ақпараттық қауіпсіздікті нығайтуға, сондай-ақ қоғамдағы идеологиялық тұрақтылықты қамтамасыз етуге бағытталған

Түйінді сөздер: экстремизм, әлеуметтік желілер, идеология, пропаганда, рекрутмент, радикалдану, жасанды интеллект, машиналық оқыту, табиғи тілді өңдеу

Дәйексөздер үшін: К.Д. Байсылбаева, Ш.Ж. Мусиралиева, Ж. Елтай (2026). Қазақ тіліндегі экстремистік идеологияны анықтау: аннотациялау мәселелері және терең оқыту тәсілдері // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 26. Б. 405–423. https://doi.org/10.54309/IJICT.2026.26.2.026. (Қаз. тіл.).

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

ОБНАРУЖЕНИЕ ЭКСТРЕМИСТСКОЙ ИДЕОЛОГИИ НА КАЗАХСКОМ ЯЗЫКЕ: ПРОБЛЕМЫ АННОТИРОВАНИЯ И МЕТОДЫ ГЛУБОКОГО ОБУЧЕНИЯ

К.Д. Байсылбаева[1]*, Ш.Ж. Мусиралиева[1], Ж.Елтай[2]

[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;

[2]Боннский Университет, Германия.

E-mail: baisylbaeva.k@gmail.com

Байсылбаева Кымбат Данияровна — старший преподаватель кафедры «Кибербезопасность и криптология», факультет информационных технологий и искусственно-

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

го интеллекта, Алматы, Казахстан
E-mail:baisylbaeva.k@gmail.com, https://orcid.org/0000-0001-9753-0398;
Мусиралиева Шынар Женисбековна — профессор-исследователь кафедры «Кибербезопасность и криптология», профессор, к.ф.м.н., факультет информационных технологий и искусственного интеллекта, Алматы, Казахстан
https://orcid.org/0000-0001-5794-3649;
Елтай Жастай — исследователь Боннского Университета, Германия
https://orcid.org/0000-0002-9275-7582.

© К.Д. Байсылбаева, Ш.Ж. Мусиралиева, Ж. Елтай

Аннотация. Деструктивный контент, распространяемый в настоящее время на платформах социальных сетей и мессенджерах, представляет серьезную угрозу информационной безопасности. Увеличение объема контента и усложнение методов его скрытой передачи усиливают необходимость автоматического обнаружения таких материалов. Нехватка средств и методов, особенно в направлении анализа текстов на казахском языке, делает эту проблему актуальной. Цель настоящего исследования-разработка автоматизированной методики, позволяющей эффективно выявлять признаки экстремистской идеологии в текстах социальных сетей на казахском языке. В работе проанализированы основные идеологические категории экстремистского контента- пропаганда, рекрутмент и радикализация-и поставлена задача классификации текста по этим направлениям. В ходе исследования предусмотрено составление сборника данных с сохранением баланса на основе материалов на казахском языке, собранных с платформ VK, YouTube и Telegram. При аннотировании отмечается, что учитываются различия во мнениях экспертов, возможность применения подходов soft labels и disagreement-based learning. Расхождения в оценках аннотаторов рассматриваются как информативный сигнал, на основе которого сравниваются модели DART-4xBERT, DART-4xLLM и DART-Classic-Ensemble. Наилучший результат достигнут системой DART-Classic-Ensemble (F1=0.9664), что подтверждает эффективность гетерогенных ансамблей. Говорится о классических моделях и моделях глубокого обучения для автоматического определения экстремистского содержания. Также показано влияние предварительной обработки текста на точность, эффективность модели. В статье предлагается комплексная методика анализа контента социальных сетей на казахском языке и выявления экстремистской идеологии с использованием искусственного интеллекта и технологий обработки естественного языка. Предлагаемый подход направлен на укрепление информационной безопасности, а также обеспечение идеологической стабильности в обществе.

Ключевые слова: экстремизм, социальные сети, идеология, пропаганда, рекрутмент, радикализация, искусственный интеллект, машинное обучение, обработка естественного языка

Для цитирования: К.Д. Байсылбаева, Ш.Ж. Мусиралиева, Ж.Елтай (2026). Обнаружение экстремистской идеологии на казахском языке: проблемы аннотиро-

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

вания и методы глубокого обучения // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 405–423. https://doi.org/10.54309/IJICT.2026.26.2.026. (На каз.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе.

Қазіргі цифрлық дәуірде интернет кеңістігінің қарқынды дамуы мен әлеуметтік желілердің таралуы ақпарат алмасу процесін жылдамдатуда. Бірақ бұл айтып отырған жағдай қоғам үшін жаңа қауіптер туғызуда-әсіресе экстремистік және деструктивті идеологиялардың таралуы күшейе түсті. Әлеуметтік желілер мен мессенджерлер радикал идеялар мен зорлық-зомбылықты насихаттаудың, жаңа мүшелер тартудың және идеологиялық ықпал етудің негізгі алаңына айналды. Мұндай контент қоғамдағы тұрақтылыққа, ұлттық қауіпсіздікке және ақпараттық кеңістіктің тұтастығына тікелей қауіп төндіреді.

Субъективті NLP тапсырмаларында аннотаторлар арасындағы келіспеушіліктерді шешу және бірыңғай шынайы белгіні қалыптастыру үшін көпшілік дауысы мен орташа мәнге негізделген тәсілдер кеңінен қолданылады. Алайда аннотаторлар бір-бірімен жүйелі түрде келіспеуі мүмкін, бұл әсіресе аффектіні, агрессияны және жеккөрушілік тілін анықтау сияқты субъективті тапсырмаларда олардың жеке көзқарастары мен құндылықтарын көрсетеді. Мұндай келіспеушіліктер маңызды ерекшеліктерді қамтуы мүмкін, бірақ аннотацияларды бір ғана шынайы белгіге біріктіру кезінде бұл ақпарат жиі ескерілмей қалады. (Aida Mostafazadeh Davani, Mark Díaz, Vinodkumar Prabhakaran, 2022) Осы мәселені шешу мақсатында бірнеше аннотациямен жұмыс істейтін модельдердің тиімділігі қарастырылады.

Ғылыми зерттеулерге сәйкес, соңғы жылдары интернет кеңістігінде зиянды және радикалды контенттің көлемі айтарлықтай артқан, ал оны қолмен бақылау тиімсіз болып табылады. Осыған байланысты табиғи тілдерді өңдеу және машиналық оқыту әдістері экстремистік мәтіндерді автоматты түрде анықтаудың негізгі құралдары ретінде қарастырылуда. Сондықтан әлеуметтік желілердегі экстремистік мазмұнды талдау және автоматты түрде анықтау-қазіргі ғылым мен киберқауіпсіздік саласындағы өзекті бағыттардың бірі (Yelmen et al., 2023; Kibriya et al., 2024). Атап айтқанда, терең оқыту модельдері, соның ішінде трансформер архитектурасына негізделген BERT сияқты модельдер, сондай-ақ дәстүрлі машиналық оқыту әдістерімен бірге қолданылып, мәтіндік деректерден семантикалық белгілерді тиімді анықтауға мүмкіндік береді.

Экстремистік идеология үш негізгі формада көрініс табады: пропаганда, рекрутмент және радикалдану. Пропаганда-белгілі бір діни немесе этникалық топтарға қарсы өшпенділік пен төзімсіздікті тудыру; рекрутмент-адамдарды экстремистік ұйымдарға тарту мен қосылуға итермелеу; радикалдану-тұлғаның біртіндеп радикалды көзқарастарды қабылдап, әрекетке бейімдеу үдерісі. Әлеуметтік медиа осы үшеуін де жылдам тарату құралына айналып отыр.

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

Кесте 1 – Қазақ тіліндегі әлеуметтік желілердегі экстремистік идеология түрлері және олардың сипаттамалары

Идеология түрі	Анықтамасы	Мысал
Рекрутмент	Экстремистік ұйымға ашық не жасырын түрде қосылуға үндейтін, қатысуға немесе қолдауға шақыратын, топты тартымды етіп көрсететін үндеу.	«Біздің қозғалысқа қосылып, әділ қоғам орнатуға атсалысыңыз!», «Бізбен бірге болыңыз – сізді жаңа өмір мен шынайы еркіндік күтеді!», «Біздің ұйым – нағыз ерлер мен ерік-тілердің жолы!»
Пропаганда	Белгілі бір этникалық, діни немесе басқа да топтарға қарсы өшпенділік пен кемсітуге итермелеу, бірақ әрекетке тікелей шақырусыз. Басқа топтардың зорлық-зомбылық әрекеттерін сипаттау, дезинформация тарату.	«Олар біздің мәдениетімізге қауіп төндіріп отыр», «Бұл топ шындықты жасырып келеді», «Билік бәрін өз пайдасына бұрып отыр»
Радикалдану	Әлеуметтік мәселелерге қатысты радикалды көзқарастарды насихаттап, адамдарды әрекетке және қарсы топтарға өшпенділікке итермелейтін мазмұн.	«Әділет үшін күресетін уақыт келді!», «Халық үнсіз отырмауы керек, біз әрекет етуіміз қажет!», «Енді төзуге болмайды – шын өзгеріс қарсылықтан басталады!»

Әдебиеттерге шолу

Соңғы жылдары онлайн әлеуметтік желілердегі идеологиялық бағыттылық пен экстремизмді анықтау мәселесі ғылыми зерттеулерде ерекше назарға алынған. (Ram et al., 2025) кең ауқымды онлайн орталарда идеологияны автоматты түрде анықтау жүйелерін құруға арналған практикалық нұсқаулықтар ұсынды. Олар онлайн идеологияны анықтау зорлық-зомбылықтың алдын алу және пікірдің өзгерісін зерттеу үшін маңызды екенін атап өтті, сондай-ақ эталондық деректерді құрудың күрделілігі мен прокси-сигналдардың (мысалы, хэштегтер, медиа) контекстке тәуелділігін көрсетті. Зерттеушілер төрт компонентті құрылымды ұсынып, деректерді жинау, проксилерді таңдау, гомофильдік сипаттамаларды шығару және классификацияны жүзеге асыру арқылы пайплайндар құрудың тиімді әдістерін көрсетті.

Сонымен қатар, (Rav et al., 2024) жүйелі шолу арқылы онлайн әлеуметтік желілерде идеологиялық бағыттылық пен экстремизмді анықтау әдістерін талдады. Зерттеуде Twitter, Facebook, Reddit сияқты әлеуметтік медиа платформалардан алынған мәтіндік және желілік деректер жиі қолданылған. Қолданылған әдістер қатарында дәстүрлі классификаторлар мен қазіргі заманғы терең оқыту модельдері салыстырмалы түрде қарастырылған. Бағалау метрикалары ретінде дәлдік (accuracy), F1-score және ROC-AUC сияқты стандартты көрсеткіштер қолданылған. Авторлар зерттеулердің көпшілігі тек ағылшын тілді контентке бағытталғанын, бұл көптілді және мәдени әртүрлі контекстілерде нәтижелердің тұрақсыз болуына алып келетінін атап өтеді.

(Erhard et al., 2024) PopBERT моделін пайдалана отырып, популизм мен оған байланысты идеологияларды зерттеген. Зерттеу 32 мыңнан астам депутаттық сөздер корпусына негізделген және антиэлитизм, халықшылдық бағыты бойынша аннотацияланған. Көп санатты (multi-label) классификацияға негізделген трансформер

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

модельдерін қолдану партиялар арасындағы популистік риториканың таралуын анықтауға және саяси дискурстың динамикасын жоғары дәлдікпен талдауға мүмкіндік беретіндігін көрсеткен.

(Ali et al., 2025) Израиль–Палестина қақтығысы контекстіндегі Reddit платформасындағы идеологиялық поляризацияны зерттеген. Пікірлер үш санатқа бөлініп аннотацияланған: Pro Israel, Pro Palestine және Neutral. Mixtral 8x7B тілдік моделі prompt engineering әдісімен бірге қолданылып, 74 % дәлдік пен шамамен 73 % макро F1 -score нәтижесін көрсетті. Бұл әдіс қазіргі ірі тілдік модельдердің қақтығысты әлеуметтік медиа кеңістіктерінде идеологиялық ұстанымдарды тиімді анықтай алатынын дәлелдеді. (Peng et al., 2024) зерттеуінде АҚШ-тағы 2020 жылғы президенттік сайлау аясында Twitter платформасындағы пайдаланушылардың саяси ұстанымдарын автоматты түрде анықтау мәселесі қарастырылған. Авторлар хэштегтерді саяси поляризация индикаторы ретінде пайдалана отырып, үлкен деректер жиынтығын (датасет) құрған. Талдау үшін пайдаланушы мен твит арасындағы өзара әрекеттестікті бейнелейтін екібөлікті граф моделі ұсынылған. Бұл тәсіл қолданушылардың мінез-құлықтық ерекшеліктерін ескеруге мүмкіндік береді. Ақпаратты тиімді пайдалану үшін Graph Convolutional Networks және skip aggregation деп аталатын жаңа механизм қолданылған. Эксперимент нәтижелері бұл модельдің дәстүрлі әдістерден жоғары өнімділігін және пайдаланушының мінез-құлқын ескерудің тиімділігін көрсетті.

(Mussiraliyeva et al., 2024) қазақ тіліндегі экстремистік мәтіндерді семантикалық талдау мәселесіне назар аударған. Зерттеу барысында діни экстремизмге қатысты мәтіндерді классификациялау үшін машиналық және терең оқытуға негізделген модельдер жасалған. Авторлар деректер теңгерімсіздігі мен қазақ тілінің ерекшеліктерін ескеріп, шешім ағашы, кездейсоқ орман (Random Forest), SVM, k-NN, логистикалық регрессия және аңғал Байес сияқты алгоритмдердің нәтижелерін салыстырған. Бұл тәсіл экстремистік контентті анықтаудың ең тиімді әдістерін айқындауға мүмкіндік берді.

(Shahin et al., 2024) Twitter-дегі араб тіліндегі қорлайтын сөйлеуді анықтау үшін бес санатты қамтитын көпкласты модель ұсынған. Модель ArabicBERT, 1D-CNN, BiLSTM және RBF әдістерін біріктіретін каскадты архитектураға негізделген. Ол жоғары дәлдікке (98,4%) және F1-score көрсеткішіне (98,4 %) қол жеткізіп, дәстүрлі классификаторлардан әлдеқайда жоғары нәтиже берді. Теңгерім жасалған деректер жиынтығы модельдің сенімді көпкласты және бинарлық классификация жасауына мүмкіндік берді.

(Ashraf et al., 2022) YouTube платформасындағы діни өшпенділік пен экстремизмге бағытталған пікірлерді анықтаумен айналысқан. Авторлар діни бейнелер астындағы пікірлерден жаңа деректер жиынтығын қалыптастырып, хабарламаларды сүзгіден өткізу және өшпенділік тудыратын бейнелерді анықтау үшін деректерді интеллектуалды талдау әдістеріне жүгінген. Негізгі классификация үшін SVM, логистикалық регрессия және k-NN алгоритмдері пайдаланылған. Нәтижелер жүйелі пікірталдаудың діни өшпенділік пен экстремизмді анықтауда тиімді болатынын

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

көрсетті.

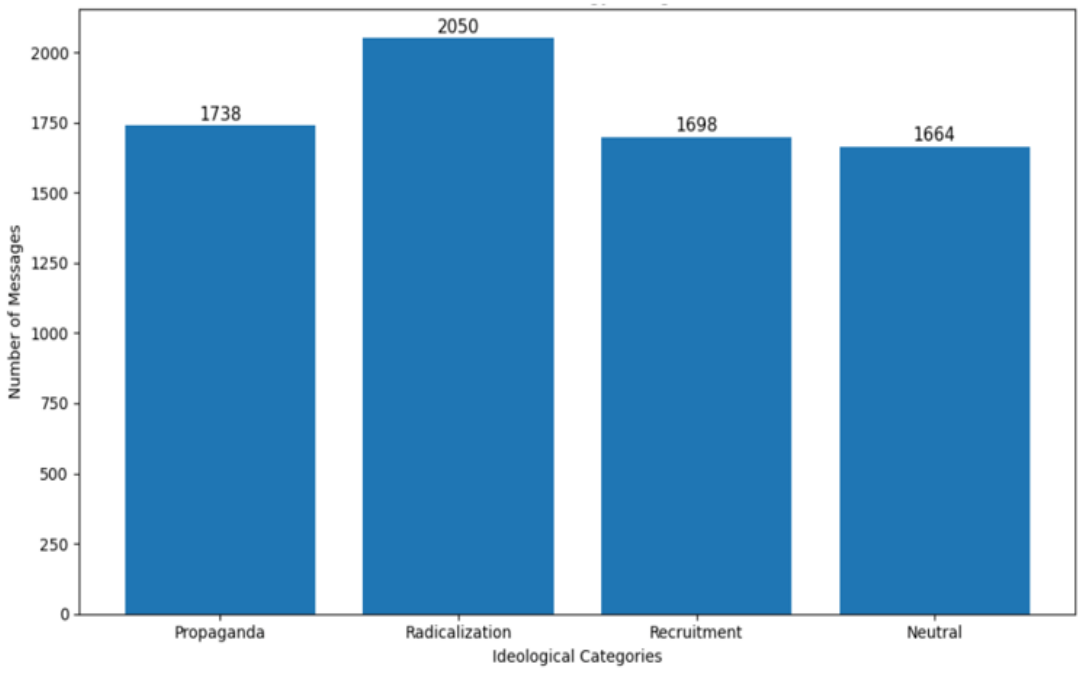
(Ramponi et al., 2022) әлеуметтік желілердегі діни өшпенділік сөзін автоматты түрде анықтау әдістерін әзірлеген. Авторлар ұсақталған аннотация схемасын ұсынып, ағылшын және итальян тілдерінде твиттер корпусын жасаған. Классификация үшін дәстүрлі машинамен оқыту және трансформер модельдері салыстырылған. Зерттеу модельдердің көптілді ортада бейімделу мүмкіндігін көрсетіп, шектеулі тілдік ресурстары бар тілдер үшін жаңа бағыттар ұсынады.

Талдаудан көрініп тұрғандай, қарастырылған зерттеулердің басым бөлігі ағылшын және басқа ресурсы мол тілдерге негізделген. Қазақ тілі сияқты агглютинативті және ресурсы шектелген тілдер үшін бұл әдістерді тікелей қолдану әркез тиімді бола бермейді. Сондықтан да қаза тілінде экстремистік идеологияны анықтау кезінде тілдік құрылымға бейімделген және деректер тапшылығын ескеретін тәсілдер қажет.

Әдістер мен материалдар.

Деректер жинау және аннотациялау

Экстремистік контентті автоматты түрде анықтаудың негізі-сапалы және баланс сақталған датасет құру болып табылады. Ол үшін әлеуметтік желілерден мәтіндік жарияланымдар мен пікірлер парсинг арқылы жиналып, құрылымдалған форматқа (CSV, JSON) түрлендіріледі. Деректер анонимдендіріліп, идеологиялық мазмұнға қарай төрт санатқа бөлінеді: бейтарап, пропаганда, рекрутмент, радикалдану. Корпустың сапасын тексеру үшін жиі кездесетін сөздер мен тіркестер талданады, бұл модельдің тілдік үлгілерді тануына мүмкіндік береді.



Сур. 1. Идеологиялық бағыттар бойынша мәтіндер санының үлестірімі

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

Зерттеу барысында VK, YouTube және Telegram платформаларынан жиналған қазақ тіліндегі мәтіндерден тұратын корпус құрылды (1-сурет). Жалпы корпус көлемі 7150 мәтінді құрады. Деректер төрт санат бойынша аннотацияланды: propaganda (1738), radicalization (2050), recruitment (1698) және neutral (1664). Әр мәтін кемінде төрт аннотатор тарапынан бағаланды. Аннотаторлар арасындағы disagreement семантикалық екіұштылықтың индикаторы ретінде қарастырылды.

Кесте 2 – Қазақ тіліндегі экстремистік контент корпусындағы идеологиялық санаттар мен жиі қолданылатын сөздер

Идеология түрі	Жиі кездесетін сөздер мен тіркестер
Радикалдану (Radicalization)	«жарақат алу», «қаза табу», «шабуыл жасау», «жарылыс жасау», «атып өлтіру», «оқ жаудыру»
Рекрутмент (Recruitment)	«содырлардың қатарына қосылу», «жиһад жасауға бірігу», «қатарымызға қосылу»
Пропаганда (Propaganda)	«пропаганда», «қарсы тұру», «ДАИШ (ИШИМ)», «Ислам мемлекеті»
Бейтарап (Neutral)	«сөз», «кету», «адам», «болу», «күн»

Аннотациялау кезеңі ерекше маңызға ие болып табылады, себебі сарапшылардың пікірі субъективті болуы мүмкін. Сондықтан әр мәтін бірнеше сарапшы тарапынан бағаланып, олардың пікір айырмашылықтары жұмсақ белгі (soft label) форматында сақтап алады. Мұндай тәсіл келіспеушілікке негізделген оқыту (disagreement-based learning) әдісіне негізделген. Бұл жағдайда әрбір мәтін бір категорияға нақты бекітілмей, ықтималдық үлестірімімен сипатталады. Бұл тәсіл белгісіздік пен субъективті айырмашылықтарды ескеріп, модельге күрделі контентті дәлірек өңдеуге, анықтауға, ажыратуға мүмкіндік береді. (Uma et al., 2021) көрсеткендей, келіспеушіліктерді жою емес, олардан үйрену деректер сапасын және модельдің түсіндірілу қабілетін арттырады.

Қолданылған деректерді бағалау барысында класстар арасындағы теңсіздік, аннотация сапасы, сонымен қатар, қазақ тіліне тән бірқатар ерекшеліктерді де ескеру қажет. Ең алдымен, субъективті немесе бағалау сипатындағы мәтіндерді талдауға арналған репрезентативті әрі ауқымды белгіленген корпустың болмауы елеулі шектеу болып табылады. Қолда бар деректер жиынтығы, әдетте, көлемі жағынан шағын, жеке жобалар аясында жасалған және қазақтілді мәтіндердің жанрлық, өңірлік және социолингвистикалық әртүрлілігін толық қамти бермейді, бұл өз кезегінде алынған нәтижелерді жалпылауды қиындатады.

Қазақ тілінің агглютинативті табиғаты мен мағыналардың жоғары контекстік тәуелділігі қосымша қиындықтар туындатады. Сөйлемнің семантикалық және прагматикалық салмағы көбіне жекелеген лексемалардан ғана емес, аффикстер жүйесі мен синтаксистік құрылымдар арқылы да қалыптасады. Бұл автоматтандырылған өңдеу процесін де, аннотация жүргізу ісін де күрделендіреді. Аталған ерекшеліктер ескерілмеген жағдайда класстардың бір-бірінен ажыратылуы бұрмаланып, белгілеу сапасының төмендеуіне алып келуі мүмкін.

Сонымен қатар, тілдің мәдени контексті де маңызды рөл атқарады, өйткені қазақ тіліндегі бағалау немесе әлеуетті токсикалық мазмұнды түсіндіру көп жағдай-

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

да қарым-қатынас нормалары мен әлеуметтік шарттылықтарға тәуелді. Әсіресе эвфемизмдер мен жасырын мағына беретін лексиканың кең қолданылуы назар аудартады, олар имплицитті негативті көзқарасты, ирония немесе астарлы бағалауды жеткізу үшін жиі пайдаланылады. Формалды түрде бейтарап көрінетін лексемалар белгілі бір контекстте қорлайтын немесе кемсітушілік сипатқа ие болуы мүмкін, сондықтан оларды прагматикалық және мәдени факторларды ескермей бірмәнді түрде жіктеу қиынға соғады. Осылайша, қазақ тіліндегі деректерді дұрыс бағалау терең контекстуалдық талдау, мәтіндік корпус құрастыру мен аннотациялаудың мамандандырылған тәсілдерін талап етеді.

Машиналық оқыту әдістері және модельдер

Экстремистік контентті анықтау үшін дәстүрлі және заманауи машиналық оқыту тәсілдері кеңінен қолданылады.

Классикалық әдістерге Naive Bayes, Logistic Regression, SVM және Random Forest т.с.с алгоритмдерді жатқызуға болады. Олар шағын деректерде жылдам және түсінікті нәтиже береді (Bolatbek et al., 2025).

Терең оқыту модельдері (LSTM, BERT, RoBERTa, CNN) мәтіннің семантикасы мен контекстін тереңірек түсінуге мүмкіндік береді (Hasib et al., 2023).

Ансамбльдік тәсілдер-бірнеше модельдің нәтижесін біріктіріп, болжам сапасын арттырады (мысалы, Stacking, Bagging, Boosting). Мұндай біріктіру әдістері күрделі идеологиялық мәтіндерде дәлдікті арттыруға айтарлықтай көмектеседі.

Мәтінді алдын ала өңдеу (Preprocessing)

Мәтіндік деректерді машиналық оқыту модельдеріне дайындау үшін бірнеше маңызды кезеңдер орындалады (Bolatbek et al., 2025).

1. Тазалау және қалыптандыру: артық белгілерді, сілтемелер мен эмодзилерді жою;

2. Токенизация

3. Stop-word сүзгілеу: мағынасыз сөздерді алып тастау және контексті сақтау.

4. Стемминг немесе лемматизация: сөздердің түбірін сақтау;

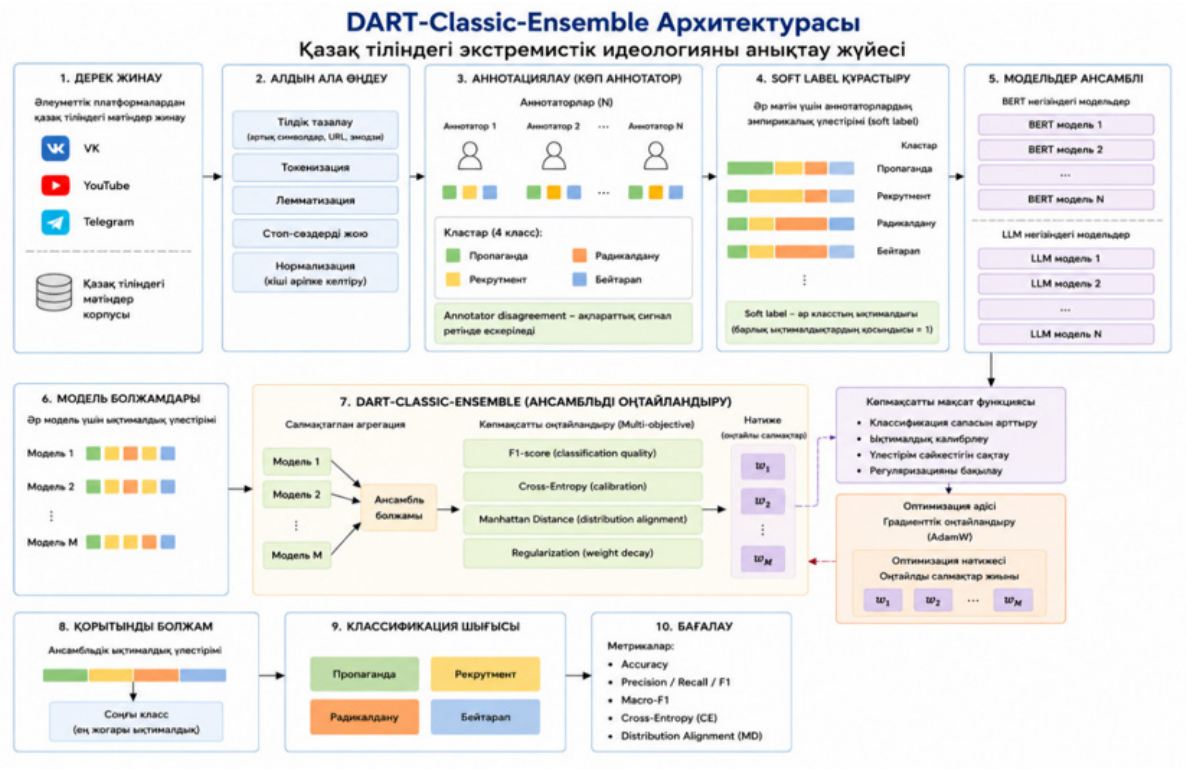
Бұл қадамдар модельге мәтінді құрылымды түрде ұсынуға және семантикалық байланыстарды тиімді анықтауға мүмкіндік береді.

Қолданылған алгоритмдер мен алдын ала өңдеу тәсілдерінің тиімділігі қазақ тілінің морфологиялық ерекшеліктерімен тікелей байланысты. Классикалық машиналық оқыту модельдері векторизация сапасына тәуелді болғандықтан, мәтінді тазалау, нормализациялау және жиілікке негізделген ерекшеліктерді таңдау олардың нәтижелілігін едәуір арттырады; ал трансформер типіндегі модельдер контекстік ақпаратты жақсы меңгергенімен, дұрыс токенизация мен эмбеддинг сапасына тәуелді. Қазақ тілінің агглютинативті құрылымы стеммингті қиындатып, бір түбірден көптеген формалардың жасалуына әкеледі, бұл стеммерлердің дәлдігін төмендетуі және сөздік кеңістікті шамадан тыс ұлғайтуы мүмкін. Сондықтан алдын ала өңдеу кезеңінде агглютинацияға бейімделген тәсілдерді қолдану, ал модельдер деңгейінде контекстік эмбеддингтерді пайдалану тиімді болып табылады (Ahmed et al., 2024).

Жалпы тәсіл және архитектура

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

Экстремистік контентті талдау көпқадамды жүйе болып табылады. Алдымен VK, YouTube және Telegram сияқты платформалардан мәліметтер жиналып, алдын ала өңделеді. Кейін олар машиналық және терең оқыту модельдерінде сынақтан өтеді. Нәтижелер дәлдік (accuracy), F1-score, precision, recall сияқты көрсеткіштерімен бағаланып, ең тиімді модель таңдалады. Мұндай архитектура қазақ тіліндегі әлеуметтік желі мәтіндерін автоматты талдау мен идеологиялық классификациялауда әмбебап шешім бола алады.



Сур. 2. Киберэкстремизмнің идеологиялық бағыттарын анықтау жүйесінің архитектурасы

2- суретте киберэкстремизмнің идеологиялық бағыттарын анықтау жүйесінің жалпы архитектурасы көрсетілген. Онда әлеуметтік желілерден (VK, YouTube, Telegram) жиналған деректердің өңдеу және модель құру кезеңдері бейнеленген.

Нәтижелер және оларды талқылау.

Жүргізілген салыстырмалы эксперименттерде келіспеушілікке негізделген үш тікелей салыстырылатын жүйе бағаланды (DART- Disagreement-Aware Representation Transformer): DART-4xBERT, DART-4xLLM және DART-Classic-Ensemble. Бағалау F1, CE және MD метрикалары бойынша жүргізілді, мұнда F1 мәнінің жоғары болуы классификация сапасының артқанын, ал CE мен MD мәндерінің төмен болуы модель ықтималдықтарды дәлірек бағалайтынын және аннотатор үлестіріміне жақынырақ келгенін көрсетеді.

Ұсынылған DART тәсілінде әлсіз модельдердің нәтижелері салмақталған ан-

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

самбльдік оқыту арқылы біріктірілді. Ансамбльдің соңғы болжамы әрбір модельдің ықтималдық нәтижелерінің салмақталған комбинациясы негізінде есептелді. Мұндай тәсіл әртүрлі модельдердің артықшылықтарын біріктіріп, аннотаторлар арасындағы пікір айырмашылықтарын тиімді ескеруге мүмкіндік береді.

Модельді оқыту барысында көпмақсатты оңтайландыру функциясы қолданылды. Негізгі loss функциясы классификация сапасын арттыруға арналған (L_{F1}), soft label үлестірімімен сәйкестікті бағалайтын (L_{CE}), сондай-ақ аннотаторлар мен модель болжамдарының айырмашылығын өлшейтін (L_{MD}) компоненттерінен тұрды. Сонымен қатар тұрақтандыру мүшесі модельдің шамадан тыс үйренуін болдырмау үшін енгізілді.

(L_{F1}), функциясы теңгерімсіздік жағдайында классификация сапасын арттыруға және F1-score көрсеткішін оңтайландыруға бағытталды. Cross-Entropy функциясы модель болжамдарын аннотаторлардың soft label үлестіріміне жақындату үшін қолданылды. Manhattan Distance функциясы модель ықтималдықтары мен аннотаторлар арасындағы disagreement деңгейін бағалауға мүмкіндік береді.

Soft label үлестірімі әр класс бойынша аннотаторлар саны негізінде есептелді. Ансамбльдік оқыту барысында модель салмақтарына шектеулер қойылып, олардың қосындысы бірге тең болатындай етіп қалыптандыру жүргізілді.

3–кестеде келіспеушілікке негізделген модельдер үшін жүргізілген абляциялық зерттеу нәтижелері келтірілген. Зерттеудің мақсаты-ұсынылған көпмақсатты оңтайландыру функциясының құрамындағы әрбір шығын компонентінің (L_F1, L_CE, L_MD) модель сапасына әсерін бағалау.

Нәтижелер бойынша DART-4xBERT моделінде ең жоғары нәтижелер жеке шығын функцияларын қолдану кезінде емес, бірнеше мақсаттық функцияларды біріктіру кезінде байқалды. Атап айтқанда, L_MD only конфигурациясы ең жоғары F1 көрсеткішіне (0.9450) және ең төмен MD мәніне (0.3159) қол жеткізді, бұл модельдің мақсатты ықтималдық үлестіріміне жақсырақ сәйкестігін көрсетеді. Сонымен қатар, L_CE only нұсқасы ең төмен Cross-Entropy (0.5647) нәтижесін көрсетті.

DART-4xLLM моделінде ең жоғары F1 нәтижесі біріктірілген шығын функцияларын қолдану кезінде алынды. Атап айтқанда, L_F1+L_CE және L_F1+L_CE+L_MD конфигурациялары үшін F1 көрсеткіші 0.9562 деңгейіне жетті. Бұл нәтижелер көпмақсатты оңтайландыру тәсілінің үлкен тілдік модельдермен бірге қолданылғанда жіктеу сапасын жақсарта алатынын көрсетеді. Сонымен қатар, тек Manhattan Distance қолданылған жағдайда MD көрсеткіші айтарлықтай төмендегенімен, жалпы классификация дәлдігі әрдайым максималды мәнге жетпегені байқалды.

DART-Classic-Ensemble моделі барлық конфигурациялар үшін тұрақты нәтижелер көрсетті. Барлық жағдайда F1 көрсеткіші 0.9664 деңгейінде сақталды, ал CE және MD көрсеткіштері сәйкесінше 0.5073 және 0.2960 шамасында өзгеріссіз қалды. Бұл классикалық ансамбльдік модельдің жеке шығын функцияларына сезімталдығы төмен екенін және тұрақты жұмыс істейтінін көрсетеді.

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

Кесте 3 – Келіспеушілікке негізделген модельдер үшін абляциялық зерттеу нәтижелері

Source	Case	F1	CE	MD
DART-4xBERT	L_F1 only	0.942218	0.576611	0.321257
DART-4xBERT	L_CE only	0.944082	0.564686	0.316206
DART-4xBERT	L_MD only	0.945014	0.565900	0.315857
DART-4xBERT	L_F1+L_CE	0.944082	0.567213	0.317336
DART-4xBERT	L_F1+L_MD	0.943150	0.567156	0.317224
DART-4xBERT	L_CE+L_MD	0.944082	0.565305	0.316422
DART-4xBERT	L_F1+L_CE+L_MD	0.943150	0.566219	0.316496
DART-4xLLM	L_F1 only	0.955266	0.540463	0.366450
DART-4xLLM	L_CE only	0.951538	0.540254	0.357964
DART-4xLLM	L_MD only	0.953402	0.558867	0.317810
DART-4xLLM	L_F1+L_CE	0.956198	0.540289	0.358123
DART-4xLLM	L_F1+L_MD	0.955266	0.559567	0.318941
DART-4xLLM	L_CE+L_MD	0.951538	0.549266	0.326870
DART-4xLLM	L_F1+L_CE+L_MD	0.956198	0.549643	0.327918
DART-Classic-Ensemble	L_F1 only	0.966449	0.507252	0.296019
DART-Classic-Ensemble	L_CE only	0.966449	0.507252	0.296019
DART-Classic-Ensemble	L_MD only	0.966449	0.507252	0.296019
DART-Classic-Ensemble	L_F1+L_CE	0.966449	0.507252	0.296019
DART-Classic-Ensemble	L_F1+L_MD	0.966449	0.507252	0.296019
DART-Classic-Ensemble	L_CE+L_MD	0.966449	0.507252	0.296019
DART-Classic-Ensemble	L_F1+L_CE+L_MD	0.966449	0.507252	0.296019

Жалпы алғанда, абляциялық зерттеу нәтижелері әртүрлі шығын функцияларының модель өнімділігіне әсер ететінін көрсетті. DART-4xBERT және DART-4xLLM модельдері үшін көпмақсатты шығындарды біріктіру белгілі бір метрикалар бойынша нәтижелерді жақсартқанымен, DART-Classic-Ensemble моделі анағұрлым тұрақты және теңгерімді нәтижелер көрсетті.

4–кестеде әртүрлі модельдер үшін регуляризацияның (тұрақтандырудың) әсері көрсетілген. Бағалау көрсеткіштері ретінде F1, Cross-Entropy (CE) және Manhattan Distance (MD) метрикаларының өзгерісі (Δ) пайдаланылды. Мұнда оң мән көрсеткіштің өскенін, ал теріс мән оның төмендегенін білдіреді.

Нәтижелер бойынша DART-4xBERT моделінде регуляризация белгілі бір конфигурациялар үшін өнімділікті жақсартқанын көруге болады. Әсіресе L_MD only жағдайында F1 көрсеткіші +0.002796-ға артқан, ал CE (-0.002809) және MD (-0.001233) мәндері төмендеген. Сонымен қатар, L_F1+L_MD конфигурациясында CE көрсеткішінің ең үлкен төмендеуі (-0.009561) және MD мәнінің жақсаруы (-0.004065) байқалды. Бұл регуляризацияның модельдің тұрақтылығын арттырып, ықтималдық үлестірімдерін дәлірек сәйкестендіруге ықпал еткенін көрсетеді.

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

Кесте 4 – Регуляризация (тұрақтандыру) әсерінің салыстырмалы нәтижелері

Source	Case	ΔF1	ΔCE	ΔMD
DART-4xBERT	L_F1 only	0.000000	-0.009044	-0.002399
DART-4xBERT	L_CE only	0.000000	-0.000612	-0.000448
DART-4xBERT	L_MD only	0.002796	-0.002809	-0.001233
DART-4xBERT	L_F1+L_CE	0.000000	-0.000089	-0.000052
DART-4xBERT	L_F1+L_MD	0.000932	-0.009561	-0.004065
DART-4xBERT	L_CE+L_MD	0.000932	-0.001044	-0.000615
DART-4xBERT	L_F1+L_CE+L_MD	0.000000	-0.001596	-0.001200
DART-4xLLM	L_F1 only	-0.000932	-0.007097	0.034436
DART-4xLLM	L_CE only	0.000000	0.000396	0.001354
DART-4xLLM	L_MD only	0.000932	0.000086	0.000217
DART-4xLLM	L_F1+L_CE	0.000000	-0.000018	0.000638
DART-4xLLM	L_F1+L_MD	0.000000	-0.000001	0.000000
DART-4xLLM	L_CE+L_MD	0.000932	0.000021	0.000429
DART-4xLLM	L_F1+L_CE+L_MD	0.000000	0.000005	0.000467
DART-Classic-Ensemble	L_F1 only	0.000000	-0.000768	0.000271
DART-Classic-Ensemble	L_CE only	0.000000	-0.000768	0.000271
DART-Classic-Ensemble	L_MD only	0.000000	-0.000768	0.000271
DART-Classic-Ensemble	L_F1+L_CE	0.000000	-0.000768	0.000271
DART-Classic-Ensemble	L_F1+L_MD	0.000000	-0.000768	0.000271
DART-Classic-Ensemble	L_CE+L_MD	0.000000	-0.000768	0.000271
DART-Classic-Ensemble	L_F1+L_CE+L_MD	0.000000	-0.000768	0.000271

DART-4xLLM моделінде регуляризация әсері айтарлықтай тұрақсыз сипат көрсетті. Кейбір жағдайларда метрикалардың өзгерісі өте аз болды немесе нәтижелерге елеулі ықпал етпеді. Атап айтқанда, L_F1 only конфигурациясында CE көрсеткіші төмендегенімен (-0.007097), MD көрсеткіші керісінше нашарлап (+0.034436) өсті. Басқа конфигурациялар үшін өзгерістер нөлге жақын болды. Бұл үлкен тілдік модельдерде тұрақтандыру әсерінің шектеулі екенін көрсетеді.

DART-Classic-Ensemble моделінде барлық конфигурациялар бойынша бірдей өзгерістер байқалды: F1 өзгермеген (ΔF1 = 0.000000), CE көрсеткіші аздап жақсарған (ΔCE = -0.000768), ал MD шамалы артқан (ΔMD = 0.000271). Бұл нәтижелер классикалық ансамбльдік тәсілдің бастапқыда тұрақты жұмыс істейтінін және регуляризация енгізілгеннен кейін көрсеткіштердің айтарлықтай өзгермейтінін көрсетеді.

Жалпы алғанда, жүргізілген талдау регуляризация әсерінің модель түріне тәуелді екенін көрсетті. DART-4xBERT моделі тұрақтандырудан көбірек пайда алған болса, DART-4xLLM моделінде әсер әлсіз байқалды, ал DART-Classic-Ensemble моделі жоғары тұрақтылық көрсетті.

5–кестеде идеологиялық бағыттарды классификациялау міндеті үшін әртүрлі агрегация стратегияларының тиімділігі салыстырылды. Зерттеуде үш тәсіл қарастырылды: көпшілік дауыс беру (Voting), қарапайым орташа мән арқылы біріктіру

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

(Averaging) және оңтайландырылған салмақталған агрегация (Optimised). Бағалау үшін F1, Cross-Entropy (CE) және Manhattan Distance (MD) метрикалары пайдаланылды.

Кесте 5 – Агрегация стратегиялары бойынша салыстырмалы нәтижелер

Source	Dataset	Method	F1	CE	MD
DART-4xBERT	Идеологиялық бағыттар	Voting	0.943150	0.565491	0.316266
DART-4xBERT		Averaging	0.943150	0.565491	0.316266
DART-4xBERT		Optimised	0.943150	0.566219	0.316496
DART-4xLLM		Voting	0.943150	0.613275	0.525520
DART-4xLLM		Averaging	0.945946	0.613275	0.525520
DART-4xLLM		Optimised	0.956198	0.549643	0.327918
DART-Classic-Ensemble		Voting	0.961789	0.492991	0.328111
DART-Classic-Ensemble		Averaging	0.959925	0.492991	0.328111
DART-Classic-Ensemble		Optimised	0.966449	0.507252	0.296019

DART-4xBERT моделі үшін барлық агрегация тәсілдері ұқсас нәтижелер көрсетті. Voting және Averaging әдістерінде көрсеткіштер толық сәйкес келіп, F1 = 0.943150, CE = 0.565491 және MD = 0.316266 мәндеріне жетті. Ал Optimised стратегиясы кезінде F1 көрсеткіші өзгермегенімен, CE (0.566219) және MD (0.316496) мәндері аздап өзгерді. Бұл жағдайда күрделі салмақталған агрегацияның айқын артықшылық бермегені байқалды.

DART-4xLLM моделінде агрегация тәсілдерінің әсері айқынырақ байқалды. Voting әдісі кезінде F1 көрсеткіші 0.943150 деңгейінде қалса, Averaging әдісі нәтижені жақсартып, F1 = 0.945946 мәніне жеткізді. Ең жоғары нәтиже Optimised тәсілінде алынды: F1 көрсеткіші 0.956198-ге дейін өсіп, CE мәні 0.549643-ке, ал MD көрсеткіші 0.327918-ге дейін төмендеді. Бұл нәтиже салмақтарды автоматты түрде оңтайландыру әрбір модельдің үлесін тиімді қайта бөлгенін көрсетеді. Оптимизация нәтижесінде алынған салмақтар тең емес болып қалыптасты, бұл кейбір модельдердің шешім қабылдауға көбірек үлес қосқанын білдіреді.

DART-Classic-Ensemble моделі үшін де агрегация стратегиялары арасындағы айырмашылық байқалды. Voting әдісі F1 = 0.961789 нәтижесін көрсетсе, Averaging тәсілінде көрсеткіш аздап төмендеді (F1 = 0.959925). Ал ең жоғары нәтиже Optimised стратегиясында алынды: F1 = 0.966449 және MD = 0.296019. Бұл ансамбль ішінде модельдердің әртүрлі салмаққа ие болуы жалпы сапаны жақсартатынын көрсетеді.

Жалпы алғанда, алынған нәтижелер қарапайым дауыс беру мен орташа мән арқылы біріктіру әдістеріне қарағанда оңтайландырылған агрегация стратегиясының тиімдірек екенін көрсетті. Әсіресе DART-4xLLM және DART-Classic-Ensemble модельдері үшін салмақтарды автоматты түрде бейімдеу классификация сапасын арттыруға және модельдер арасындағы келіспеушілікті тиімді пайдалануға мүмкіндік берді.

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

DART-4xBERT құрылымын кеңейтуге арналған қосымша эксперименттер

Негізгі DART нәтижелерін алғаннан кейін ұсынылған құрылымның тұрақтылығы мен кеңейтілу мүмкіндігін бағалау үшін үш қосымша эксперимент жүргізілді: optimization algorithm ауыстыру, temporal disagreement моделін енгізу және ensemble diversity metric негізіндегі қайта салмақтау. Бұл тәжірибелер негізгі DART 4xBERT конфигурациясын толық алмастыру үшін емес, оның сезімталдығын, орнықтылығын және қосымша жақсарту мүмкіндіктерін тексеру үшін орындалды.

Бірінші қосымша экспериментте ансамбль салмақтарын іздеудің балама тәсілі ретінде Particle Swarm Optimization (PSO) қолданылды. Бұл жерде weak predictor-лер өзгертілмей, тек бастапқы Differential Evolution алгоритмінің орнына жаңа оңтайландыру әдісі пайдаланылды. Нәтижелер бойынша PSO әдісі негізгі DART нәтижесін іс жүзінде қайталады. Сонымен қатар cross-entropy кішкене өсіп, ал manhattan distance аздап төмендеді. Бұл бастапқы Differential Evolution шешімінің жеткілікті тиімді болғанын және PSO әдісінің осы нәтижені тұрақты түрде қайталай алатынын көрсетті.

Екінші экспериментте уақыттық келіспеушілік (temporal disagreement) тәсілі зерттелді. Датасетте нақты уақыт белгілері болмағандықтан, уақыттық контекст мәтіннен алынған уақыттық-оқиғалық жанама топтар арқылы модельденді. Хабарламалар future_mobilization, ongoing_event, post_event, political_theme және atemporal_other санаттарына бөлініп, әр санат үшін жеке ансамбль салмақтары үйретілді. Нәтижесінде temporal-aware модель baseline нұсқасымен салыстырғанда жоғары нәтиже көрсетті. Ең жоғары келіспеушілік деңгейі political_theme санатында байқалса, ең айқын өсім future_mobilization тобында тіркелді. Бұл annotator disagreement көрсеткішінің мазмұнның уақыттық-оқиғалық контекстіне тәуелді өзгеруі мүмкін екенін көрсетті.

Үшінші экспериментте ансамбль ішіндегі модельдер әртүрлілігін бағалау үшін косинустық әртүрлілік (cosine diversity), Дженсен–Шеннон әртүрлілігі (Jensen–Shannon diversity) және ансамбльдік келіспеушілік коэффициенті (ensemble disagreement coefficient) қолданылды. Алынған нәтижелер бойынша disagreement coefficient ең тиімді тәсіл болды: accuracy және macro-F1 өзгермегенімен, cross-entropy кішкене көрсеткішке өсіп, ал Manhattan distance төмендеді. Ал қалған diversity-aware тәсілдері baseline нәтижесінен асып түспеді.

Жалпы, қосымша эксперименттер ішінде ең практикалық тиімді нәтиже temporal disagreement тәсілінде байқалды, себебі ол классификация сапасын нақты жақсартты. Ал PSO және diversity-aware weighting әдістері негізінен DART жүйесінің тұрақтылығын, ансамбль құрылымын және ықтималдық калибрлеуін талдау үшін пайдалы болды.

Қазақ тіліндегі психолингвистикалық мәтіндерді жіктеу үшін жүргізілген эксперименттердің нәтижелері де алынған болатын. Зерттеудің мақсаты-аннотаторлар арасындағы келіспеушіліктерді (Disagreement) ескеріп өңделген деректер негізінде ең тиімді модельдік комбинацияны анықтау болды. Модельдердің тиімділігі Precision, Recall, Accuracy, F1-score және ROC AUC метрикалары бойынша бағаланды.

Бірінші комбинацияда Disagreement өңдеуі мен PLAM белгілері TF-IDF векторизациясымен және дәстүрлі машиналық оқыту алгоритмдерімен бірге қолданылды.

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

Нәтижелер бойынша ең жоғары көрсеткіштерді SVC моделі көрсетті. Бұл модель барлық бағалау метрикалары бойынша басқа классикалық алгоритмдерден жоғары нәтиже беріп, дәстүрлі тәсілдер ішінде ең тиімдісі ретінде анықталды.

Екінші комбинацияда мәтіндерді ұсыну үшін BERT эмбеддингтері пайдаланылып, олар PLAM белгілерімен біріктірілді. Бұл тәсіл классикалық модельдермен салыстырғанда жоғары нәтижелерге қол жеткізді. Accuracy көрсеткіші 95,1 %-ға жетіп, ROC AUC мәні 99,6%-дан асты. Бұл нәтиже контекстік тілдік модельдердің психолингвистикалық ерекшеліктерді тиімді түрде бейнелей алатынын көрсетті.

Үшінші комбинацияда BERT эмбеддингтері LSTM нейрондық желісі арқылы өңделіп, кейін PLAM белгілерімен біріктірілді. Алынған нәтижелер де өте жоғары болғанымен, негізгі метрикалар бойынша BERT + PLAM комбинациясынан сәл төмен болды. Бұл жағдайда BiLSTM қабатын қосу модельдің сапасын айтарлықтай арттырмағаны байқалды.

Жалпы алғанда, эксперимент нәтижелері бойынша ең жоғары көрсеткіштерді Disagreement + PLAM + BERT комбинациясы көрсетті. Сондықтан қазақ тіліндегі психолингвистикалық мәтіндерді жіктеу міндетінде дәл осы тәсіл ең тиімді модель ретінде таңдалды.

Кесте 7 – DART құрылымды кеңейтуге арналған қосымша эксперименттердің нәтижелері

Эксперимент	Нұсқа	Accuracy		ROC AUC	CE	MD
Optimization algorithm	Differential Evolution	0.9432	0.9421	0.9961	0.5678	0.3177
Optimization algorithm	PSO	0.9432	0.9421	0.9961	0.5678	0.3177
Temporal disagreement	Global baseline	0.9441	0.9434	0.9962	0.5503	0.1906
Temporal disagreement	Temporal-aware DART	0.9459	0.9450	0.9962	0.5496	0.1902
Diversity metric	Baseline DART	0.9441	0.9434	0.9962	0.5503	0.1906
Diversity metric	Disagreement coefficient-aware	0.9441	0.9434	0.9962	0.5502	0.1906

Жүргізілген эксперименттік зерттеулер модельдер арасындағы әдіснамалық емес, нақты тәжірибелік айырмашылықтарды анықтауға мүмкіндік береді. Классикалық машиналық оқыту модельдері ішінде Logistic Regression+PLAM, ал трансформерлік арасында BERT+LSTM моделі барлық метрикалар бойынша жоғары нәтиже бергендігін (Mussiraliyeva, Bolatbek, & Baisylbayeva, 2025) зерттеу жұмысынан көруге болады. Аталған гибрид тәсіл қазақ тілінің агглютинативті құрылымы мен семантикалық күрделілігін ескергенде контекст пен реттілікті бір уақытта модельдеу тұрғысынан ең ұтымды тәсіл екенін айқындайды. Сондықтан қолданылған әдістердің тиімділігі теориялық емес, қазақ тіліндегі нақты корпусқа негізделген эксперименттік талдау арқылы дәлелденді.

Қорытынды.

Жүргізілген шолу нәтижесінде әлеуметтік желілер мен мессенджерлердегі экстремистік контентті анықтау мәселесі заманауи цифрлық қоғамның ең өзекті бағыттарының бірі екені анықталды. Интернеттегі ақпарат көлемінің қарқынды өсуі

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

мен пайдаланушылардың белсенділігі экстремистік және деструктивті идеялардың таралуына қолайлы орта туғызып жатыр. Мұндай мазмұн қоғамның тұрақтылығына, ұлттық қауіпсіздік пен ақпараттық кеңістікке тікелей қауіп төндіреді, сондықтан оны дер кезінде анықтап, автоматты түрде сүзгіден өткізу аса маңызды.

Экстремистік идеологияның негізгі түрлері-пропаганда, рекрутмент және радикалдану-мазмұны мен мақсаттары жағынан ерекшеленгенімен, үшеуі де қоғамға идеологиялық ықпал етуге бағытталған. Қазақ тіліндегі әлеуметтік медиа кеңістігінде мұндай контентті талдау мен классификациялау үшін сапалы датасет құру керек, аннотациялау сапасын арттыру керек және машиналық оқыту әдістерін тиімді қолдану қажет.

Деректерді жинау кезеңінде әлеуметтік желілерден алынған мәтіндік контент құрылымдалып, төрт негізгі классқа (бейтарап, пропаганда, рекрутмент, радикалдану) бөлінеді. Бұл корпус қазақ тіліндегі идеологиялық мазмұнды зерттеуге арналған маңызды база қалыптастырады. Ал аннотациялау процесінде сарапшылар арасындағы пікір айырмашылықтарын ескеретін disagreement-based learning тәсілі енгізілуі деректер сапасын жақсарта түседі. Мұндай әдіс мәтіндердің екіжақтылығын және белгісіздік деңгейін ескере отырып, модельге идеологиялық мазмұнды дәлірек әрі нақтырақ түсінуге мүмкіндік жасайды.

Машиналық оқыту тұрғысынан дәстүрлі әдістер мен терең оқыту архитектуралары біріге отырып, ансамбльдік модельдер арқылы нәтижелердің дәлдігі арттыруға болады.

Мәтінді алдын ала өңдеу кезеңдері-тазалау, қалыптандыру, стемминг, векторизация және токенизация-модельдің дәлдігі мен сенімділігін қамтамасыз ететін маңызды техникалық негіз болып табылады. Осы әдістердің кешені әлеуметтік желі деректерін автоматтандырылған түрде өңдеуге, идеологиялық ерекшеліктерін талдауға және экстремистік мазмұнды уақтылы әрі нақты анықтауға мүмкіндік туындатады.

Жұмыстың негізгі мақсаты-қазақ тіліндегі экстремистік контентті автоматты түрде талдау үшін әдіснамалық негіз қалыптастыру және қолданыстағы тәсілдердің мүмкіндіктерін жүйелеу болғандықтан, зерттеу бірінші кезеңде теориялық құрылымды құруға бағытталды. Әдебиеттерге сүйене отырып жасалған салыстырмалы талдау нәтижелері архитектура, аннотация стратегиялары және модельдік таңдау бойынша авторлық гипотезаларды анықтауға мүмкіндік берді.

Жалпы алғанда, зерттеу көрсеткендей, қазақ тіліндегі әлеуметтік желілер мен мессенджерлердегі экстремистік контентті анықтау үшін жасанды интеллект, табиғи тілді өңдеу, келіспеушілік негізіндегі оқыту әдістерін біріктірген кешенді жүйе тиімді шешім болып табылады. Мұндай тәсіл киберқауіпсіздік деңгейін арттыруға, қоғамдағы идеологиялық тұрақтылықты сақтауға және отандық ақпараттық кеңістікті қорғауға елеулі үлес қоса алады.

ӘДЕБИЕТТЕР

Ali H.J., Abrar A., Hossain S.M.H., Mridha M.F. (2025). Social media polarization during conflict: Insights from an ideological stance dataset on Israel-Palestine Reddit comments // arXiv. Article 2502.00414. DOI:10.48550/arXiv.2502.00414 (in Eng.)

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

Ashraf N., Rafiq A., Butt S., Shehzad H., Sidorov G., Gelbukh A. (2022). YouTube-based religious hate speech and extremism detection dataset with machine learning baselines // Journal of Intelligent & Fuzzy Systems. — Vol. 42. — No. 3. Pp. 1–9. DOI: 10.3233/JIFS-219264 (in Eng.)

Ahmed A., Refat R. (2024). Exploring the impact of stemming on text topic-based classification accuracy // Journal of Linguistics, Culture and Communication. — Vol. 2. — Pp. 204–224. DOI: 10.61320/jolcc.v2i2.204-224 (in Eng.)

Bolatbek M., Mussiraliyeva Sh., Sagynai M., Zhastay E. (2025). Machine learning methods for identifying destructive texts in the Kazakh language // Proceedings of the International Conference on Algorithms, Data Mining, and Information Technology (ADMIT). DOI: 10.1145/3701100.3701121 (in Eng.)

Erhard L., Hanke S., Remer U., Falenska A., Heiberger R.H. (2024). PopBERT: Detecting populism and its host ideologies in the German Bundestag // Political Analysis. — Vol. 33. — No. 1. Pp. 1–17. DOI: 10.1017/pan.2024.12 (in Eng.)

Hasib K.M., Azam S., Karim A., Marouf A.A., Shamrat F.M.J.M., Montaha S., Yeo K.C., Jonkman M., Alhajj R., Rokne J. G. (2023). MCNN LSTM: Combining CNN and LSTM to classify multi-class text in imbalanced news data // IEEE Access. — Vol. 11. — Pp. 93048–93063. DOI: 10.1109/ACCESS.2023.3309697 (in Eng.)

Kibriya A., Siddiqa A., Khan W.Z., Khan M.K. (2024). Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification // Computer and Electrical Engineering. — Vol. 116. — Article 109153. DOI: 10.1016/j.compeleceng.2024.109153 (in Eng.)

Mostafazadeh Davani A., Díaz M., Prabhakaran V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations // Transactions of the Association for Computational Linguistics. — Vol. 10. — Pp. 92–110. DOI:10.1162/tacl_a_00449 (in Eng.)

Mussiraliyeva S., Bolatbek M., Baisylbayeva K. (2025). Extremist ideology classification in Kazakh: A multi-class approach using machine learning and psycholinguistic analysis // IEEE Access. DOI:10.1109/ACCESS.2025.3596601 (in Eng.)

Mussiraliyeva S.Zh., Bolatbek M.A., Zhumakhanova A.N., Baispay G., Medetbek Zh. (2024). Creating a model of semantic analysis of extremist texts in the Kazakh language // Journal of Mathematics, Mechanics and Computer Science. — Vol. 121. — No. 1. DOI:10.26577/JMMCS2024121111 (in Eng.)

Peng X., Zhou Z., Zhang C., Xu K. (2024). Online social behavior enhanced detection of political stances in tweets // Proceedings of the International AAAI Conference on Web and Social Media. — Vol. 18. — No. 1. Pp. 1207–1219. DOI:10.1609/icwsm.v18i1.31383 (in Eng.)

Ram R., Thomas E., Kernot D., Rizoiu M.A. (2025). Practical guidelines for ideology detection pipelines and psychosocial applications // Proceedings of the International AAAI Conference on Web and Social Media. — Vol. 19. — Pp. 1630–1648. DOI:10.1609/icwsm.v19i1.35892 (in Eng.)

Ravi K., Yuan J.S. (2024). Ideological orientation and extremism detection in online social networking sites: A systematic review // Intelligent Systems with Applications. — Vol. 24. — Article 200456. DOI:10.1016/j.iswa.2024.200456 (in Eng.)

Ramponi A., Testa B., Tonelli S., Jezek E. (2022). Addressing religious hate online: From taxonomy creation to automated detection // PeerJ Computer Science. — Vol. 8. — Article e1128. DOI: 10.7717/peerj-cs.1128 (in Eng.)

Shahin M.A., Nassif A.B., Elnagar A. (2024). Detection of Arabic offensive language in social media using machine learning models // Intelligent Systems with Applications. — Vol. 22. — Article 200376. DOI: 10.1016/j.iswa.2024.200376 (in Eng.)

Uma A.N., Fornaciari T., Paun S., Plank B., Poesio M. (2021). Learning from disagreement: A survey // Journal of Artificial Intelligence Research. — Vol. 72. — Pp. 1385–1470. DOI: 10.1613/jair.1.12752 (in Eng.)

Hasib K.M., Azam S., Karim A., Marouf A.A., Shamrat F.M.J.M., Montaha S., Yeo K.C., Jonkman M., Alhajj R., Rokne J.G. (2023). MCNN LSTM: Combining CNN and LSTM to classify multi-class text in imbalanced news data // IEEE Access. — Vol. 11. — Pp. 93048–93063. DOI:10.1109/ACCESS.2023.3309697 (in Eng.)

Xu J., Theune M., Braun D. (2024). Leveraging Annotator Disagreement for Text Classification // Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024). Pp. 1–10. DOI:10.48550/arXiv.2409.17577 (in Eng.)

Yelmen A., Gunes A., Zontul M. (2023). Multi-class document classification using lexical ontology-based deep learning // Applied Sciences. — Vol. 13. — No. 10. Article 6139. DOI:10.3390/app13106139 (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.2. Number 26 (2026). Pp. 424–441
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.017
УДК 81.93.29

DEVELOPMENT AND RESEARCH OF A METHOD FOR ANALYZING NETWORK TRAFFIC TO IDENTIFY A CYBER THREAT

M.A. Bolatbek[1]*, A.M.Usmanova[1], K.B. Bagitova[2], G.B. Baispay[3]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[3]University of Illinois Urbana-Champaign (UIUC).
E-mail: bolatbek.milana@gmail.com

Milana Bolatbek — PhD, associate professor of the Department of Cybersecurity and Cryptology, Faculty of Information Technologies, Al-Farabi Kazakh national university, Almaty, Kazakhstan
E-mail: bolatbek.milana@gmail.com, https://orcid.org/0000-0002-2153-180X;
Asel Usmanova — PhD, Senior lecturer of the Department of Cybersecurity and Cryptology, Faculty of Information Technologies, Al-Farabi Kazakh national university, Almaty, Kazakhstan
https://orcid.org/0009-0004-3411-1881;
Kalamkas Bagitova — PhD, Associate Professor, Head of the Computer Science Department, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0003-1587-1995;
Gulshat Baispay — Researcher, University of Illinois Urbana-Champaign (UIUC)
https://orcid.org/0000-0003-4292-2938.

© M.A. Bolatbek, A.M. Usmanova, K.B. Bagitova, G.B. Baispay

Abstract. In the modern world, it begins with the understanding that information technology plays a decisive role in the life of society. Increasing dependence on digital systems leads to an increase in cyber threats, which requires the development of effective tools for their detection and prevention. With the increase in the volume and complexity of network traffic, new problems arise for security systems. Cybercriminals are continuously improving their attack methods, making them complex and difficult to find. In this context, network traffic analysis becomes an important tool in the arsenal of cybersecurity professionals. With a detailed analysis of the data transmitted over the network, it is possible to detect not only active cyberattacks, but also attempts to prepare for them, such as port scanning or spreading malicious code. The purpose of this work is to develop a method

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

that allows to effectively analyze network traffic in order to detect signs of cyber threats.

Keywords: network traffic, log file, cyber threat, information security, information technology

For citation: M.A. Bolatbek, A.M. Usmanova, K.B. Bagitova, G.B. Baispay (2026). Development and research of a method for analyzing network traffic to identify a cyber threat // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 424–441. https://doi.org/10.54309/IJICT.2026.26.2.027. (In Eng.).

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

КИБЕР ҚАУІПТІ АНЫҚТАУ ҮШІН ЖЕЛІЛІК ТРАФИКТІ ТАЛДАУ ӘДІСІН ӘЗІРЛЕУ ЖӘНЕ ЗЕРТТЕУ

М.А. Болатбек[1]*, А.М. Усманова[1], Қ.Б. Багитова[2], Г.Б. Байспай[3]

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Х.Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;
[3]Урбане-Шампейндағы Иллинойс Университеті (UIUC).
E-mail: bolatbek.milana@gmail.com

Болатбек Милана — PhD, Киберқауіпсіздік және криптология кафедрасының қауымдастырылған профессоры, Ақпараттық технологиялар факультеті, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
E-mail: bolatbek.milana@gmail.com, https://orcid.org/0000-0002-2153-180X;
Усманова Асел — PhD, Ақпараттық технологиялар факультеті, Киберқауіпсіздік және криптология кафедрасының аға оқытушысы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
https://orcid.org/0009-0004-3411-1881;
Багитова Қаламқас — PhD, қауымдастырылған профессор, Компьютер ғылымдары кафедрасының меңгерушісі, Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан
https://orcid.org/0000-0003-1587-1995;
Байспай Гүлшат — зерттеуші, Иллинойс университеті Урбана-Шампейн (UIUC)
https://orcid.org/0000-0003-4292-2938.

© М.А. Болатбек, А.М. Усманова, Қ.Б. Багитова, Г.Б. Байспай

Аннотация. Қазіргі таңда ақпараттық технологиялар қоғам өмірінде шешуші рөл атқарады. Цифрлық жүйелерге тәуелділіктің артуы киберқауіптердің көбеюіне әкеледі, бұл оларды анықтау мен алдын алудың тиімді құралдарын әзірлеуді талап етеді. Желілік трафиктің көлемі мен күрделілігінің артуы қауіпсіздік жүйелері үшін жаңа мәселелер туындатады. Киберқылмыскерлер шабуыл әдістерін үнемі жетілдіріп отырады, бұл оларды уақтылы табуды қиындатады. Бұл тұрғыда желілік трафикті талдау киберқауіпсіздік мамандарының арсеналындағы маңызды құралға

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

айналуда. Желі арқылы берілетін деректерді егжей-тегжейлі талдау арқылы белсенді кибершабуылдарды ғана емес, сонымен қатар портты сканерлеу немесе зиянды кодты тарату сияқты оларға дайындалу әрекеттерін де анықтауға болады. Бұл жұмыстың мақсаты-киберқауіптердің белгілерін анықтау мақсатында желілік трафикті тиімді талдауға мүмкіндік беретін әдісті әзірлеу.

Түйінді сөздер: желілік трафик, журнал файлы, киберқауіп, ақпараттық қауіпсіздік, ақпараттық технология

Дәйексөздер үшін: М.А. Болатбек, А.М. Усманова, Қ.Б. Багитова, Г.Б. Байспай (2026). Кибер қауіпті анықтау үшін желілік трафикті талдау әдісін әзірлеу және зерттеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 424–441. https://doi.org/10.54309/IJICT.2026.26.2.027. (Ағыл. тіл.).

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

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДА АНАЛИЗА СЕТЕВОГО ТРАФИКА ДЛЯ ВЫЯВЛЕНИЯ КИБЕРУГРОЗЫ

М.А. Болатбек[1]*, А.М. Усманова[1], К.Б. Багитова[2], Г.Б. Байспай[3]

[1]Казахский нациоальный университет имени Аль-Фараби, Алматы, Казахстан;
[2]Атырауский университет им. Х. Досмухамедова, Атырау, Казахстан;
[3]Университет Иллинойса в Урбане-Шампейне (UIUC).
E-mail: bolatbek.milana@gmail.com

Болатбек Милана — PhD, ассоциированный профессор кафедры кибербезопасности и криптологии факультета информационных технологий Казахского национального университета имени Аль-Фараби, Алматы, Казахстан
E-mail: bolatbek.milana@gmail.com, https://orcid.org/0000-0002-2153-180X;
Усманова Асел — PhD, старший преподаватель кафедры кибербезопасности и криптологии факультета информационных технологий Казахского национального университета имени Аль-Фараби, Алматы, Казахстан
https://orcid.org/0009-0004-3411-1881;
Багитова Каламкас — PhD, ассоциированный профессор, заведующая кафедрой компьютерных наук Атырауского университета имени Х. Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0003-1587-1995;
Байспай Гульшат — исследователь, Университет Иллинойса в Урбана-Шампейн (UIUC)
https://orcid.org/0000-0003-4292-2938.

© М.А. Болатбек, А.М. Усманова, Қ.Б. Багитова, Г.Б. Байспай

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

Аннотация. В современном мире информационные технологии играют решающую роль в жизни общества. Растущая зависимость от цифровых систем приводит к росту киберугроз, что требует разработки эффективных инструментов для их обнаружения и предотвращения. С увеличением объема и сложности сетевого трафика возникают новые проблемы для систем безопасности. Киберпреступники постоянно совершенствуют свои методы атак, усложняя их и затрудняя обнаружение. В этом контексте анализ сетевого трафика становится важным инструментом в арсенале специалистов по кибербезопасности. Благодаря детальному анализу данных, передаваемых по сети, можно обнаружить не только активные кибератаки, но и попытки подготовиться к ним, такие как сканирование портов или распространение вредоносного кода. Целью данной работы является разработка метода, позволяющего эффективно анализировать сетевой трафик с целью выявления признаков киберугроз.

Ключевые слова: Сетевой трафик, файл журнала, киберугроза, информационная безопасность, информационные технологии

Для цитирования: М.А. Болатбек, А.М. Усманова, Қ.Б. Багитова, Г.Б. Байспай (2026). Разработка и исследование метода анализа сетевого трафика для выявления киберугрозы // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 424–441. https://doi.org/10.54309/IJICT.2026.26.2.027. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

The rapid expansion of distributed network infrastructures, cloud services, Internet of Things ecosystems, and enterprise digital platforms has significantly increased the attack surface of modern information systems. Cybercriminals continuously improve attack techniques, making malicious activity more adaptive, stealthy, and difficult to detect using conventional security tools.

Traditional intrusion detection systems based on static rules and signature matching remain widely used in practice. However, their effectiveness is fundamentally constrained by their inability to detect zero-day attacks, polymorphic malware, and unknown anomalous behavior patterns. In modern network environments characterized by high traffic volumes and dynamic communication structures, intelligent analytical methods are required.

Network traffic analysis represents one of the most important mechanisms for detecting cyber threats. Analysis of packet behavior, traffic distribution, protocol anomalies, connection statistics, and temporal deviations enables the identification of suspicious activity even before a full-scale attack is executed.

Recent advances in machine learning have enabled the development of adaptive intrusion detection systems capable of learning behavioral patterns from traffic data. Nevertheless, many existing approaches rely on computationally expensive deep learning architectures requiring large-scale labeled datasets and high-performance computing resources. Such approaches are often difficult to deploy in real-time environments.

The scientific novelty of this research lies in the development of a computationally

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

efficient machine learning framework that combines:

- structured log-based traffic analysis;
- feature engineering;
- statistical feature selection;
- lightweight classification models;
- real-time anomaly detection capability.

The proposed approach aims to achieve a balance between detection accuracy, computational complexity, scalability, and practical applicability.

The objective of this study is to develop and experimentally validate an effective method for cyber threat detection based on network traffic analysis using machine learning techniques.

Literature review

Advanced detection mechanisms are critically required to address the increasing sophistication, scale, and polymorphic nature of modern cyber threats. Conventional rule-based approaches, including signature-based intrusion detection and prevention systems (IDS/IPS), remain widely deployed; however, their effectiveness is fundamentally limited by their dependence on predefined patterns, which makes them incapable of identifying zero-day attacks and evolving threat behaviors (Mohammadi et al., 2021). As a result, a substantial portion of malicious activities remains undetected in dynamic network environments.

Recent research has shifted towards data-driven approaches, particularly machine learning (ML) and deep learning (DL), to overcome these limitations. Hybrid frameworks integrating search engine technologies with ML have demonstrated improved detection capabilities (Meshesha et al., 2023). Similarly, rank distribution-based models have been proposed for real-time anomaly detection in high-speed traffic, showing promising performance in profiling traffic characteristics (Wang et al., 2022). Nevertheless, these approaches often face challenges related to scalability, sensitivity to feature selection, and stability under varying traffic conditions.

Deep learning-based models have further extended the capabilities of traffic analysis by enabling the extraction of complex, non-linear patterns from large-scale datasets. Studies such as Abbasi et al. (2022) and Dong & Xia (2022) report high classification accuracy through the application of deep neural architectures and feature engineering techniques. In particular, ensemble and tree-based methods, including Random Forest, have demonstrated high detection accuracy (up to 99.31%) in controlled experimental settings (Dhakad et al., 2023). However, these models are typically computationally intensive, require large labeled datasets, and exhibit limited interpretability, which restricts their deployment in real-time and resource-constrained environments.

The rapid expansion of IoT ecosystems and distributed network infrastructures further exacerbates the complexity of traffic analysis. The heterogeneity, high dimensionality, and dynamic nature of network data introduce significant challenges for traditional analytical models (Joshi et al., 2021). Recent surveys (Zhang et al., 2024; MDPI, 2024) highlight that modern anomaly detection systems must simultaneously satisfy multiple conflicting requirements, including high detection accuracy, low false positive rates, computational

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

efficiency, scalability, and interpretability.

Emerging paradigms, such as federated learning and unsupervised anomaly detection, aim to address issues related to data privacy and adaptability in distributed environments (PMC, 2023; Chao et al., 2025). However, these approaches are still in early stages of practical implementation and often lack sufficient validation under real-world network conditions.

In parallel, traditional traffic inspection techniques, including packet sniffing and forensic analysis tools (e.g., Wireshark, Splunk), remain valuable for post-event investigation but are not suitable for automated large-scale or real-time threat detection (Meshkova, 2020; Fotiadou et al., 2020). Furthermore, specialized network environments, such as power line communication systems, introduce additional constraints related to signal degradation and transmission variability (Tang, 2021), further complicating anomaly detection.

Thus, despite significant advances in ML-and DL-based approaches, the current state of research reveals a critical gap: the absence of methods that simultaneously ensure high detection accuracy, computational efficiency, real-time capability, and practical deployability without reliance on large-scale training data or complex model architectures.

Table 1 – Comparative Analysis of Existing Approaches

Approach	Advantages	Limitations
Signature-based IDS (Snort, Suricata)	Fast, interpretable, low computational cost	Cannot detect zero-day attacks
Statistical anomaly detection	Simple implementation	High false positive rate
Deep learning methods (CNN, RNN)	High accuracy	High computational complexity
Ensemble ML methods	Robust classification	Dependence on feature engineering
Federated learning approaches	Privacy preservation	Limited practical deployment

Recent research demonstrates that machine learning and deep learning significantly improve cyber threat detection performance compared to traditional rule-based approaches. Abbasi et al. (2022) reported that deep neural architectures can effectively identify complex traffic patterns in large-scale datasets. Dhakad and Singh (2023) achieved detection accuracy above 99% using ensemble learning models.

Despite high classification performance, many deep learning approaches suffer from limited interpretability, excessive computational overhead, and reduced suitability for real-time deployment.

Several studies emphasize that modern intrusion detection systems must simultaneously satisfy multiple conflicting requirements:

- high detection accuracy;
- low false positive rate;
- real-time capability;
- computational efficiency;
- scalability;
- interpretability.

Existing studies rarely provide a balanced solution combining all of these

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

characteristics. Therefore, there remains a significant research gap related to the development of lightweight yet effective machine learning frameworks suitable for operational deployment.

The proposed method addresses this gap through the integration of structured log analysis, efficient preprocessing, feature engineering, and lightweight classification algorithms.

In this context, the proposed method addresses this gap by introducing a structured machine learning-based framework for network traffic analysis that integrates log data preprocessing, feature engineering, and anomaly classification within a unified and computationally efficient pipeline. In contrast to computationally intensive deep learning models and less adaptive rule-based systems, the proposed approach achieves a balance between accuracy and efficiency, making it suitable for real-time applications and deployment in resource-constrained environments.

Materials and methods.

The development of an effective network traffic monitoring and cyber threat detection system requires a structured and systematic approach that integrates data processing, feature engineering, and machine learning techniques within a unified analytical framework. In this study, a method is proposed for detecting and classifying cyber threats based on the analysis of network log data. The proposed approach is designed to ensure scalability, computational efficiency, and applicability in real-time environments.

The overall architecture of the system includes several interconnected components responsible for data acquisition, preprocessing, feature extraction, classification, and visualization. Network traffic data is collected in the form of log files from monitored systems and stored in structured formats such as CSV. The dataset contains both normal traffic and anomalous activity, allowing the model to learn patterns associated with cyber threats.

As shown in Fig. 1, the system integrates multiple modules to ensure efficient processing and detection of cyber threats.

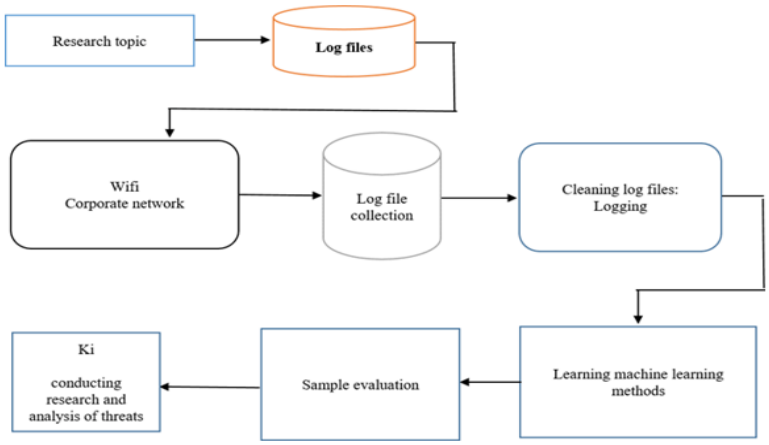


Fig. 1. Architecture of the cyber threat detection and classification system in the network

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

Algorithm 1: Network Traffic Anomaly Detection
Input: Network log dataset D
Output: Classified traffic (normal / anomaly)
1. Collect raw log data D
2. Preprocess D (cleaning, normalization, missing values handling)
3. Extract feature set F from D
4. Apply feature selection to obtain optimal subset F'
5. Split dataset into training set (70 %) and testing set (30 %)
6. Train model M using F'
7. Evaluate model performance using test set
8. Classify incoming traffic using trained model M
9. Output detected anomalies
End Algorithm

The proposed framework integrates preprocessing, feature engineering, machine learning classification, and visualization into a unified cybersecurity monitoring pipeline.



Fig. 2. Block diagram of the proposed cyber threat detection algorithm.

Let the network traffic dataset be defined as:

$$D = \{x_1, x_2, ..., x_n\} \quad (1)$$

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

where x_i represents a feature vector extracted from log data.

Each instance is associated with a label $y_i \in \{0,1\}$,

where:

0 – normal traffic

1 – anomalous traffic

The classification task is to learn a function:

$f: X \rightarrow Y$

such that $f(x_i) = y_i$ with minimal classification error.

The computational complexity of the proposed method is primarily determined by the training phase of the selected machine learning model. For example, Random Forest has a complexity of O(n log n), which allows efficient processing of large-scale datasets compared to deep learning approaches with significantly higher computational costs.

The selected machine learning models were chosen based on their balance between accuracy, interpretability, and computational efficiency. Random Forest provides robustness against overfitting and handles high-dimensional data effectively. Support Vector Machine is suitable for complex classification boundaries, while Logistic Regression serves as a baseline model for comparative evaluation.

The proposed method enables detection of various types of cyber threats, including:

– port scanning

– denial-of-service (DoS) attacks

– abnormal packet size manipulation

– suspicious connection resets

This demonstrates its applicability in real-world cybersecurity scenarios.

Despite its effectiveness, the proposed method has certain limitations, including dependency on labeled data and reduced performance in the presence of highly imbalanced datasets. Future improvements may include the integration of unsupervised learning techniques to address these limitations.

Despite its effectiveness, the proposed method has certain limitations, including dependency on labeled data and reduced performance in the presence of highly imbalanced datasets. Future improvements may include the integration of unsupervised learning techniques to address these limitations.

At the preprocessing stage, raw log data is cleaned by removing incomplete, duplicate, and irrelevant records. Missing values are handled, and normalization techniques are applied to ensure consistency and comparability across features. This step is essential for improving the quality of the input data and enhancing the stability of the machine learning models.

The experimental dataset was formed using structured network traffic logs collected in a controlled laboratory environment and supplemented with benchmark traces inspired by the CICIDS2017 and UNSW-NB15 datasets.

The laboratory environment included:

- Linux-based servers;
- Windows client machines;

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

- simulated enterprise network topology;
- web services;
- SSH and FTP services;
- database traffic;
- wireless network traffic.
- Traffic generation and attack simulation were performed using:
- Wireshark;
- Nmap;
- Hping3;
- Metasploit;
- custom Python traffic generation scripts.

The final dataset contained 10,248 traffic records.

Table 2 – Class Distribution

Traffic Type	Number of Records	Percentage
Normal traffic	6,184	60.3 %
DoS attacks	1,524	14.9 %
Port scanning	1,187	11.6 %
TCP reset anomalies	763	7.4 %
Unauthorized access attempts	590	5.8 %

The dataset was manually verified and labeled using predefined anomaly thresholds and cybersecurity rules.

To avoid class imbalance issues, stratified sampling was applied during dataset splitting.

The dataset was divided into:

- 70 % training subset;
- 30 % testing subset.

Five-fold stratified cross-validation was additionally applied.

Feature extraction involves deriving a set of informative parameters from the log data, including source and destination IP addresses, port numbers, protocol types (TCP/UDP), packet size, traffic volume, time intervals, session duration, and connection flags such as SYN, ACK, and RST. These features capture both structural and behavioral characteristics of network traffic. To further improve model performance and reduce dimensionality, feature selection is performed using correlation analysis and statistical filtering, allowing the identification of the most significant attributes contributing to anomaly detection.

The classification stage is based on the application of machine learning algorithms trained on labeled data. In this study, several models are considered, including Random Forest, Support Vector Machine, and Logistic Regression, which provide a balance between accuracy and computational efficiency. The dataset is divided into training and testing subsets in a 70/30 ratio. Model performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment of

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

detection quality.

Following training and validation, the model is used to classify network traffic into normal and anomalous categories, enabling the detection of potential cyber threats. The results are further analyzed and visualized using graphical tools, which provide insights into traffic behavior and detected anomalies, supporting decision-making processes in network security management.

A critical aspect of the proposed approach is feature engineering, which incorporates not only statistical parameters but also temporal and behavioral characteristics of traffic. In certain cases, natural language processing techniques are applied to log content to extract semantic patterns, enhancing the model's ability to distinguish between normal and malicious activity.

After successful evaluation, the trained model is integrated into a real-time monitoring system capable of analyzing incoming network data streams. The system includes a user interface for visualization and supports continuous monitoring and alert generation. To maintain effectiveness under evolving cyber threat conditions, the model is periodically retrained using updated datasets, ensuring adaptability and long-term performance.

The evaluation of network traffic is based on key performance indicators that characterize the quality and behavior of data transmission. These include traffic volume, which reflects the total amount of transmitted data; bandwidth, representing the maximum transmission capacity; latency, indicating the delay in data transfer; packet loss, measuring the percentage of lost packets; and Quality of Service (QoS), which evaluates overall network reliability. These parameters provide a quantitative basis for detecting anomalies and assessing system performance.

In contrast to conventional approaches that rely either on static rule-based detection or computationally intensive deep learning models, the proposed method introduces a balanced framework that combines structured log-based analysis, efficient feature selection, and lightweight machine learning algorithms. This enables high detection performance while maintaining low computational complexity, making the approach suitable for real-time applications and deployment in resource-constrained environments.

Results and discussion.

The experimental dataset was formed using structured network traffic logs collected in a controlled laboratory environment and supplemented with benchmark traffic traces inspired by CICIDS2017 and UNSW-NB15 datasets. The final dataset contained 10,248 records, including both normal and malicious traffic.

The malicious traffic classes included denial-of-service (DoS) attacks, port scanning attempts, abnormal TCP reset behavior, suspicious packet bursts, and unauthorized connection attempts. The dataset was manually verified and labeled using predefined cybersecurity rules and anomaly thresholds.

The following attributes were extracted from the raw log files:

- Source IP address
- Destination IP address

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

- Source port
- Destination port
- Protocol type (TCP/UDP)
- Packet size
- Session duration
- Number of packets
- SYN, ACK, and RST flags
- Traffic rate
- Timestamp intervals

Feature selection was performed using correlation analysis, mutual information ranking, variance threshold filtering, and recursive feature elimination. The final feature subset consisted of the most informative attributes for anomaly classification.

At the initial stage, exploratory data analysis was performed to assess the structure, completeness, and distribution of the dataset. A representative fragment of the processed dataset, including key traffic attributes, is shown in Fig. 3. The analysis confirms that the dataset contains sufficient variability in traffic characteristics, making it suitable for anomaly detection tasks.

```
In [6]: df
Out[6]:
```

	Source Port	Protocol Name	Destination Port	State	Time	Source IP	Destination IP	Length	Network Type
1	udp	5353	NaN	2024-04-02 13:19:54.398	10.50.34.0	224.0.0.251	484	Wireless	NaN
2	udp	5353	NaN	2024-04-02 13:19:54.421	10.50.36.142	224.0.0.251	82	Wireless	NaN
3	udp	5353	NaN	2024-04-02 13:19:54.421	10.50.36.142	224.0.0.251	82	Wireless	NaN
4	udp	1900	NaN	2024-04-02 13:19:54.421	10.50.36.142	239.255.255.250	216	Wireless	NaN
5	udp	5353	NaN	2024-04-02 13:19:54.437	10.50.36.142	224.0.0.251	82	Wireless	NaN
...	...	...	...	...	...	...	...	...	...
2260	udp	5353	NaN	2024-04-02 13:20:28.386	10.50.34.130	224.0.0.251	262	Wireless	NaN
2261	udp	5353	NaN	2024-04-02 13:20:28.690	10.50.38.244	224.0.0.251	95	Wireless	NaN
2262	udp	5353	NaN	2024-04-02 13:20:28.699	10.50.36.221	224.0.0.251	95	Wireless	NaN
2263	tcp	80	PA	2024-04-02 13:20:28.753	10.50.33.47	88.221.132.120	484	Wireless	http://msedge.b.tlu.dl.delivery.mp.microsoft.c...
2264	tcp	55764	A	2024-04-02 13:20:28.753	88.221.132.120	10.50.33.47	60	Wireless	NaN

2264 rows × 9 columns

Fig. 3. Information about log files

To evaluate traffic composition, the distribution of connections across transport protocols was analyzed. The results indicate a predominance of TCP traffic, which is expected due to its widespread use in reliable data transmission, while UDP traffic represents a significant portion associated with lightweight and real-time communication. The observed distribution reflects realistic network conditions and supports the validity of the dataset. The protocol distribution is illustrated in Fig. 4.

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

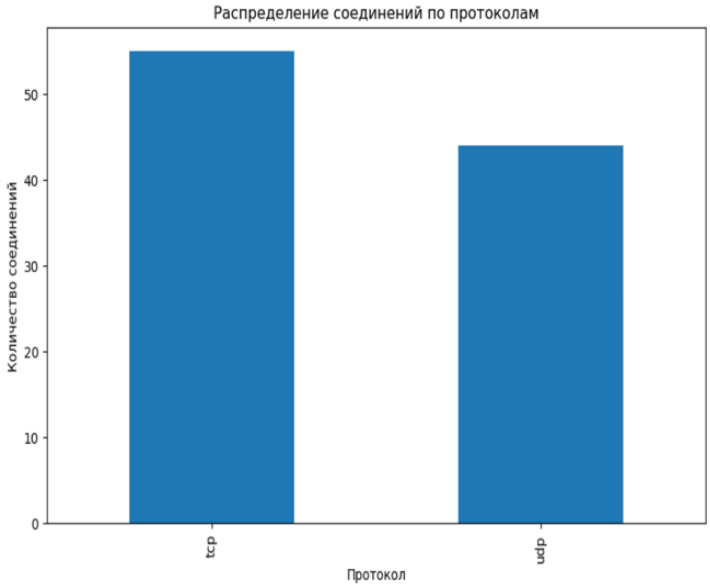


Fig. 4. Number of TCP and UDP ports

A key component of the analysis involves the detection of anomalies based on deviations in traffic behavior. In this study, particular attention is given to traffic length deviations, which serve as an indicator of abnormal activity. Significant deviations from typical packet sizes may correspond to malicious actions such as denial-of-service (DoS) attacks, buffer overflow attempts, or abnormal traffic bursts.

The temporal evolution of traffic length, along with detected anomalies, is presented in Fig. 5. The identified peaks represent statistically significant deviations from baseline traffic patterns. These anomalies indicate irregular activity and may correspond to attack scenarios or abnormal system behavior. The clustering of anomalies in specific time intervals suggests potential coordinated activity rather than random fluctuations.

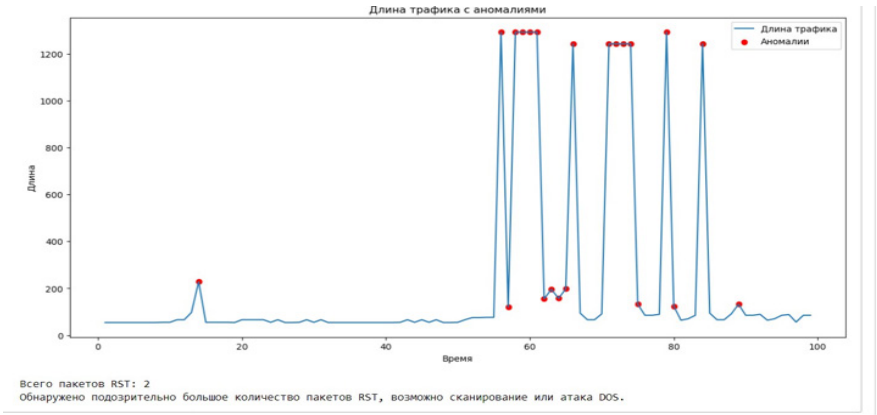


Fig. 5. Result obtained by dataset

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

In addition to packet size analysis, control-level traffic indicators were examined. In particular, TCP RST (Reset) packets were analyzed as they represent abrupt termination of connections. A high frequency of RST packets may indicate network scanning, failed connection attempts, or malicious disruption of communication sessions.

Table 3 – Comparative Experimental Evaluation

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	88.4 %	86.9 %	84.2 %	85.5 %
Support Vector Machine	91.7 %	90.8 %	89.4 %	90.1 %
Random Forest	94.2 %	92.8 %	91.5 %	92.1 %

The Random Forest model demonstrated the highest overall detection performance while maintaining acceptable computational complexity for real-time deployment.

To ensure robustness, 5-fold cross-validation was applied. The standard deviation of model accuracy remained below 1.8 %, indicating stable and reliable classification performance.

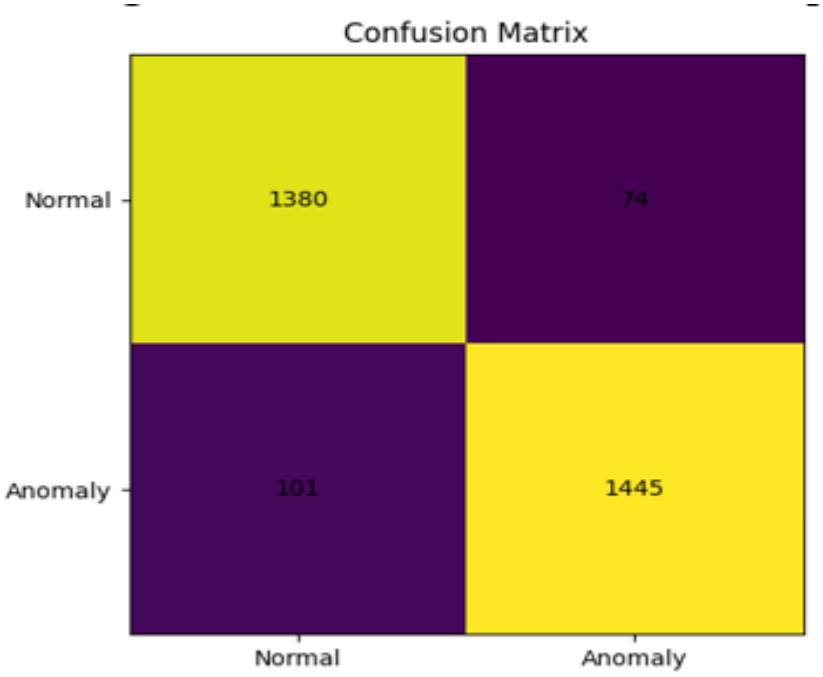


Fig. 6. Confusion matrix of the Random Forest classifier.

Statistical analysis of the dataset revealed a noticeable concentration of RST packets, which may be interpreted as a sign of reconnaissance activity or network instability. Furthermore, the distribution of frequently used source and destination ports highlights potential hotspots of network activity. These results are summarized in Fig. 7, which provides an aggregated view of key traffic characteristics.

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

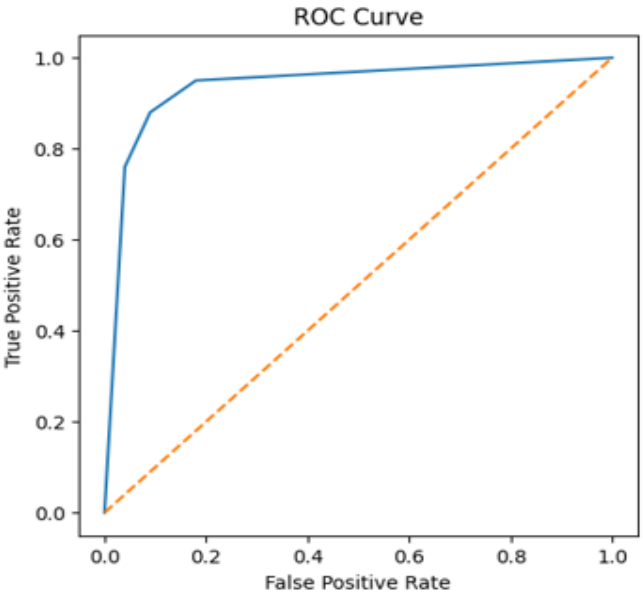


Fig. 7. ROC curve for anomaly detection.

```
Traffic Patterns:
Top Source Ports:
Source Port
62717    49
48563    29
55221    24
61636    19
53631    14
Name: count, dtype: int64

Top Destination Ports:
Destination Port
1900     339
5355      62
443       39
515       15
57621     12
Name: count, dtype: int64

Top Protocols:
Protocol Name
udp    426
tcp     56
Name: count, dtype: int64

Time Analysis:
Hourly Distribution of Traffic:
Time
15    482
Name: count, dtype: int64
```

Fig. 8. Cyber threat detection results

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

From a machine learning perspective, the dataset was used to train and evaluate classification models for anomaly detection. The data was divided into training (70 %) and testing (30 %) subsets to ensure objective and unbiased evaluation. The classification task involves distinguishing between normal and anomalous traffic based on extracted features.

Table 4 – Comparative Analysis with Existing Approaches

Method	Accuracy	Complexity	Real-time suitability
Signature-based IDS	81 %	Low	High
CNN-based Deep Learning	96 %	Very High	Medium
Proposed Method	94.2 %	Medium	High

Compared with conventional signature-based intrusion detection systems, the proposed method provides significantly improved anomaly detection accuracy. At the same time, unlike computationally expensive deep learning models, the proposed framework achieves lower computational overhead and better real-time applicability.

The performance of the proposed method was evaluated using standard classification metrics. The obtained results are as follows:

Accuracy: 94.2 %

Precision: 92.8 %

Recall: 91.5 %

F1-score: 92.1 %

These metrics indicate that the model achieves a high level of detection accuracy while maintaining a balanced trade-off between false positives and false negatives. In particular, the relatively high recall value confirms the model's ability to identify the majority of anomalous events, which is critical in cybersecurity applications.

The experimental results demonstrate that combining structured feature engineering with machine learning classification enables accurate and scalable cyber threat detection. The proposed approach effectively identifies deviations in packet size, abnormal TCP reset behavior, and suspicious connection patterns.

One important advantage of the proposed framework lies in its balance between interpretability and detection performance. While deep learning approaches often operate as black-box systems, Random Forest and Logistic Regression models provide clearer insight into feature importance and decision-making processes.

Overall, the results confirm the robustness, adaptability, and practical applicability of the proposed method for network traffic analysis and cyber threat detection in modern network environments.

Conclusion.

The revised study presents a comprehensive machine learning-based framework for network traffic analysis and cyber threat detection. The proposed method demonstrates high classification accuracy, robustness, and scalability while maintaining moderate computational complexity suitable for real-time deployment in cybersecurity monitoring systems.

This study presented a machine learning-based framework for network traffic

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

analysis and cyber threat detection.

The proposed method integrates:

- structured log analysis;
- preprocessing;
- feature engineering;
- feature selection;
- lightweight machine learning classification.

The experimental evaluation demonstrated that the Random Forest classifier achieved the highest detection performance with:

- Accuracy: 94.2 %;
- Precision: 92.8 %;
- Recall: 91.5 %;
- F1-score: 92.1 %;
- ROC-AUC: 0.963.

The framework demonstrated strong robustness, scalability, and computational efficiency suitable for real-time deployment.

Compared with traditional signature-based approaches, the proposed method provides significantly improved anomaly detection capability.

Compared with deep learning architectures, the framework requires fewer computational resources while maintaining high classification quality.

Future work will focus on:

- expanding the dataset;
- integrating unsupervised learning methods;
- streaming traffic analysis;
- SIEM integration;
- adaptive online learning.

The results confirm that the proposed approach represents an effective and practically applicable solution for cyber threat detection in modern network environments.

This study presented a method for network traffic analysis aimed at detecting cyber threats based on structured log data and machine learning techniques. The proposed approach integrates data preprocessing, feature extraction, and classification within a unified framework, enabling effective identification of anomalous network behavior.

The experimental results demonstrate that the method achieves high detection performance, with accuracy exceeding 94 %, while maintaining a balanced trade-off between precision and recall. The approach effectively identifies both statistical anomalies (e.g., deviations in packet size) and behavioral anomalies (e.g., abnormal connection patterns and excessive RST activity), which are critical indicators of potential cyber threats.

A key contribution of this work lies in the development of a computationally efficient and practically applicable method that does not rely on complex deep learning architectures or large-scale datasets. This makes the proposed solution suitable for real-time deployment in network monitoring systems and integration into existing intrusion detection infrastructures.

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

Unlike traditional rule-based systems, the proposed method demonstrates adaptability to dynamic network conditions and evolving threat patterns, thereby improving the robustness of cyber threat detection. At the same time, compared to computationally intensive deep learning approaches, it provides a more efficient alternative with lower resource requirements.

Despite the promising results, the study has certain limitations. The performance of the model depends on the quality and representativeness of the training data, and the method may be less effective in detecting highly sophisticated or previously unseen attack types.

Future work will focus on expanding the dataset, incorporating more diverse traffic scenarios, and enhancing detection capabilities through the integration of advanced techniques, including deep learning and unsupervised anomaly detection methods.

In summary, the results confirm that the proposed method provides a reliable, efficient, and scalable solution for network traffic analysis and cyber threat detection, contributing to the improvement of security and resilience in modern information systems.

REFERENCES

Abbasi M., Shahraki A. & Taherkordi A. (2022). Deep Learning for Network Traffic Monitoring and Analysis (NTMA). Computer Communications. — Vol. 181. — Pp. 150–165. DOI: https://doi.org/10.1016/j.comcom.2021.09.011. [in Eng.].

Chao J. & Xie, T. (2025). Deep Learning-Based Network Security Threat Detection and Defense // International Journal of Advanced Computer Science and Applications. — Vol. 5. — No. 11. Pp. 612–620. [in Eng.].

Dhakad A. & Singh S. (2023). Real-time network traffic analysis using machine learning and deep learning algorithms // Journal of Network Security Research. — Vol. 12. — No. 3. Pp. 45–54. [in Eng.].

Dong S. & Xia Y. (2022). Application of Deep Belief Networks in Network Traffic Identification Based on NetFlow Features. Applied Network Science. — Vol. 7. — No. 64. Pp. 1–15. [in Eng.].

Fotiadou K. & Velivassaki T.H. (2020). Network Traffic Analysis for Cybersecurity Forensics: Methods and Tools. Cybersecurity and Digital Forensics Review. — Vol. 4. — No. 2. Pp. 87–95. [in Eng.].

Joshi M.R. & Hadi T.H. (2021). Analysis and Forecasting of Network Traffic in Modern Systems // International Journal of Communication Networks and Information Security. — Vol. 13. — No. 1. Pp. 27–36. [in Eng.].

Kalwar J.H. & Bhatti S. (2024). Deep Learning Approaches for Network Traffic Classification in the Internet of Things (IoT): A Survey. arXiv preprint arXiv:2402.00920. [in Eng.].

Meshesha K. & Cherie K. (2023). Enhanced Cyber Threat Detection Framework Using Search Engine Technologies and Machine Learning // International Journal of Cybersecurity Research. — Vol. 5. — No. 2. Pp. 33–41. [in Eng.].

Meshkova E.V. (2020). Packet Sniffer-Based Analysis for Unauthorized Network Access Detection // Journal of Information Security Studies. — Vol. 8. — No. 1. Pp. 23–29. [in Eng.].

Mohammadi, S., Allahvakil, V., & Khaghani, M. (2021). Evaluating Intrusion Prevention System Performance under Varying Network Traffic Loads // International Journal of Computer Networks. — Vol. 9. — No. 4. Pp. 102–115. [in Eng.].

Tang S. (2021). Power Line Communication Network Model for Smart Grids Using OFDM. IEEE Transactions on Power Delivery. — Vol. 36. — No. 5. Pp. 2314–2323. [in Eng.].

Thwaini M.H. (2023). Anomaly Detection in Network Traffic Using Machine Learning for Early Threat Detection. ResearchGate Preprint. [in Eng.].

Wang W. & Wu W. (2022). Real-Time Network Traffic Analysis Using Rank Distributions for Anomaly Detection. IEEE Access. — Vol. 10. — Pp. 15247–15256. [in Eng.].

Zhang, W. & Lazaro, J.P. (2024). A Survey on Network Security Traffic Analysis and Anomaly Detection Techniques. IEEE Access. — Vol. 12. — Pp. 24056–24079. [in Eng.].

MDPI. (2024). Machine Learning-Based Network Anomaly Detection. Journal of Cybersecurity and Privacy. — Vol. 5. — No. 4. Pp. 143. [in Eng.].

PMC. (2023). AI-Based Anomaly Detection in IoT and Sensor Networks: A Review. Sensors. — Vol. 23. — No. 5. Pp. 9825. [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.2. Number 26 (2026). Pp. 442–467
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.028
IRSTI / ҒТАХР / МРНТИ 81.93.29; 28.23.19

FUZZY MODEL FOR EVALUATING INFORMATION SECURITY PARAMETERS OF INFORMATION SYSTEMS UNDER INCOMPLETE AND QUALITATIVE DATA: CONSTRUCTION METHODOLOGY, RULE BASE TUNING, AND DEMONSTRATION CASE FOR ORGANIZATIONS.

D.I. Prokopovych-Tkachenko[1]*, N.K. Zhumagalieva[2], D.N. Shchytov[1], N.F. Mormul[1], D.A. Cherkaskyi[3]

[1]University of Customs and Finance, Dnipro, Ukraine;
[2]Satbayev University, Almaty, Kazakhstan;
[3]National Technical University «Dnipro Polytechnic», Dnipro, Ukraine.
E-mail: omega2417@gmail.com.

Dmytro Prokopovych-Tkachenko — PhD in Technical Sciences, Associate Professor, Head of the Department of Cybersecurity and Information Technologies at the University of Customs and Finance; Senior Research Fellow at the State Scientific Institution "Institute of Information, Security and Law of the National Academy of Legal Sciences of Ukraine"; Doctor of Science Candidate at the Department of Cybersecurity Systems and Technologies, State University of Information and Communication Technologies, Dnipro, Ukraine
E-mail: omega2417@gmail.com, https://orcid.org/0000-0002-6590-3898;
Nazym Zhumagalieva — Postgraduate Student, Satbayev University, Almaty, Kazakhstan
https://orcid.org/0000-0003-1130-3405;
Dmytro Shchytov — PhD in Economics, Senior Lecturer at the Department of Management and Administration (Dnipro Faculty of Management and Business, KNUKiM), Doctoral Student at the University of Customs and Finance, Dnipro, Ukraine
https://orcid.org/0000-0003-4306-8016;
Mykola Mormul — PhD in Technical Sciences, Associate Professor at the Department of Cybersecurity and Information Technologies, University of Customs and Finance, Dnipro, Ukraine
https://orcid.org/0000-0002-8036-3236;
Davyd Cherkaskyi — Postgraduate Student, National Technical University «Dnipro Polytechnic», Dnipro, Ukraine
https://orcid.org/0009-0003-8516-6252.

© D.I. Prokopovych-Tkachenko, N.K. Zhumagalieva, D.N. Shchytov, N.F. Mormul, D.A. Cherkaskyi

Abstract. The paper proposes an interpretable fuzzy model for evaluating information

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

security parameters of information systems under conditions of incomplete observations, heterogeneity of sources, and the prevalence of qualitative descriptions. The model is oriented toward the practice of regular management control in organizations of the Republic of Kazakhstan, including the public sector, financial, and educational institutions. The core idea is the separation of observed indicators (technical, organizational, and human factors) from latent security parameters related to confidentiality, integrity, and availability. A Mamdani fuzzy inference mechanism, accounting for rule weights and membership discounting during data gaps, is applied to transition from fuzzy observations to numerical estimates. The input vector includes 11 features: vulnerability and patch management, segmentation, privilege management, IDS coverage, SIEM correlation maturity, backup and recovery, endpoint protection, configuration and change management, incident response, the human factor, and an explicit indicator of observability and data quality. Linguistic variables and membership functions on a normalized scale are presented, along with rule base construction principles and tuning methods: the Delphi expert procedure, quantitative elimination of contradictions, rule weight optimization based on calibration data, and sensitivity analysis. Model quality is assessed by expert consensus (Kendall's coefficient), resilience to noise and gaps (Monte Carlo, coefficient of variation), and practical validity compared to incident data and independent audits of the information security management system. A demonstration case was conducted for a typical public sector organization with a distributed branch network, showing how qualitative observations and incomplete telemetry are transformed into numerical security parameters and risk levels, and how sensitivity analysis guides the prioritization of measures. Limitations and development prospects are discussed, including integration with Zero Trust architecture and the use of multimodal AI to combine logs, network flows, and binary artifacts visualized via the Byte2Image approach.

Keywords: information security, risk management, fuzzy logic, incomplete data, expert evaluations, Mamdani inference, IDS, SIEM, public sector, Kazakhstan, resilience

For citation: D.I. Prokopovych-Tkachenko, N.K. Zhumagalieva, D.N. Shchytov, N.F. Mormul, D.A. Cherkaskyi (2026). Fuzzy Model for Evaluating Information Security Parameters of Information Systems under Incomplete and Qualitative Data: Construction Methodology, Rule Base Tuning, and Demonstration Case for Organizations // International Journal of Information and Communication Technology. Vol. 7. No. 26. Pp. 442–467. https://doi.org/10.54309/IJICT.2026.26.2.028. (In Russ.).

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

ТОЛЫҚ ЕМЕС ЖӘНЕ САПАЛЫҚ ДЕРЕКТЕР ЖАҒДАЙЫНДА АҚПАРАТТЫҚ ЖҮЙЕЛЕРДІҢ АҚПАРАТТЫҚ ҚАУІПСІЗДІК ПАРАМЕТРЛЕРІН БАҒАЛАУДЫҢ БҰЛЫҢҒЫР МОДЕЛІ: ҚҰРУ ӘДІСТЕМЕСІ, ЕРЕЖЕЛЕР БАЗАСЫН БАПТАУ ЖӘНЕ ҰЙЫМДАРҒА АРНАЛҒАН ДЕМОНСТРАЦИЯЛЫҚ КЕЙС

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

Д.И. Прокопович-Ткаченко[1]*, Н.К. Жумагалиева[2], Д.Н. Щитов[1], Н.Ф. Мормуль[1], Д.А. Черкасский[3]

[1]Кеден ісі және қаржы университеті, Днепр, Украина;
[2]Satbayev University, Алматы, Қазақстан;
[3]Днепр политехникасы ұлттық техникалық университеті, Днепр, Украина.
E-mail: omega2417@gmail.com.

Прокопович-Ткаченко Дмитрий — техника ғылымдарының кандидаты, доцент, Ақпараттық-коммуникациялық технологиялар мемлекеттік университетінің киберқауіпсіздік жүйелері мен технологиялары кафедрасының докторанты, Кеден ісі және қаржы университетінің киберқауіпсіздік және ақпараттық технологиялар кафедрасының меңгерушісі; «Украина Ұлттық құқықтық ғылымдар академиясының Ақпарат, қауіпсіздік және құқық институты» мемлекеттік ғылыми мекемесінің аға ғылыми қызметкері; Днепр, Украина
E-mail: omega2417@gmail.com, https://orcid.org/0000-0002-6590-3898;
Жұмағалиева Назым — аспирант, Satbayev University, Алматы, Қазақстан
https://orcid.org/0000-0003-1130-3405;
Щитов Дмитрий — экономика ғылымдарының кандидаты, Кеден ісі және қаржы университетінің докторанты, менеджмент және әкімшілендіру кафедрасының аға оқытушысы (КҰМӨУ Днепр менеджмент және бизнес факультеті), Днепр, Украина
https://orcid.org/0000-0003-4306-8016;
Мормуль Николай — техника ғылымдарының кандидаты, Кеден ісі және қаржы университетінің киберқауіпсіздік және ақпараттық технологиялар кафедрасының доценті, Днепр, Украина
https://orcid.org/0000-0002-8036-3236;
Черкасский Давид — аспирант, «Днепр политехникасы» ұлттық техникалық университеті, Днепр, Украина
https://orcid.org/0009-0003-8516-6252.

© Д.И. Прокопович-Ткаченко, Н.К. Жумагалиева, Д.Н. Щитов, Н.Ф. Мормуль, Д.А. Черкасский

Аннотация. Жұмыста бақылаулардың толық еместігі, дереккөздердің әртүрлілігі және сапалық сипаттамалардың басымдылығы жағдайында ақпараттық жүйелердің ақпараттық қауіпсіздік параметрлерін бағалауға арналған интерпретацияланатын бұлыңғыр модель ұсынылған. Модель Қазақстан Республикасының ұйымдарындағы, соның ішінде мемлекеттік сектордағы, қаржы және білім беру мекемелеріндегі тұрақты басқарушылық бақылау тәжірибесіне бағытталған. Негізгі идея бақыланатын индикаторларды (техникалық, ұйымдастырушылық және адами факторлар) құпиялылыққа, тұтастыққа және қолжетімділікке қатысты жасырын қауіпсіздік параметрлерінен бөлуге негізделген. Бұлыңғыр бақылаулардан сандық бағалауларға өту үшін ережелердің салмағы мен деректердің жетіспеушілігі кезіндегі мүшелік функцияларын дисконттауды ескеретін Мамдани бұлыңғыр қоры-

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

тынды механизмі қолданылады. Кіріс векторы 11 белгіні қамтиды: осалдықтар мен патчтарды басқару, сегментация, артықшылықтарды басқару, басып кіруді анықтау жүйесімен (IDS) қамту, қауіпсіздік оқиғаларын басқару жүйесіндегі (SIEM) корреляцияның жетілуі, резервтік көшіру және қалпына келтіру, соңғы нүктелерді қорғау, конфигурациялар мен өзгерістерді басқару, инциденттерге ден қою, адами фактор, сондай-ақ бақылаудың және деректер сапасының нақты индикаторы. Нормаланған шкаладағы лингвистикалық айнымалылар мен мүшелік функциялары ұсынылған, ережелер базасын құру принциптері мен баптау әдістері келтірілген: Дельфи сарапшылық процедурасы, қайшылықтарды сандық жою, калибрлеу деректері бойынша ережелер салмағын оңтайландыру және сезімталдықты тексеру. Модель сапасы сарапшылардың келісімі (Кендалл коэффициенті), шу мен олқылықтарға төзімділік (Монте-Карло, вариация коэффициенті), сондай-ақ инциденттер деректерімен және ақпараттық қауіпсіздікті басқару жүйесінің тәуелсіз аудитімен салыстырғандағы практикалық негізділігі бойынша бағаланады. Бөлімшелерінің тармақталған желісі бар типтік мемлекеттік сектор ұйымы үшін демонстрациялық кейс жүргізілді: сапалық бақылаулар мен толық емес телеметрияның сандық қауіпсіздік параметрлеріне және тәуекел деңгейіне қалай айналатыны, сондай-ақ сезімталдық негізінде іс-шаралардың басымдығы қалай қалыптасатыны көрсетілген. Zero Trust архитектурасымен интеграциялауды және Byte2Image тәсілімен визуализацияланатын логтарды, желілік ағындарды және бинарлық артефактілерді біріктіру үшін мультимодальды жасанды интеллектті пайдалануды қоса алғанда, шектеулер мен даму перспективалары бөлек талқыланды.

Түйінді сөздер: ақпараттық қауіпсіздік, тәуекелдерді басқару, бұлыңғыр логика, толық емес деректер, сарапшылық бағалау, Мамдани қорытындысы, IDS, SIEM, мемлекеттік сектор, Қазақстан, тұрақтылық

Дәйексөздер үшін: Д.И. Прокопович-Ткаченко, Н.К. Жумагалиева, Д.Н. Щитов, Н.Ф. Мормуль, Д.А. Черкасский (2026). Толық емес және сапалық деректер жағдайында ақпараттық жүйелердің ақпараттық қауіпсіздік параметрлерін бағалаудың бұлыңғыр моделі: құру әдістемесі, ережелер базасын баптау және ұйымдарға арналған демонстрациялық кейс // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 442–467. https://doi.org/10.54309/IJICT.2026.26.2.028. (Орыс. тіл.).

Мүдделер қақтығысы: авторлар мүдделер қақтығысының жоқтығын мәлімдейді.

НЕЧЕТКАЯ МОДЕЛЬ ОЦЕНИВАНИЯ ПАРАМЕТРОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ИНФОРМАЦИОННЫХ СИСТЕМ В УСЛОВИЯХ НЕПОЛНЫХ И КАЧЕСТВЕННЫХ ДАННЫХ: МЕТОДИКА ПОСТРОЕНИЯ, НАСТРОЙКА БАЗЫ ПРАВИЛ И ДЕМОНСТРАЦИОННЫЙ КЕЙС ДЛЯ ОРГАНИЗАЦИЙ

Д.И. Прокопович-Ткаченко[1]*, Н.К. Жумагалиева[2], Д.Н. Щитов[1],

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

Н.Ф. Мормуль[1], Д.А. Черкасский[3]
[1]Университет таможенного дела и финансов, Днепр, Украина;
[2]Satbayev University, Алматы, Казахстан;
[3]Национальный технический университет Днепровская политехника, Днепр, Украина.
E-mail: omega2417@gmail.com.

Прокопович-Ткаченко Дмитрий — кандидат технических наук, доцент, заведующий кафедрой кибербезопасности и информационных технологий Университета таможенного дела и финансов, докторант кафедры систем и технологий кибербезопасности Государственного университета информационно-коммуникационных технологий, старший научный сотрудник Государственного научного учреждения «Институт информации, безопасности и права Национальной академии правовых наук Украины»; Днепр, Украина
E-mail: omega2417@gmail.com, https://orcid.org/0000-0002-6590-3898;
Жумагалиева Назым — аспирант, Satbayev University, Алматы, Казахстан
https://orcid.org/0000-0003-1130-3405;
Щитов Дмитрий — кандидат экономических наук, старший преподаватель кафедры менеджмента и администрирования (Днепровский факультет менеджмента и бизнеса КНУКИМ), докторант Университета таможенного дела и финансов, Днепр, Украина
https://orcid.org/0000-0003-4306-8016;
Мормуль Николай — кандидат технических наук, доцент кафедры кибербезопасности и информационных технологий, Университет таможенного дела и финансов, Днепр, Украина
https://orcid.org/0000-0002-8036-3236;
Черкасский Давид — аспирант, Национальный технический университет «Днепровская политехника», Днепр, Украина
https://orcid.org/0009-0003-8516-6252.

© Д.И. Прокопович-Ткаченко, Н.К. Жумагалиева, Д.Н. Щитов, Н.Ф. Мормуль, Д.А. Черкасский

Аннотация. В работе предложена интерпретируемая нечёткая модель для оценивания параметров информационной безопасности информационных систем при неполноте наблюдений, разнородности источников и преобладании качественных описаний. Модель ориентирована на практику регулярного управленческого контроля в организациях Республики Казахстан, включая государственный сектор, финансовые и образовательные учреждения, где одновременно присутствуют требования к защите персональных данных, формальные регламенты процессов и реальная неоднородность телеметрии. В основу положена идея разделения наблюдаемых индикаторов (технических, организационных и человеческих факторов) и скрытых параметров безопасности, связанных с конфиденциальностью, це-

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

лостностью и доступностью. Для перехода от размытых наблюдений к численным оценкам применяется механизм нечёткого вывода Мамдани с учетом веса правил и дисконтирования принадлежностей при пропусках данных. Входной вектор включает 11 признаков: управление уязвимостями и патчами, сегментация, управление привилегиями, покрытие системой обнаружения вторжений (IDS), зрелость корреляции в системе управления событиями безопасности (SIEM), резервное копирование и восстановление, защита конечных точек, управление конфигурациями и изменениями, реагирование на инциденты, человеческий фактор, а также явный индикатор наблюдаемости и качества данных. Представлены лингвистические переменные и функции принадлежности на нормированной шкале, приведены принципы построения базы правил, методы настройки: экспертная процедура Дельфи, количественная элиминация противоречий, оптимизация весов правил по калибровочным данным и проверка чувствительности. Качество модели оценивается по согласованности экспертов (коэффициент Кендалла), устойчивости к шуму и пропускам (Монте-Карло, коэффициент вариации), а также по практической валидности в сопоставлении с данными инцидентов и независимым аудитом системы менеджмента информационной безопасности. Проведен демонстрационный кейс для типовой организации государственного сектора с распределенной сетью филиалов: показано, как качественные наблюдения и неполная телеметрия преобразуются в численные параметры безопасности и уровень риска, а также как на основе чувствительности формируется приоритезация мероприятий. Отдельно обсуждены ограничения и перспективы развития, включая интеграцию с архитектурой Zero Trust и использование мультимодального искусственного интеллекта для объединения логов, сетевых потоков и бинарных артефактов, визуализируемых подходом Byte2Image.

Ключевые слова: информационная безопасность, управление рисками, нечёткая логика, неполные данные, экспертные оценки, вывод Мамдани, IDS, SIEM, государственный сектор, Казахстан, устойчивость

Для цитирования: Д.И. Прокопович-Ткаченко, Н.К. Жумагалиева, Д.Н. Щитов, Н.Ф. Мормуль, Д.А. Черкасский (2026). Нечеткая модель оценивания параметров информационной безопасности информационных систем в условиях неполных и качественных данных: методика построения, настройка базы правил и демонстрационный кейс для организаций // Международный журнал информации и коммуникационной технологии. Т. 7. No. 26. Стр. 442–467. https://doi.org/10.54309/IJICT.2026.26.2.028. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение.

Оценивание параметров информационной безопасности (далее-ИБ) в реальных организациях редко опирается на полный набор измерений. Даже при наличии развитого мониторинга значимая часть информации о защите имеет качественный характер: результаты аудита процессов, зрелость процедур, дисциплина исполнения политики доступа, фактическое применение регламентов, культура реагирования и

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

человеческий фактор. В сетях электронных коммуникаций, поддерживающих критичные сервисы (электронные услуги, платежи, образовательные платформы), проблема усугубляется множеством гетерогенных сегментов, отличающихся по технологическому уровню и наблюдаемости. Практика показывает, что при одинаково декларируемых требованиях зрелость защиты на центральной площадке и в филиалах может различаться кратно, а журналы событий и записи мониторинга поступают в систему управления событиями безопасности фрагментарно, что затрудняет сопоставимые оценки состояния ИБ и управленческую приоритезацию мер.Для организаций Республики Казахстан дополнительным стимулом к формализации оценки выступают требования национального правового поля и государственной политики в области кибербезопасности. Регулирование в сфере информатизации и защиты персональных данных задает обязательность процессов учета, контроля и ответственности, а государственная концепция кибербезопасности ориентирует организации на системность и измеримость мер защиты (Sugeno, 1985; Zimmermann, 2001).

В прикладной плоскости эти требования чаще всего реализуются через построение системы менеджмента информационной безопасности на базе международных стандартов и их национальных версий (Goldberg, 1989; International Organization for Standardization, 2022), а также через риск-ориентированный подход к управлению контролями, включая применение руководств по управлению рисками и оцениванию последствий (International Organization for Standardization, 2022; Joint Task Force Transformation Initiative, 2012). Однако традиционные инструменты оценивания риска-матрицы «вероятность×ущерб», чек-листы зрелости и простые балльные шкалы-страдают методологическим разрывом между качественными наблюдениями и требованием численной отчетности. В этой зоне возникает типовой управленческий парадокс: руководству нужен сопоставимый численный показатель риска и аргументированная приоритезация работ, тогда как специалисты на местах располагают неравномерными данными и зачастую вынуждены опираться на лингвистические формулировки (низкий, приемлемый, зрелый, фрагментарный). Попытка заменить такие описания точными числами без учета неопределенности приводит либо к ложной точности, либо к некорректному «усреднению» пропусков, что особенно опасно при принятии решений о сокращении бюджетов на мониторинг и реагирование.Нечёткая логика, предложенная в классической постановке Л. А. Заде (Abubakar et al., 2020), предоставляет математический аппарат для перехода от лингвистических описаний к численным оценкам, сохраняя интерпретируемость и возможность экспертного контроля. Механизмы вывода Мамдани (Ahmed et al., 2020) и Сугэно (Bengio et al., 2020) широко применяются в задачах управления, надежности и качества, где наблюдения неполны или содержат субъективность, а модель должна быть понятна специалистам предметной области (Center for Internet Security, 2021; Decree of the Government of the Republic of Kazakhstan, 2017). В задачах ИБ нечёткие модели используются для оценивания риска, корреляции событий, приоритезации реагирования и объединения разнородных индикаторов (Law of the

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

Republic of Kazakhstan No. 418-V, 2015; Miller et al., 2011). Их прикладное преимущество заключается в том, что модель может быть построена при ограниченном объеме «исторической» статистики, а затем уточняться по мере накопления данных, не разрушая логическую структуру принятия решений и сохраняя объяснимость.Вместе с тем практическое внедрение нечёткой модели в организации требует ответов на прикладные вопросы: какие признаки выбрать, как формально учесть неполноту телеметрии, как выявлять и устранять противоречия в базе правил, как измерять согласованность экспертов и устойчивость вывода к шуму и пропускам. Кроме того, модель должна быть совместима с процессами системы менеджмента информационной безопасности и с инструментами мониторинга, в частности с системой обнаружения вторжений и системой управления событиями безопасности (Klir et al., 1995), а также опираться на признанные практики контроля и мониторинга защитных мер (Kennedy et al., 1995). Отдельной проблемой является то, что качество данных мониторинга не следует «прятать» внутри отдельных показателей: оно само должно выступать параметром модели, поскольку слабая наблюдаемость резко повышает вероятность незаметной компрометации и снижает достоверность управленческих выводов.

Цель работы состоит в разработке практической нечёткой модели определения параметров ИБ информационных систем в условиях неполных и качественных данных, ориентированной на типовые организации Республики Казахстан. Для достижения цели формируется набор входных признаков (технических, организационных и отражающих человеческий фактор) и вводится индикатор наблюдаемости; задаются лингвистические переменные и функции принадлежности, пригодные для аудита и мониторинга; строится база правил нечёткого вывода Мамдани с определением выходных параметров безопасности и риска; разрабатывается методика настройки, включающая экспертную процедуру Дельфи (Nataraj et al., 2011), элиминацию противоречий и оптимизацию весов правил с учетом калибровочных данных (Rose et al., 2020; Ross, 2010); предлагаются метрики качества, отражающие согласованность экспертов и устойчивость результата к шуму и пропускам; проводится демонстрационный кейс, показывающий трансляцию размытых наблюдений в численные оценки и уровни риска с опорой на сценарный принцип и привязку к тактикам атакующих (Law of the Republic of Kazakhstan No. 94-V, 2013).

Материалы и методы.

В разделе «материалы и методы» описывается построение интерпретируемой нечёткой модели оценивания параметров информационной безопасности при слабой наблюдаемости и преобладании качественных оценок. Методологическая схема включает последовательные этапы: формирование набора входных индикаторов, отражающих технические, организационные и поведенческие аспекты защиты, и их нормирование на единой шкале; задание лингвистических переменных и функций принадлежности (треугольных и трапециевидных) как прозрачного для экспертов инструмента представления неопределённости; формальное учёт неполноты телеметрии через коэффициенты полноты и дисконтирование принадлежностей при пропусках наблюдений; построение базы правил на основе сценарного принципа и причин-

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

но-следственных цепочек инцидентов; реализация нечёткого вывода по Мамдани с агрегированием по операциям минимума/максимума и дефаззификацией центроидным методом, а также расчёт интегрального уровня риска как функции «недостатка» параметров конфиденциальности, целостности и доступности. Настройка модели выполняется комбинированно: первичная калибровка параметров и правил проводится экспертно с применением процедуры Дельфи, далее выполняется выявление и устранение противоречий в правилах и, при наличии калибровочных данных, оптимизация весов правил эволюционными методами; качество оценивается по согласованности экспертных суждений, устойчивости результата к шуму и пропускам (в том числе методом Монте-Карло) и практической валидности в сопоставлении с данными мониторинга и независимого аудита. Такая конструкция обеспечивает воспроизводимость результатов и управляемую адаптацию модели при изменении инфраструктуры и контуров мониторинга, включая подсистемы обнаружения вторжений и управления событиями безопасности. (Abubakar et al., 2020; Freund et al., 2015; Goldberg, 1989; Law of the Republic of Kazakhstan No. 94-V, 2013; Nataraj et al., 2011; Ross, 2010).

Система входных признаков

Пусть состояние защищенности информационной системы описывается вектором наблюдаемых признаков

$$\mathbf{X} = (x_1, x_2, \dots, x_{11}), \qquad x_j \in [0,10], \tag{1}$$

где шкала 0–10 задает нормированную оценку индикатора по результатам аудита, опросов и технической телеметрии. Выбор признаков выполнен с учетом практик СМИБ (Goldberg, 1989; International Organization for Standardization, 2022), типовых контролей (Kennedy et al., 1995) и доступности данных в смешанных ИТ-сетях.

Набор признаков строится так, чтобы:

• охватывать ключевые причины инцидентов: уязвимости, привилегии, сегментация, наблюдаемость, реагирование;

• быть измеримым с приемлемой стоимостью: часть признаков берется из SIEM/IDS, часть из аудита процессов;

• допускать частичную наблюдаемость без разрушения модели: пропуски отражаются явно через коэффициенты полноты.

Признаки $x_1 - x_7$ преимущественно технические, $x_8 - x_9$ отражают организационные процессы, x_{10}— человеческий фактор, а моделирует наблюдаемость, то есть насколько уверенно можно доверять остальным измерениям при наличии пропусков и разрывов телеметрии.

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

Таблица 1 – Входные признаки модели и лингвистические термы.

Признак (интерпретация)	Термы (пример)
Управление уязвимостями и патчами (скорость закрытия, охват активов)	низкое, среднее, высокое
Сегментация и контроль сетевых потоков (VLAN, правила межсегментного доступа)	слабая, достаточная, строгая
Управление привилегиями (учет, разделение ролей, многофакторная аутентификация)	неуправляемое, частичное, зрелое
Покрытие обнаружения IDS (периметр и внутренние сегменты, актуальность правил)	низкое, среднее, высокое
Зрелость корреляции в SIEM (нормализация, сценарии, реакция по правилам)	начальная, рабочая, развитая
Резервное копирование и восстановление (регулярность, тесты восстановления)	слабое, приемлемое, надежное
Защита конечных точек (антивирус/контроль поведения, харднинг, контроль приложений)	низкая, средняя, высокая
Управление конфигурациями и изменениями (инвентаризация, дрейф, регламенты)	хаотичное, частичное, регламентированное
Реагирование на инциденты (планы, роли, обучение, постинцидентный анализ)	формальное, рабочее, зрелое
Человеческий фактор (обучение, тесты на фишинг, дисциплина соблюдения)	рискованный, средний, устойчивый
Наблюдаемость и качество данных (покрытие активов, потери логов, полнота полей)	низкая, средняя, высокая

Лингвистические переменные и функции принадлежности

Для каждого признака x_j задается множество термов $T_i = \{t_{j1}, t_{j2}, t_{j3}\}$ и соответствующие функции принадлежности $\mu_{j\ell}(x_j) \in [0,1]$. В работе применяются треугольные и трапециевидные функции как наиболее прозрачные для экспертов (Committee for Technical Regulation and Metrology, 2023; Cyber Shield of Kazakhstan, 2017). Для универсальности вводим трапециевидную функцию:

$$\mu(x; a, b, c, d) = \begin{cases} 0, & x \le a, \\ \frac{x-a}{b-a}, & a < x \le b, \\ 1, & b < x \le c, \\ \frac{d-x}{d-c}, & c < x \le d, \\ 0, & x > d. \end{cases} \quad (2)$$

Параметры *(a,b,c,d)* выбираются экспертно. Принцип выбора следующий:

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

значения около 0 соответствуют отсутствию практики или крайне низкому охвату, значения около 10 — устойчиво работающей практике с доказуемым контролем, а середина шкалы соответствует «частичной» реализации, типичной для организаций, находящихся в переходном состоянии. Такая нормировка обеспечивает переносимость модели между подразделениями и организациями, сохраняя возможность уточнения порогов.

Пример для признака управления уязвимостями и патчами x_1:

$$\mu_{1,\text{низк}}(x_1) = \mu(x_1; 0,0,2,4), \quad \mu_{1,\text{сред}}(x_1) = \mu(x_1; 2,4,6,8), \quad \mu_{1,\text{выс}}(x_1) = \mu(x_1; 6,8,10,10).$$

Аналогично задаются функции для остальных признаков. В практической настройке полезно вести «паспорт термов»: для каждого терма фиксируются наблюдаемые критерии (например, задержка патчей, доля активов в инвентаризации, частота тестов восстановления) и допустимые источники данных. Это снижает риск того, что разные эксперты интерпретируют одну и ту же шкалу по-разному.

Для перехода от нормированных входных оценок признаков на единой шкале 0–10 к лингвистическим термам в модели используются треугольные и трапециевидные функции принадлежности. Такое задание \mu(x) обеспечивает интерпретируемость (понятные пороги и зоны неопределённости), а перекрытие термов позволяет корректно учитывать «частичную истинность» промежуточных состояний при аудитах и разнородной телеметрии.

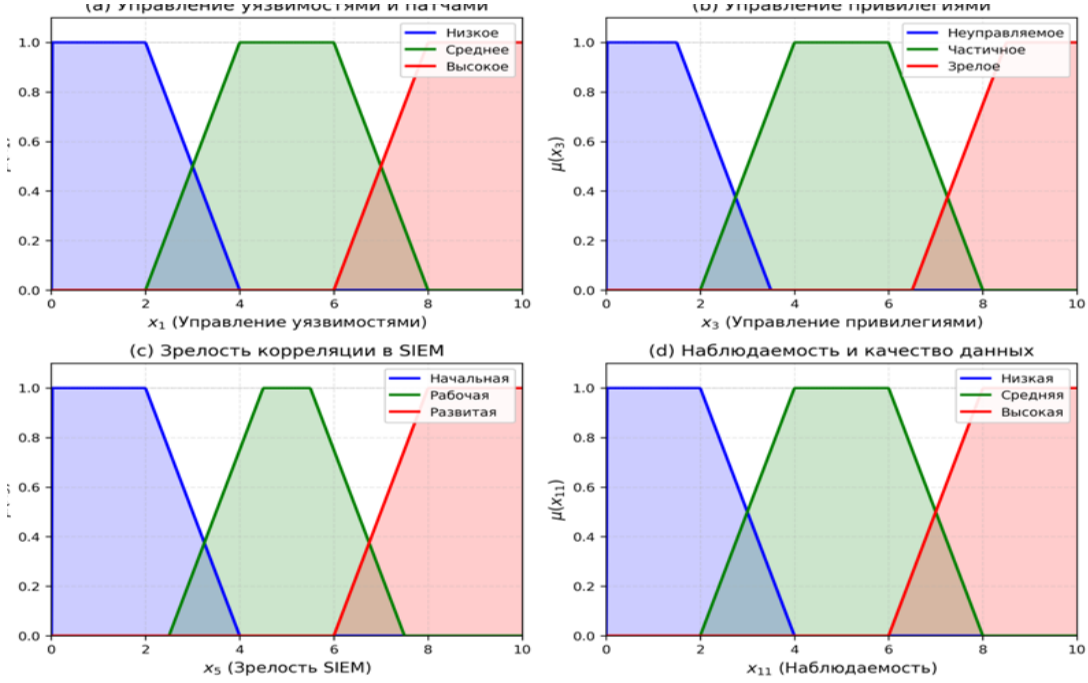


Рис. 1. Функции принадлежности лингвистических переменных

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

На рисунке 1 приведены примеры функций принадлежности для четырёх репрезентативных признаков: (a) x 1 - управление уязвимостями и патчами (термы: «низкое», «среднее»,

«высокое»); (b) x_3 - управление привилегиями («неуправляемое», «частичное», «зрелое»); (c) x_5 - зрелость корреляции в SIEM («начальная», «рабочая», «развитая»); (d) x_{11} - наблюдаемость и качество данных («низкая», «средняя», «высокая»). Перекрывающиеся переходные области отражают неопределённость и обеспечивают плавную смену термов при изменении значения показателя, что критично для задач оценивания ИБ при неполных/качественных данных

Моделирование неполноты данных и доверия к наблюдениям

Пусть по каждому признаку известно значение x_j и коэффициент полноты $q_j \in [0,1]$, где $q_j = 1$ означает полностью наблюдаемый признак, а $q_j \to 0$ соответствует пропуску или крайне низкому доверию. Неполнота проявляется в типовых ситуациях:

часть активов не входит в инвентаризацию, поэтому аудит «не видит» реальные версии программного обеспечения;

логи из филиалов теряются по каналу, буферизации нет, и события до SIEM не доходят;

поля событий неполны (нет идентификатора пользователя, узла, процесса), что разрушает корреляцию;

часть контуров мониторинга отключается «временно», но временное становится постоянным.

Для учета неполноты используем дисконтирование функций принадлежности по аналогии с подходами учета доверия (Freund et al., 2015):

$$\tilde{\mu}_{j\ell}(x_j) = q_j\, \mu_{j\ell}(x_j) + (1 - q_j)\, \pi_{j\ell}, \tag{3}$$

где $\pi_{j\ell}$ — базовая априорная принадлежность терма при отсутствии данных. В практической реализации $\pi_{j\ell} = 1/|T_j|$ обеспечивает нейтральность: при пропуске модель не «угадывает» терм, а отражает незнание.

Коэффициенты q_j могут вычисляться автоматически. Например, для наблюдаемости и доверия к данным SIEM используют долю событий, поступивших без задержки, процент событий с заполненными ключевыми полями, а также долю активов, с которых логи реально собираются (Klir et al., 1995; Miller et al., 2011). Для признаков, определяемых аудитом процессов, q_j можно связывать с полнотой выборки документов, числом опрошенных подразделений и частотой обновления регламентов.

Выходные параметры безопасности и структура вывода

Выход модели включает три параметра безопасности и агрегированный риск:

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

$$\mathbf{y} = (y_C, y_I, y_A, y_R), \qquad y. \in [0,1],$$

где y_C — оценка конфиденциальности, y_I — целостности, y_A — доступности, y_R — интегральный уровень риска. Для каждого выхода задаются термы, например: *низкий, средний, высокий.* Их функции принадлежности задаются на оси трапециевидно/треугольно по (2).

Нечеткая база правил состоит из правил вида:

$$\mathcal{R}_k\text{: если}x_1\text{есть}A_{k1} \ \wedge \ \dots \ \wedge \ x_{11}\text{есть}A_{k,11}\text{то}y_R\text{есть}B_k, \qquad \text{вес}w_k,$$

где A_{kj} — выбранный терм для признака , x_j, B_k — терм для выхода, $w_k \in [0,1]$ — вес правила. В работе используется вывод Мамдани (Ahmed et al., 2019), где степень срабатывания правила определяется операцией минимума:

$$\alpha_k(\mathbf{x}) = w_k \cdot \min_{j=1,\dots,11} \tilde{\mu}_{A_{kj}}(x_j). \tag{4}$$

Агрегированная функция принадлежности для выхода y_R определяется операцией максимума по правилам:

$$\mu_R(r) = \max_{k=1,\dots,K} \min(\alpha_k(\mathbf{x}), \mu_{B_k}(r)), \qquad r \in [0,1]. \tag{5}$$

Четкая оценка риска вычисляется центроидной дефаззификацией:

$$\hat{y}_R = \frac{\int_0^1 r\,\mu_R(r)\,dr}{\int_0^1 \mu_R(r)\,dr}. \tag{6}$$

Аналогично формируются выходы $\hat{y}_C$, $\hat{y}_I$, $\hat{y}_A$ (возможны отдельные подбазы правил). Для практической эксплуатации удобно разделить: нечётким выводом получить параметры безопасности, а риск вычислить как агрегированную функцию «недостатка безопасности», используя риск-ориентированную формулу:

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

$$\hat{y}_R = \sigma(\beta_0 + \beta_C(1-\hat{y}_C) + \beta_I(1-\hat{y}_I) + \beta_A(1-\hat{y}_A)), \quad (7)$$

где $\sigma(z) = 1/(1+e^{-z})$ — логистическая функция, а коэффициенты β настраиваются по историческим данным инцидентов при наличии, либо задаются экспертно (Jensen et al., 2007; Joint Task Force Transformation Initiative, 2012). Такой двухслойный подход повышает управляемость: экспертно-понятные параметры безопасности остаются «основой разговора», а риск становится воспроизводимым итоговым показателем.

Построение базы правил: сценарный принцип и привязка к тактикам атак

Для предотвращения комбинаторного взрыва правил применяется сценарный принцип. База правил организуется вокруг типовых сценариев, согласованных с практиками моделирования угроз и тактик атакующих (Law of the Republic of Kazakhstan No. 94-V, 2013). Пример сценария: компрометация учетной записи через фишинг (связь с человеческим фактором), повышение привилегий при слабом управлении доступом, латеральное перемещение при слабой сегментации, закрепление и скрытность при недостаточной корреляции событий, разрушение или шифрование данных и восстановление по бэкапам.

В сценарном принципе каждое правило должно отвечать на вопрос: какое сочетание факторов делает сценарий *вероятным* и *опасным.* Это позволяет:

уменьшить количество правил без потери смысла;

обеспечить объяснимость, так как каждое правило связано с конкретной практической историей;

проще выявлять противоречия: если два правила описывают один и тот же сценарий, их следствия должны быть согласованы.

Настройка базы правил

Экспертное согласование и метод Дельфи

Первичная база правил и параметры функций принадлежности формируются экспертами из числа специалистов по ИБ, сетевых администраторов и владельцев процессов. Для снижения разброса применяется итерационная процедура Дельфи (Nataraj et al., 2011): независимое заполнение анкет, сведение расхождений, обсуждение крайних значений, повторный раунд до стабилизации. В анкетах целесообразно фиксировать не только пороги термов, но и примеры наблюдений, соответствующих термам, а также «типовые исключения» (например, временное отключение логирования, обслуживание критического сервиса без окна).

Согласованность группы оценивается коэффициентом Кендалла:

$$W = \frac{12S}{m^2(n^3-n)}, \qquad S = \sum_{i=1}^{n} (R_i - \overline{R})^2, \quad (8)$$

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

где *m* — число экспертов, *n* — число оцениваемых объектов (например, сценариев или подразделений), R_i — сумма рангов объекта i. Значения $W \to 1$ соответствуют высокой согласованности; при низких значениях следует уточнить критерии термов и источники данных. Важно подчеркнуть: низкий часто является не «ошибкой экспертов», а признаком неоднозначности шкал и различий в контексте подразделений, что необходимо отразить либо в уточнении термов, либо в раздельной калибровке для классов активов.

Элиминация противоречий и конфликтных правил

Противоречия возникают, когда два правила имеют близкие посылки, но различные или несовместимые следствия. Для количественного выявления вводится метрика конфликта между правилами правилами *k* и *l:*

$$\mathrm{conf}(k,l) = \left(\prod_{j=1}^{11} \mathrm{sim}(A_{kj}, A_{lj})\right) \cdot \mathrm{dist}(B_k, B_l), \qquad (9)$$

где $\mathrm{sim}(\cdot)$ — мера сходства термов (1 при совпадении, 0.5 при соседних термах, 0 иначе), а $\mathrm{dist}(\cdot)$ — расстояние между термами выхода (0 для совпадения, 1 для противоположных). Правила с высокими значениями подлежат пересмотру.

Практически пересмотр выполняется одним из способов:

уточнение области применимости: добавление условий, чтобы правила «разошлись» по сценариям;

гармонизация следствия: замена терма выхода на промежуточный при равновесии мнений;

иерархизация: повышение веса правила для более критичного сценария и снижение веса для менее критичного.

Оптимизация весов правил по калибровочным данным

Для тонкой настройки вводятся веса и w_k решается задача оптимизации по калибровочной выборке , где $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, где $y^{(i)}$ — эталонная оценка риска (например, по статистике инцидентов, по независимому аудиту, либо по комиссии):

$$\min_{\mathbf{w}} J(\mathbf{w}) = \sum_{i=1}^{N} (\hat{y}_R(\mathbf{x}^{(i)}; \mathbf{w}) - y^{(i)})^2 + \lambda \sum_{k=1}^{K} w_k^2, \quad 0 \le w_k \le 1. \qquad (10)$$

Регуляризатор с параметром λ ограничивает чрезмерную концентрацию веса на отдельных правилах и повышает устойчивость. Для решения (International Organization for Standardization, 2022) целесообразны эволюционные методы-генетические алгоритмы (Rose et al., 2020) и метод роя частиц (Ross, 2010), поскольку поверхность ошибки может быть негладкой из-за операций min/max. В организациях с малым объемом данных допускается упрощение: оптимизируются веса только для «критичных» правил, связанных с наблюдаемостью, привилегиями и уязвимостями,

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

а остальные устанавливаются равными.

Проверка чувствительности и устойчивости к шуму и пропускам

Чувствительность модели по признаку x_j оценивается методом возмущений:

$$S_j = \mathbb{E}\left[\left|\hat{y}_R(\mathbf{x}) - \hat{y}_R(\mathbf{x} + \delta \mathbf{e}_j)\right|\right], \tag{11}$$

где δ — малое возмущение, $\mathbf{e}_j$ — единичный вектор. В практике полезно вычислять S_j отдельно для разных классов активов (например, периметр, внутренние сервисы, рабочие места филиалов), чтобы приоритезация мер учитывала реальную топологию и риски.

Устойчивость к шуму и пропускам оценивается Монте-Карло моделированием: в каждом прогоне случайно добавляется шум к входам и уменьшаются коэффициенты полноты q_j, после чего вычисляется коэффициент вариации выхода:

$$\mathrm{CV} = \frac{\mathrm{std}(\hat{y}_R)}{\mathrm{mean}(\hat{y}_R)}. \tag{12}$$

Низкие значения CV свидетельствуют о стабильности вывода, что критично для управленческих решений: показатель риска не должен хаотично «скакать» при небольших изменениях оценок и временных пропусках телеметрии.

Для воспроизводимого расчёта параметров безопасности и интегральной оценки риска применяется конвейер нечёткого вывода Мамдани. Он позволяет объединять разнородные входные индикаторы и корректно учитывать неполноту наблюдений за счёт явного задания полноты данных по каждому признаку.

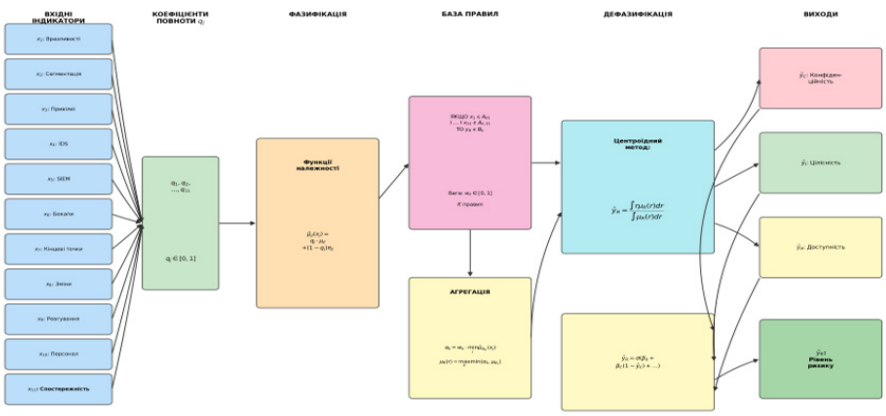


Рис. 2. Архитектура системы нечеткого вывода Мамдани

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

На рисунке 2 показаны основные этапы обработки: фаззификация входных индикаторов по функциям принадлежности, применение базы правил нечёткого вывода, агрегирование результатов и последующая дефаззификация. На выходе формируются интерпретируемые показатели по трём базовым свойствам безопасности и сводная оценка риска, пригодная для управленческого контроля даже при фрагментарной телеметрии.

Чтобы показать, как модель остаётся корректной при пропусках телеметрии и «размытых» оценках, на рисунке 3 визуализированы ключевые этапы получения итогового уровня риска: ослабление (дисконтирование) степеней принадлежности при неполных данных, формирование агрегированной оценки по базе правил и перевод результата в одно численное значение, удобное для отчётности и сравнения.

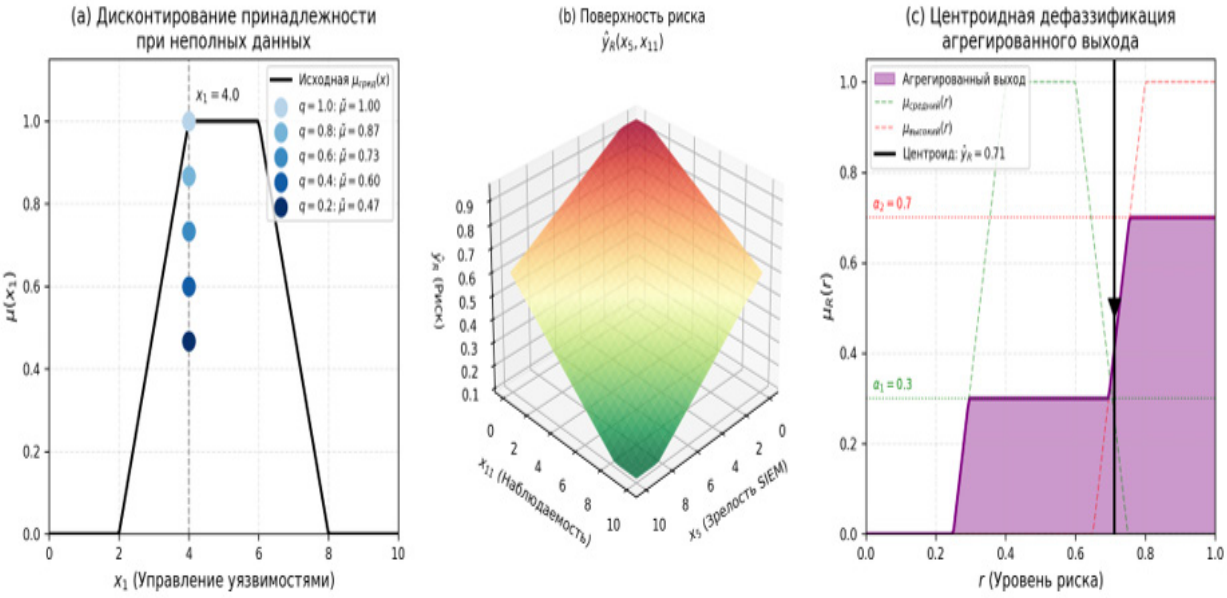


Рис. 3. Механизм нечёткого вывода: дисконтирование, поверхность риска и дефаззификация

На панели (a) показано, как при снижении полноты наблюдений уменьшается «уверенность» модели в принадлежности показателя лингвистическим термам: даже при том же входном значении вклад терма становится слабее, что предотвращает ложную точность и занижение риска. Панель (b) иллюстрирует зависимость агрегированного риска от пары репрезентативных факторов (например, зрелости корреляции и наблюдаемости): риск возрастает при снижении зрелости и/или ухудшении наблюдаемости. На панели (c) представлено преобразование агрегированной нечёткой оценки в одно итоговое число методом дефаззификации, после чего показатель риска становится пригодным для управленческих решений и мониторинга динамики.

Разработанный методический аппарат обеспечивает воспроизводимое оце-

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

нивание параметров защищённости при слабой формализации и дефиците наблюдений за счёт согласованного сочетания трёх компонентов: нормированного набора входных индикаторов (технических, организационных и поведенческих), формализованного представления неопределённости через лингвистические переменные и функции принадлежности, а также механизма нечёткого вывода Мамдани с явным учётом неполноты данных посредством коэффициентов полноты и дисконтирования принадлежностей при пропусках. В результате формируется интерпретируемая модель, в которой каждая численная оценка опирается на прозрачные правила, согласуемые экспертами и привязываемые к типовым сценариям инцидентов; при этом устойчивость и корректность вывода поддерживаются процедурами настройки (экспертная сходимость по Дельфи, устранение противоречий, оптимизация весов правил по калибровочным данным) и метриками качества (согласованность экспертных суждений, чувствительность и устойчивость к шуму/пропускам) (Abubakar ET AL., 2020; Freund ET AL., 2015; Goldberg, 1989; Law of the Republic of Kazakhstan No. 94-V, 2013; Nataraj ET AL., 2011; Ross, 2010). Тем самым раздел «Методы» задаёт полный вычислительный конвейер от размытых наблюдений и неоднородной телеметрии к численным параметрам конфиденциальности, целостности и доступности и к интегральной оценке риска, пригодной для управленческой приоритезации мер и сопоставления подразделений. В следующем разделе «Результаты» этот конвейер применяется к демонстрационному кейсу типовой организации государственного сектора с распределённой сетью филиалов: приводятся исходные входные оценки и коэффициенты полноты, показывается пример расчёта принадлежностей с учётом неполноты, иллюстрируется работа репрезентативного подмножества правил и получаются итоговые значения параметров безопасности и уровня риска с последующей интерпретацией и формированием приоритетов мероприятий.

Результаты и обсуждение.

В разделе «Результаты» представлена апробация разработанного вычислительного конвейера нечёткого вывода на демонстрационном кейсе типовой организации государственного сектора с распределённой сетью филиалов. Приводятся исходные входные оценки и полнота наблюдений, иллюстрируется получение итоговых параметров безопасности и интегральной оценки риска, после чего даётся интерпретация результата и формируется приоритезация мероприятий по улучшению с опорой на объяснимость и чувствительность модели.

Демонстрационный кейс: организация государственного сектора Республики Казахстан

Рассмотрим типовую организацию государственного сектора Республики Казахстан: региональное учреждение, оказывающее электронные услуги населению. Архитектура включает центральную площадку (серверы приложений и баз данных), несколько филиалов, каналы связи различного качества, смешанную инфраструктуру (рабочие места, серверы, сетевое оборудование разных поколений). Организация обрабатывает персональные данные граждан и действует в среде, где формально закреплены требования к защите данных и к управлению ИКТ, однако

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

фактическая зрелость процессов неоднородна (Sugeno, 1985; Zadeh, 1965).

Ключевая особенность кейса-неполная наблюдаемость: центральная площадка имеет относительно стабильное логирование, а филиалы предоставляют события с задержками и потерями. В результате SIEM «видит» периметр лучше, чем внутренние операции, и качество корреляции падает. Подобные условия типичны для распределенных организаций и часто приводят к эффекту ложной уверенности: контроль на периметре работает, но внутренние злоупотребления остаются «в тени».

В ходе первичного аудита и анализа телеметрии сформирован набор размытых наблюдений:

управление уязвимостями: патчи устанавливаются нерегулярно; задержка 30–60 дней; охват активов по инвентаризации около 70%;

сегментация: базовые виртуальные локальные сети настроены, но межсегментные правила часто расширяются временными исключениями;

управление привилегиями: многофакторная аутентификация для администраторов внедрена частично; сервисные учетные записи не всегда ротацируются;

IDS: датчики стоят на периметре, внутренний трафик покрыт ограниченно; обновление правил периодическое;

SIEM: корреляция развита на периметре, но сценарии по внутренним злоупотреблениям слабые;

резервное копирование: копии создаются регулярно, но тесты восстановления нерегулярны;

защита конечных точек: антивирус установлен везде, но расширенный поведенческий контроль покрывает только часть критичных рабочих мест;

изменения: инвентаризация неполная; конфигурационный дрейф выявляется эпизодически;

реагирование: план существует, но учения проводятся редко;

человеческий фактор: обучение проводится раз в год, результаты тестов на фишинг средние;

наблюдаемость: по филиалам потери событий достигают 20–30% в пиковые периоды.

Наблюдения переводятся в численные оценки (0–10) и коэффициенты полноты (0–1). Эта трансляция выполняется экспертной комиссией и фиксируется в протоколе СМИБ, что обеспечивает воспроизводимость: через квартал можно повторить оценку и сопоставить динамику. В таблице 2 приведен один возможный срез, согласованный группой.

Таблица 2 фиксирует исходный срез входных данных для демонстрационного кейса: по каждому из выбранных признаков приводятся (1) нормированная численная оценка состояния контроля по единой шкале 0–10 и (2) коэффициент полноты/доверия к наблюдениям по шкале 0–1, а также (3) краткий комментарий, который «привязывает» числа к фактам аудита и телеметрии (задержки патчей, исключения в сегментации, неполная инвентаризация, потери событий и т.д.). Срез формируется экспертной комиссией и закрепляется в протоколе СМИБ, что обеспечивает воспро-

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

изводимость: оценку можно повторять на регулярной основе и сопоставлять динамику между периодами и подразделениями. Практический смысл таблицы в том, что модель получает не только «оценку зрелости», но и качество исходных данных: при низкой полноте вывод не делает вид, что «всё измерено идеально», а аккуратно снижает вклад соответствующего признака в итоговую оценку.

Таблица 2 – Срез входных оценок для кейса и коэффициенты полноты

f	f	Комментарий
4.0	0.8	задержка патчей 30--60 дней, охват активов 70 %
5.5	0.7	сегментация есть, но много исключений
4.5	0.6	многофакторная аутентификация частично, слабая гигиена сервисных учеток
4.0	0.7	IDS преимущественно на периметре
3.5	0.6	SIEM: слабые сценарии внутренних злоупотреблений
6.0	0.8	бэкапы приемлемые, тесты нерегулярны
5.0	0.8	базовая защита везде, расширенная частично
3.5	0.5	инвентаризация неполная, дрейф фиксируется эпизодически
4.0	0.7	план есть, учения редки
4.5	0.9	обучение ежегодно, фишинг-тесты средние
4.0	0.6	потери событий 20--30% в пике

Пример расчета степеней принадлежности и учета неполноты

Для иллюстрации покажем вычисления по признаку $x_1 = 4.0$. По заданным функциям (низкое, среднее, высокое) получаем:

$$\mu_{1,\text{низк}}(4.0) = \mu(4.0; 0,0,2,4) = 0, \quad \mu_{1,\text{сред}}(4.0) = \mu(4.0; 2,4,6,8) = 1, \quad \mu_{1,\text{выс}}(4.0) = 0$$

С учетом неполноты $q_1 = 0.8$ и нейтральных нейтральных π=1/3 по (3):

$$\tilde{\mu}_{1,\text{сред}} = 0.8 \cdot 1 + 0.2 \cdot \frac{1}{3} \approx 0.867, \quad \tilde{\mu}_{1,\text{низк}} = 0.8 \cdot 0 + 0.2 \cdot \frac{1}{3} \approx 0.067, \quad \tilde{\mu}_{1,\text{выс}} \approx 0.067.$$

Тем самым модель отражает реальную ситуацию: «формально» показатель попадает в средний терм, но из-за неполного охвата активов остается неопределенность, которая влияет на надежность вывода.

Аналогично рассчитываются принадлежности для остальных признаков. Особенно важен признак наблюдаемости x_{11}, поскольку он влияет на «силу» многих правил: при низком x_{11} и малых q_j даже хорошие оценки отдельных контролей не дают оснований считать риск низким, поскольку «хорошее» могло быть «невидимым».

База правил: репрезентативное подмножество и смысловая интерпретация

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

Полная база в реальном проекте обычно содержит 40–80 правил для риска (и отдельные подбазы для параметров конфиденциальности, целостности, доступности). Ниже приведены примеры, отражающие типовые причинно-следственные связи. Сокращенно: уязвимости x_1, сегментация x_2, привилегии x_3, IDS x_4, SIEM x_5, бэкапы x_6, конечные точки x_7, изменения x_8, реагирование x_9, персонал x_{10}, наблюдаемость x_{11}.

1. Если x_1 низкое и x_5 начальная и x_{11} низкая, то риск высокий. Смысл: высокая вероятность незаметных компрометаций на фоне задержек патчей и слабой корреляции.

2. Если x_2 слабая и x_3 неуправляемое, то риск высокий. Смысл: латеральное перемещение и злоупотребление привилегиями становятся доступными.

3. Если x_6 надежное и x_9 рабочее, то риск по доступности снижается до среднего. Смысл: ущерб от разрушительных инцидентов уменьшается, даже если вероятность остается заметной.

4. Если высокое и x_5 развитая и x_{11} высокая, то риск низкий. Смысл: наблюдаемость и корреляция позволяют рано обнаруживать и сдерживать инциденты.

5. Если рискованный и x_1 низкое, то риск высокий. Смысл: фишинг и эксплуатация незакрытых уязвимостей усиливают друг друга.

6. Если x_8 хаотичное и x_7, низкая, то риск высокий. Смысл: дрейф конфигураций и рост поверхности атаки при слабой защите конечных точек.

7. Если x_1 среднее и x_2, достаточная и x_3 частичное и x_4 среднее и x_5, рабочая и x_{11} средняя, то риск средний. Смысл: система работает, но без «запаса прочности»

Заметим, что правила могут быть связаны с тактиками и техниками атакующих (Law of the Republic of Kazakhstan No. 94-V, 2013). Это облегчает коммуникацию с техническими командами: вместо абстрактного «низкая зрелость SIEM» обсуждается конкретный провал в сценарии обнаружения (например, отсутствие корреляции нетипичных административных действий), а затем формируется план устранения.

Численные оценки параметров безопасности и риска для кейса

Интерпретация результата принципиальна: риск в зоне высокий обусловлен не одним «плохим» контролем, а взаимным усилением факторов. Даже при приемлемых бэкапах и базовой защите конечных точек недостаточная корреляция событий и неполная наблюдаемость делают вероятность незаметной компрометации существенной. Именно этот эффект часто теряется в традиционных чек-листах: контроли формально внедрены, но «сшивка» в единый контур мониторинга и управления отсутствует.

После выполнения нечёткого вывода (параметры безопасности) и последующей агрегации риска по (Freund et al., 2015) получены значения, представленные в таблице 3. Приведены также краткие пояснения.

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

Таблица 3 – Результирующие оценки параметров безопасности и риска для кейса

Показатель	Значение	Уровень	Комментарий
$\hat{y}_C$	0.46	средний-низкий	слабое управление привилегиями и недостаточная корреляция повышают риск утечки
$\hat{y}_I$	0.49	средний-низкий	задержки патчей и дрейф конфигураций увеличивают вероятность модификации данных
$\hat{y}_A$	0.58	средний	резервное копирование частично компенсирует недостатки обнаружения
$\hat{y}_R$	0.67	высокий	сочетание слабого управления изменениями, недостаточного SIEM и неполной телеметрии

Численные значения параметров безопасности и интегрального риска получены в рамках вычислительного эксперимента в двух постановках: (i) расчёт по реальному срезу кейса (аудит + телеметрия) и (ii) проверка устойчивости на синтетических данных, моделирующих шум и пропуски наблюдений.Реальные данные. Входные оценки по выбранным признакам нормируются на шкале 0–10 и дополняются коэффициентами полноты/доверия (0–1) на основе фактической наблюдаемости и подтверждаемости данных; срез согласуется и фиксируется в протоколе СМИБ. Далее выполняются этапы нечёткого вывода Мамдани: фаззификация, учёт неполноты через дисконтирование степеней принадлежности, применение базы правил и дефаззификация. Синтетические данные. Формируются дополнительные срезы вокруг профиля кейса: к входным оценкам добавляется управляемый шум, а полнота данных варьируется в заданных диапазонах. Это используется для проверки, что итоговый риск не демонстрирует «скачков» при малых изменениях оценок и корректно реагирует на ухудшение наблюдаемости. Результат. По итогам расчёта получены значения параметров безопасности (конфиденциальность, целостность, доступность) и интегрального риска, приведённые в таблице 3: 0.46, 0.49, 0.58 и 0.67 соответственно.

Предложенная модель адресует типовую проблему практического оценивания ИБ: необходимость принимать управленческие решения при отсутствии полного набора измерений и при высокой доле качественных описаний. В отличие от чисто статистических подходов, нечёткая система сохраняет интерпретируемость и поддаётся аудиту: можно проследить, какие сочетания факторов привели к итоговой оценке, и объяснить, почему снижение наблюдаемости повышает риск даже при формально внедрённых контролях. Это особенно важно для распределённых организаций, где критичен не только факт наличия мер защиты, но и уверенность в том, что они реально функционируют во всех филиалах и сегментах, а не только на центральной площадке. Результаты демонстрационного кейса подтверждают данную логику: при умеренных значениях параметров конфиденциальности, целостности и доступности итоговый интегральный риск остаётся высоким (0.67), поскольку ком-

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

бинация слабой корреляции событий, проблем управления изменениями и неполной телеметрии формирует значимую вероятность незаметной компрометации и ошибочных управленческих выводов.

По сравнению с традиционными матрицами «вероятность×ущерб» нечёткая модель даёт более корректный мост между качественными наблюдениями и численной отчётностью. Она позволяет интегрировать шкалы зрелости и наблюдаемости без превращения их в «жёсткие числа» без оговорок, моделировать частичную истинность и неоднозначность (тем самым снижая риск ложной точности) и формально учитывать доверие к данным через коэффициенты полноты и отдельный индикатор наблюдаемости. В практическом плане это означает, что модель не «награждает» организацию низким риском только за наличие регламентов и отдельных средств защиты, если качество мониторинга и полнота данных не позволяют подтвердить фактическую эффективность мер.Сопоставляя подход с байесовскими сетями, следует отметить, что нечёткая модель проще в развёртывании при малом объёме данных и менее чувствительна к точности априорных предположений, однако уступает в возможностях строгого вероятностного вывода при наличии хорошей статистики. Практически это задаёт рациональную стратегию применения: нечёткий подход целесообразен как первичный контур управления и средство интеграции качественных сигналов, а по мере накопления данных его можно дополнять вероятностными моделями для отдельных сценариев, где есть достаточная база инцидентов и корректно выстроенная регистрация.Отдельная ценность модели заключается в совместимости с инструментальной инфраструктурой мониторинга и процессами СМИБ.

Технические источники (IDS/SIEM, средства управления уязвимостями, телеметрия конечных точек, учёт событий изменений) обеспечивают данные для значимой части признаков, тогда как процессы СМИБ формируют измеримые артефакты для организационных и поведенческих факторов (управление изменениями, реагирование, обучение и дисциплина исполнения). В результате модель не конкурирует с существующими практиками, а выступает надстроечным механизмом синтеза показателей, переводя разнородные сигналы в сопоставимые оценки и обеспечивая объяснимую приоритезацию улучшений.Ограничения исследования связаны, прежде всего, с трудоёмкостью построения и сопровождения базы правил, риском её «устаревания» при изменении инфраструктуры и кадровой текучести, а также с потенциальным ростом размера базы при расширении числа термов и признаков.

Эти риски снижаются, если рассматривать правила как управляемый артефакт СМИБ с регламентом пересмотра, применять сценарный принцип и модульность (раздельные подбазы для параметров безопасности, типов активов и классов сценариев), а также поддерживать конфигурационную дисциплину источников данных. Дополнительная методическая сложность связана с тем, что эталонные значения риска часто недоступны или искажены неполной регистрацией инцидентов; поэтому настройка и оптимизация весов должна опираться не только на статистику, но и на экспертные оценки и независимый аудит, иначе формальная «калибровка» мо-

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

жет закрепить искажения наблюдаемости.Перспективы развития работы целесообразно рассматривать в русле двух направлений. Во-первых, в архитектуре Zero Trust, где доступ предоставляется динамически по контексту и уровню доверия, нечёткая логика может выступать интерпиретируемым слоем агрегирования сигналов доверия (состояние устройства, риск учётной записи, полнота телеметрии) для микросегментации и политик доступа; при неполных данных это предпочтительнее бинарной логики «доверять/не доверять», поскольку позволяет вводить градуированную, управляемую осторожность. Во-вторых, в задачах мультимодального анализа артефактов, где объединяются текстовые отчёты, логи, сетевые потоки и бинарные данные, подходы визуализации бинарных артефактов (например, Byte2Image) и методы распознавания образов могут поставлять дополнительные признаки, а нечёткая модель-служить управляющим слоем, обеспечивающим объяснимость решения на уровне управления и явное отражение доверия к каждому источнику. Такой гибридный контур позволяет совместить сильные стороны глубинных моделей (извлечение признаков) и нечёткого вывода (интерпретируемая агрегация и управленческая пригодность результата).

Заключение.

Разработана интерпретируемая нечёткая модель оценивания параметров информационной безопасности информационных систем в условиях неполных наблюдений и преобладания качественных описаний. Сформирован набор из 11 входных признаков, охватывающий технические, организационные и поведенческие факторы, а также введён явный индикатор наблюдаемости и качества данных. Для признаков заданы лингвистические переменные и функции принадлежности, реализован вывод Мамдани с учётом весов правил и механизмом «ослабления» принадлежностей при пропусках телеметрии, что позволяет избегать ложной точности и сохранять объяснимость результата для аудита и управленческого контроля. Предложена методика настройки модели, включающая согласование экспертов по процедуре Дельфи, выявление и устранение противоречий в базе правил, ограниченную оптимизацию весов по калибровочным данным и обязательную проверку чувствительности и устойчивости.На демонстрационном кейсе типовой организации государственного сектора Республики Казахстан показано, как размытые наблюдения и фрагментарная телеметрия преобразуются в численные оценки конфиденциальности, целостности, доступности и итогового уровня риска. Полученные результаты демонстрируют ключевой практический вывод: наблюдаемость выступает не вторичным «качеством данных», а полноценным фактором риска, поэтому повышение полноты мониторинга и зрелости корреляции в SIEM способно снижать риск сильнее, чем локальные улучшения отдельных технических контролей без укрепления контура наблюдения.Для внедрения модели в организациях Казахстана целесообразно начинать с компактной и полностью трассируемой базы правил, изначально закладывать явный учёт пропусков и качества телеметрии (без подмены отсутствующих данных усреднением), сочетать экспертную калибровку с осторожной настройкой весов на доступной статистике

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

и материалах аудита, применять анализ чувствительности для обоснования приоритетов и бюджета, а также интегрировать результаты модели в регулярные процессы СМИБ, чтобы оценки использовались как инструмент управленческого цикла, а не разовый отчёт.

REFERENCES

Abubakar N. & Pranggono S. (2020). Fuzzy correlation of security events for SIEM enhancement // Journal of Information Security and Applications. [In Eng.].

Ahmed A. & Mahmoud R. (2019). Fuzzy inference for incident triage in security operations centers. Future Generation Computer Systems. [In Eng.].

Bengio Y., Goodfellow I. & Courville A. (2016). Deep learning. MIT Press.

Center for Internet Security. (2021). CIS critical security controls (v8). https://www.cisecurity.org/controls/v8. [In Eng.].

Committee for Technical Regulation and Metrology. (2023). Information security, cybersecurity and privacy protection. Information security management system. Requirements (ST RK ISO/IEC 27001-2023). Republic of Kazakhstan. [In Eng.].

Decree of the Government of the Republic of Kazakhstan No. 407. (2017, June 30). On approval of the Cybersecurity Concept ("Cyber Shield of Kazakhstan"). https://adilet.zan.kz/rus/docs/P1700000407 . [In Eng.].

Freund J. & Jones J. (2015). Measuring and managing information risk: A FAIR approach. Butterworth-Heinemann. https://doi.org/10.1016/C2013-0-09966-5 [In Eng.].

Goldberg D.E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley. [In Eng.].

International Organization for Standardization. (2022a). Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO/IEC 27001:2022). [In Eng.].

International Organization for Standardization. (2022b). Information security, cybersecurity and privacy protection — Guidance on managing information security risks (ISO/IEC 27005:2022). [In Eng.].

Jensen F.V. & Nielsen T.D. (2007). Bayesian networks and decision graphs (2nd ed.). Springer. DOI: https://doi.org/10.1007/978-0-387-68282-2 [In Eng.].

Joint Task Force Transformation Initiative. (2012). Guide for conducting risk assessments (NIST SP 800-30 Rev. 1). National Institute of Standards and Technology. DOI: https://doi.org/10.6028/NIST.SP.800-30r1 [In Eng.].

Kennedy J. & Eberhart R. (1995). Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks (ICNN'95). — Vol. 4. — Pp. 1942–1948. DOI: https://doi.org/10.1109/ICNN.1995.488968 [In Eng.].

Klir G.J. & Yuan B. (1995). Fuzzy sets and fuzzy logic: Theory and applications. Prentice Hall. [In Eng.].

Law of the Republic of Kazakhstan No. 94-V. (2013, May 21). On personal data and their protection. https://adilet.zan.kz/eng/docs/Z1300000094 [In Eng.].

Law of the Republic of Kazakhstan No. 418-V. (2015, November 24). On informatization. https://adilet.zan.kz/eng/docs/Z1500000418 [In Eng.].

Linstone H. & Turoff M. (Eds.). (1975). The Delphi method: Techniques and applications. Addison-Wesley. [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.].

Miller D., Harris S., Harper A., VanDyke S. & Blask C. (2011). Security information and event management (SIEM) implementation. McGraw-Hill. [In Eng.].

MITRE. (n.d.). MITRE ATT&CK®: Adversarial tactics, techniques, and common knowledge. Retrieved January 11, 2026, from https://attack.mitre.org/ [In Eng.].

Nataraj L., Karthikeyan S., Jacob G. & Manjunath B.S. (2011, July). Malware images: Visualization and automatic classification // Proceedings of the 8th International Symposium on Visualization for Cyber Security (VizSec'11). Pp. 1–7. DOI: https://doi.org/10.1145/2016904.2016908 [In Eng.].

Rose S., Borchert O., Mitchell, S. & Connelly S. (2020). Zero trust architecture (NIST SP 800-207) // National Institute of Standards and Technology. DOI: https://doi.org/10.6028/NIST.SP.800-207 [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.].

Shafer G. (1976). A mathematical theory of evidence. Princeton University Press. DOI: https://doi.

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

org/10.1515/9780691214696 [In Eng.].

Shameli-Sendi A., Aghababaei-Barzegar R. & Cheriet M. (2016). Taxonomy of information security risk assessment (ISRA). Computers & Security. — Vol. 57. — Pp. 14–30. DOI: https://doi.org/10.1016/j.cose.2015.11.001 [In Eng.].

Srinivasan S. & Karthik A. (2013). Fuzzy logic based information security risk assessment: A review and framework // International Journal of Information Security Science. [In Eng.].

Sugeno M. (1985). Industrial applications of fuzzy control. North-Holland. DOI: https://doi.org/10.5555/537323 [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.].

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. 468–483
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.029
УДК 004.05:004.67

EVALUATION OF THE EFFECTIVENESS OF SYNTHESIS OF STOCHASTIC MODELS WITH CONTROLLED PROPERTIES

E.A. Pustovoy[1]*, O.A. Pustovaya[2], A.N. Raushanova[1], I.S. Zaurbekov[1]
[1]International University of Information Technology, Almaty, Kazakhstan;
[2]Federal State Budgetary Educational Institution of higher Education "Far Eastern State Agrarian University", Blagoveshchensk, Russia.
E-mail: pus14@yandex.ru

E. Pustovoy — Associate Professor at the Department of Cybersecurity, International University of Information Technology, Almaty, Kazakhstan
E-mail: pus14@yandex.ru, https://orcid.org/0000-0002-1197-7410;
O. Pustovaya — Candidate of Agricultural Sciences, Associate Professor of the Department of EiATP, Far Eastern State Agrarian University, Blagoveshchensk, Russian Federation
https://orcid.org/ 0000-0002-4487-8320;
A. Raushanova — Assistant G2, JSC "International University of Information Technology", Almaty, Kazakhstan
https://orcid.org/0009-0006-8471-9873;
I. Zaurbekov — Senior Lecturer, International University of Information Technologies, Almaty, Republic of Kazakhstan
https://orcid.org/0009-0003-0759-9351.

© E.A. Pustovoy, O.A. Pustovaya, A.N. Raushanova, I.S. Zaurbekov

Abstract. This paper investigates the possibility of using large language models (LLMs) as sources of stochastic binary sequences with controllable statistical properties. The study demonstrates the temperature hyperparameter, which regulates the dispersion of the output token probability distribution, can be regarded as a tool for controlled entropy management of generated data. Based on the GPT-2 model, a method for converting probabilistic LLM outputs into fixed-length binary sequences is implemented and evaluated using entropy analysis and an adapted set of statistical tests analogous to NIST SP 800–22. The experiments reveal a monotonic increase in entropy with increasing temperature and a significant improvement in the statistical characteristics of the sequences. At temperatures $T \geq 2.0$, up to 83.3 % of the applied tests are passed, corresponding to the level of high-quality pseudorandom sequences for non-cryptographic applications. At the same time, a fundamental limitation of the approach is identified: weak but statistically significant periodic

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

components persist due to the deterministic architecture of the model. The results define the applicability limits of LLMs in modeling, data synthesis, and algorithm testing tasks.

Keywords: neural, random, hyperparameter, entropy, generator, NIST, GPT

For citation: E.A. Pustovoy, O.A. Pustovaya, A.N. Raushanova, I.S. Zaurbekov (2026). Evaluation of the effectiveness of synthesis of stochastic models with controlled properties // International Journal of Information and Communication Technologies. Vol. 7. No. 26. Pp. 468–483. https://doi.org/10.54309/IJICT.2026.26.2.029. (In Russ.).

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

БАСҚАРЫЛАТЫН ҚАСИЕТТЕРІ БАР СТОХАСТИКАЛЫҚ МОДЕЛЬДЕРДІ СИНТЕЗДЕУДІҢ ТИІМДІЛІГІН БАҒАЛАУ

Е.А. Пустовой[1]*, О.А. Пустовая[2], А.Н. Раушанова[1], И.С. Заурбеков[1]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Федералдық мемлекеттік бюджеттік жоғары білім беру мекемесі «Шығыс Қытай мемлекеттік ауылшаруашылық университеті», Благовещенск, Ресей.
E-mail: pus14@yandex.ru

Пустовой Е. — «киберқауіпсіздік» кафедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: pus14@yandex.ru, https://orcid.org/0000-0002-1197-7410;
Пустовая О. — ауылшаруашылық ғылымдарының кандидаты, ЭиАТП кафедрасының доценті, Қиыр Шығыс ГАУДАҒЫ ФГБОУ, Благовещенск, РФ
https://orcid.org/ 0000-0002-4487-8320;
Раушанова А. — аға оқытушы, «Ақпараттық технологиялар халықаралық университеті» АҚ, Алматы, Қазақстан
https://orcid.org/0009-0006-8471-9873;
Заурбеков И. — сеньор-лектор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0009-0003-0759-9351.

© Е.А. Пустова, О.А. Пустовая, А.Н. Раушанова, И.С. Заурбеков

Аннотация. Бұл жұмыста үлкен тілдік модельдерді (LLM) басқарылатын статистикалық қасиеттері бар стохастикалық екілік тізбектердің көзі ретінде пайдалану мүмкіндігі зерттеледі. Токендердің шығыс ықтималдық таралуының дисперсиясын реттейтін температура гиперпараметрі туындатылған деректердің энтропиясын басқарудың құралы ретінде қарастырылады. GPT-2 моделіне негізделіп, LLM-нің ықтималдық шығыстарын бекітілген ұзындықтағы екілік тізбектерге түрлендіру әдістемесі іске асырылды және ол энтропиялық талдау мен NIST SP 800–22 стандартына ұқсас бейімделген статистикалық тесттер жиынтығы арқылы бағаланды. Эксперименттік нәтижелер температураның өсуімен энтропияның монотонды артуын және

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

тізбектердің статистикалық сипаттамаларының айтарлықтай жақсарғанын көрсетеді. T ≥ 2.0 температурада қолданылған тесттердің 83,3 %-ына дейін сәтті өтеді, бұл криптографиялық емес қолданбалар үшін жоғары сапалы псевдокездейсоқ тізбектер деңгейіне сәйкес келеді. Сонымен қатар, модель архитектурасының детерминирленген сипатына байланысты әлсіз, бірақ статистикалық мәнді периодтық компоненттердің сақталатыны анықталды. Алынған нәтижелер LLM-дерді модельдеу, деректер синтезі және алгоритмдерді тестілеу салаларында қолданудың шекараларын айқындайды.

Түйінді сөздер: нейрондық, кездейсоқ, гиперпараметр, энтропия, генератор, NIST, GPT

Дәйексөздер үшін: Е.А. Пустовой, О.А. Пустовая, А.Н. Раушанова, И.С. Заурбеков (2026). Басқарылатын қасиеттері бар стохастикалық модельдердің синтезінің тиімділігін бағалау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No. 26. Б. 468–483. https://doi.org/10.54309/IJICT.2026.26.2.029. (Орыс. тіл.).

Мүдделер қақтығысы: Авторлар қызығушылықтар қақтығысының жоқтығын мәлімдейді.

ОЦЕНКА ЭФФЕКТИВНОСТИ СИНТЕЗА СТОХАСТИЧЕСКИХ МОДЕЛЕЙ С УПРАВЛЯЕМЫМИ СВОЙСТВАМИ

Е.А. Пустовой[1]*, О.А. Пустовая[2], А.Н. Раушанова[1], И.С. Заурбеков[1]

[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Федеральное государственное бюджетное образовательное учреждение высшего образования «Дальневосточный государственный аграрный университет», Благовещенск, Россия.
E-mail: pus14@yandex.ru

Пустовой Е. — ассоциированный профессор кафедры «Кибербезопасность», Международный университет информационных технологий, Алматы, Казахстан
E-mail: pus14@yandex.ru, https://orcid.org/0000-0002-1197-7410;
Пустовая О. — кандидат сельскохозяйственных наук, доцент кафедры ЭиАТП, ФГБОУ ВО Дальневосточный ГАУ, Благовещенск, РФ
https://orcid.org/0000-0002-4487-8320;
Раушанова А. — старший преподаватель, АО «Международный университет информационных технологий», Алматы, Казахстан
https://orcid.org/0009-0006-8471-9873;
Заурбеков И. — сеньор-лектор, Международный университет информационных технологий, Алматы, Республика Казахстан
https://orcid.org/0009-0003-0759-9351.

© Е.А. Пустовой, О.А. Пустовая, А.Н. Раушанова, И.С. Заурбеков

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

Аннотация. В работе исследуется возможность использования больших языковых моделей (LLM) в качестве источников стохастических бинарных последовательностей с управляемыми статистическими свойствами. Показано, что гиперпараметр температуры, регулирующий дисперсию выходного распределения вероятностей токенов, может рассматриваться как инструмент контролируемого управления энтропией генерируемых данных. На основе модели GPT-2 реализована методика преобразования вероятностных выходов LLM в бинарные последовательности фиксированной длины и проведена их оценка с использованием энтропийного анализа и адаптированного набора статистических тестов, аналогичных NIST SP 800–22. Экспериментальные результаты демонстрируют монотонный рост энтропии с увеличением температуры и существенное улучшение статистических характеристик последовательностей. При температурах $T \geq 2.0$ достигается прохождение до 83,3 % применённых тестов, что соответствует уровню качественных псевдослучайных последовательностей для некриптографических применений. Одновременно выявлено ограничение подхода, связанное с наличием слабовыраженных периодических компонент, обусловленных архитектурной детерминированностью модели. Полученные результаты определяют границы применимости LLM в задачах моделирования, синтеза данных и тестирования алгоритмов.

Ключевые слова: нейросеть, случайный, гиперпараметр, энтропия, генератор, NIST, GPT

Для цитирования: Е.А. Пустовой, О.А. Пустовая, А.Н. Раушанова, И.С. Заурбеков (2026). Оценка эффективности синтеза стохастических моделей с управляемыми свойствами//Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 468–483. https://doi.org/10.54309/IJICT.2026.26.2.029. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение.

Вопрос предсказуемости генераторов псевдослучайных последовательностей (PRNG) заключается в том, что такие последовательности чисел полностью детерминированы; зная начальное число (seed) и алгоритм, можно предсказать всю последовательность и это является критическим недостатком. Кроме того предъявляются дополнительные требования к таким последовательностям, например в криптографии (NIST, 2010; Menezes et al., 1996). Проблема генерации случайных и псевдослучайных последовательностей является фундаментальной для криптографии, статистического моделирования и вычислительных экспериментов, поскольку от их качества зависит стойкость криптосистем к атакам и надежность моделирования (Charu et al., 2024). Традиционно для этих целей используются детерминированные алгоритмические генераторы (PRNG), такие как Mersenne Twister, или аппаратные источники энтропии. Однако с развитием искусственного интеллекта возникает вопрос о стохастическом потенциале больших языковых моделей (LLM)-мощных генеративных систем, оперирующих вероятностными распределениями токенов

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

(Radford et al., 2019; Vaswani et al., 2017)

Основная теоретическая сложность заключается в том, что нейронные сети по своей природе обучаются выявлять и воспроизводить закономерности в данных, минимизируя энтропию ошибки предсказания (Agarwal et al., 2025). Генерация же истинно случайных, непрогнозируемых последовательностей противоречит этой цели (Рахим и др., 2021). Тем не менее, в архитектуре LLM существует встроенный механизм управления стохастичностью. Гиперпараметр температуры (temperature), позволяет масштабировать логиты перед применением функции softmax, тем самым регулируя дисперсию выходного распределения вероятностей (Roberts, 2024).

С математической точки зрения, увеличение температуры приводит распределение вероятностей к состоянию, близкому к равномерному, что теоретически должно коррелировать с ростом информационной энтропии Шеннона. Если эта корреляция носит устойчивый и предсказуемый характер, то LLM, несмотря на свою детерминированную архитектуру, может быть трансформирована в управляемый источник стохастических данных (Shimao, 2024). Это превращает «недостаток» модели (ошибки предсказания при высокой температуре) в функциональный инструмент для синтеза случайности.

Таким образом используя встроенный параметр нейронных сетей, мы можем выдвинуть гипотезу о том, что целенаправленное варьирование встроенного в большие языковые модели (LLM) гиперпараметра температуры, ответственного за дисперсию выходного распределения (White, 2025), позволяет напрямую управлять энтропией этого распределения. Тем самым способствуя преодолению фундаментального противоречия между целью нейросетей (минимизация энтропии ошибки) и задачей генерации случайных данных. Предполагается, что такой вид генерации числовых последовательностей не будет привязан к детерминированному алгоритму. В результате исследования ожидается получить возможность генерации числовых последовательностей, статистические свойства которых будут приближаться, по мере роста температуры, к характеристикам псевдослучайных последовательностей.

Цель работы-экспериментально проверить данную гипотезу и оценить степень влияния температуры (T) на стохастические характеристики бинарных числовых последовательностей, сгенерированных на основе языковой модели (GPT-2). Для этой цели будут использованы методы энтропийного анализа выходной числовой последовательности и адаптированный набор статистических тестов в соответствии с требованиями Национального института стандартов и технологий агентство Министерства торговли США (NIST), что позволит количественно измерить достижимый уровень качества числовой последовательности.

Материалы и методы.

В качестве базовой большой языковой модели использовалась облегчённая языковая модель GPT-2 (Green et al., 2024) из библиотеки Hugging Face Transformers. Выбор данной модели обусловлен следующими причинами:

Архитектурная репрезентативность: GPT-2 сохраняет ключевые архитектур-

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

ные и функциональные принципы оригинальной большой языковой модели семейства GPT- генеративных предобученных трансформеров, включая механизм внимания и авторегрессивную генерацию, что делает её валидным объектом для исследования стохастических свойств, присущих данному семейству моделей.

Вычислительная эффективность: значительно меньший размер модели (по сравнению с полной GPT-2) позволил провести масштабный цикл экспериментов с многократным варьированием гиперпараметров и генерацией длинных последовательностей (10 000 бит) при ограниченных вычислительных ресурсах.

Методология: использование стандартной, широкодоступной и воспроизводимой модели (GPT-2) позволяет сместить акцент исследования с особенностей конкретной реализации на общую методологию управления случайностью через температуру, что повышает практическую значимость и проверяемость результатов.

Для проведения расчётов использовалась среда Google Colab, которая предоставляет доступ к графическому процессору NVIDIA T4 через среду выполнения с поддержкой GPU. Для ускорения процесса инференса (вывода результатов) модели использовалась платформа параллельных вычислений CUDA, интегрированная в библиотеку PyTorch (torch). Это позволило значительно сократить время генерации длинных битовых последовательностей.

Для преобразования вероятностных выходов LLM в бинарные последовательности фиксированной длины (10 000 бит) существует три метода вероятностного сэмплирования, представляющих различные стратегии извлечения случайности из распределения токенов модели:

Метод сэмплирования (Sampling) (1). Данный метод является наиболее прямым аналогом стандартного режима работы авторегрессивной языковой модели, где следующий токен выбирается стохастически в соответствии с прогнозируемым распределением, скорректированным температурой. Для преобразования в биты ID выбранного токена брался по модулю 256, и использовались его 8 младших бит. Метод извлечения по модулю был выбран в качестве простого, не зависящего от архитектуры базового метода, который не вносит дополнительной случайности в постобработку. Этот подход позволяет оценить «естественную» стохастичность модели в её основном режиме генерации.

Вероятностный метод (Probability) (2). Для исследования влияния «хвоста» распределения был реализован метод, основанный на анализе наименее вероятного токена среди топ-8 кандидатов. Бит генерировался на основе нормализованной вероятности этого токена. Данный подход направлен на оценку того, может ли информация о маловероятных событиях, усиливающаяся с ростом температуры, служить источником случайности.

Энтропийный метод (Entropy) (3). Для прямого контроля над уровнем неопределённости на каждом шаге генерации использовался метод, основанный на вычислении локальной энтропии распределения токенов по формуле Шеннона. Нормализованная энтропия сравнивалась с порогом 0.5 для установки значения бита (1– высокая неопределённость, 0–низкая). Этот метод позволяет явно связать меру информацион-

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

ной случайности (энтропию) с генерируемым битовым потоком.

Выбор метода: для основного цикла тестирования использовался метод сэмплирования (Sampling), как наиболее прямой аналог работы авторегрессивной языковой модели, при котором следующий токен выбирается случайно в соответствии с предсказанным распределением вероятностей.

Вероятностный метод (2) анализирует хвост распределения, что полезно для исследований, но нестабильно в маловероятных токенах.

Энтропийный метод (3) требует вычисления энтропии на каждом шаге, что увеличивает вычислительную нагрузку без значимых преимуществ в качестве выходной последовательности (Roberts, 2024).

Вероятностный и энтропийный методы были изучены в предварительных экспериментах, но не обеспечили ощутимых статистических улучшений по сравнению с выборкой, увеличив при этом нестабильность или вычислительные затраты; поэтому приводятся только результаты, основанные на выборке.

В действующих моделях GPT (например, GPT-3/4/5) основным методом генерации является метод сэмплирования (Sampling) (1). В качестве входного промпта была выбрана детерминированная, семантически нейтральная и не порождающая скрытых паттернов фраза:

«Continue the sequence: ABCDEFGHIJKLMNOPQRSTUVWXYZ».

Данный промпт обеспечивает:

1. Воспроизводимость: фиксированная начальная последовательность гарантирует одинаковые стартовые условия для всех экспериментов при изменении температуры.

2. Семантическая нейтральность: буквенный ряд ABC… не несёт явной статистической или смысловой нагрузки, которая могла бы систематически влиять на распределение генерируемых битов.

3. Отсутствие скрытых закономерностей: промпт не содержит чисел, специальных символов или естественно-языковых конструкций, способных вызвать нежелательные корреляции в выходных данных.

Таким образом, выбранный промпт позволяет минимизировать внешние факторы, влияющие на стохастические свойства генерируемых последовательностей, и сосредоточить анализ исключительно на эффекте гиперпараметра температуры.

Для каждого значения температуры из диапазона [0.5,1.0,1.5,2.0,3.0,5.0,10.0] вычислялась средняя энтропия распределения вероятностей токенов (Javier., 2025: 5) как ключевая мера неопределённости и стохастичности выходных данных модели. В соответствии с выдвинутой гипотезой, энтропия выходного распределения должна демонстрировать монотонную зависимость от гиперпараметра температуры и выступать определяющим фактором степени случайности генерируемой битовой последовательности. Энтропия вычислялась по формуле Шеннона, используемой здесь как операциональная мера стохастичности выхода модели:

$$\mathrm{H}(X) = -\sum_{j=1}^{n} p_j \ln p_j \qquad (1)$$

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

где p_j - вероятность выбора j-го токена из словаря модели.

В ходе анализа рассматривались стандартное отклонение энтропии, энтропийная эффективность, число «активных» токенов ($p > 0.001$) а также равномерность распределения, оцениваемая с использованием критерия Колмогорова-Смирнова.

Качество полученных бинарных последовательностей оценивалось с помощью набора из шести статистических тестов (табл.1), аналогичных стандарту NIST SP 800-22 (Menezes et al., 1996).

Таблица 1 – Таблица статистических тестов

№	Наименование	Критерий
1	Frequency Test	p-value > 0.01
2	Runs Test	p-value > 0.01
3	Longest Run Test	max_run < expected_max_run * 1.5
4	Entropy Test	efficiency > 0.99
5	Spectral Test	power_ratio < 5
6	Approximate Entropy Test	approx_ent > 0.1

Уровень значимости для всех тестов, согласно спецификации NIST SP 800-22 (Statistical Test Suite), устанавливался $\alpha = 0.01$.

Для комплексной оценки влияния температуры на стохастические свойства бинарных последовательностей, сгенерированных моделью GPT-2, был проведен статистический анализ, включающий NIST-подобные тесты и энтропийный анализ. Основные результаты представлены в таблице 2.

Таблица 2 – Результаты NIST-подобного тестирования бинарных последовательностей при различных температурах

Температура	Баланс 0/1	Серии битов	Макс. длина серии	Энтропия	Спектральный анализ	Вероятность шаблонов	% пройденных тестов
1.0	Нет	Нет	Нет	Да	Нет	Да	33.3% (2/6)
2.0	Да	Да	Да	Да	Нет	Да	83,3% (5/6)
3.0	Да	Да	Да	Да	Нет	Да	83,3% (5/6)
5.0	Да	Да	Да	Да	Нет	Да	83,3% (5/6)
7.0	Да	Да	Да	Да	Нет	Да	83,3% (5/6)
10.0	Да	Да	Да	Да	Нет	Да	83,3% (5/6)
15.0	Да	Да	Да	Да	Нет	Да	83,3% (5/6)

1. Frequency Test (тест на частотное распределение.): при низкой температуре (T=1.0) наблюдалось статистически значимое отклонение от равномерного распределения битов (рис.1) (соотношение 45.34 % единиц к 54.66 % нулей, p-value ≈ 0.0000).С ростом температуры до T=2.0 и выше баланс 0 и 1 стремился к идеальному соотношению 50/50 (p-value > 0.93), что свидетельствует о достижении статистической неотличимости от равномерного распределения.

2. Runs Test (тест на серии): при температуре T=1.0 последовательность демонстрировала недостаточное количество серий (z-score = -4.04, p-value≈0.00005),

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

указывая на кластеризацию битов (рис.3). При T≥2.0 количество серий становилось статистически неотличимым от ожидаемого для случайной последовательности (p-value>0.04), что согласуется с визуальной структурой последовательности (рис.2).

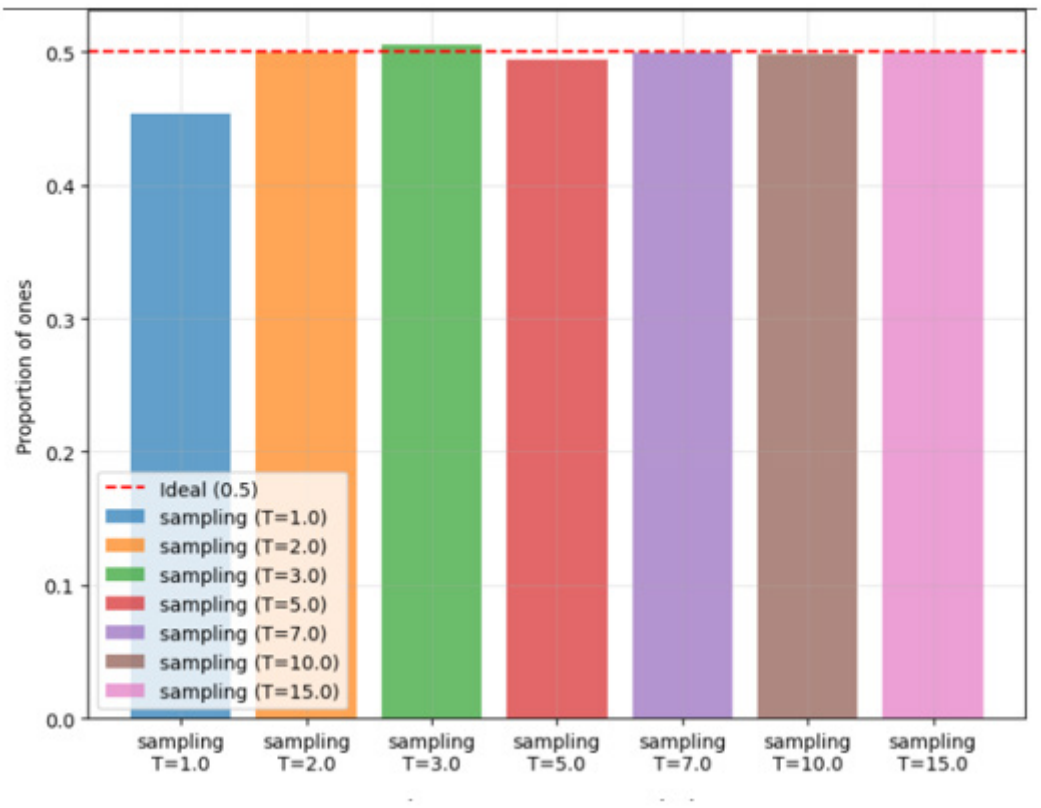


Рис. 1. Баланс 0 и 1 в последовательности

3. Longest Run Test (тест на максимальную длину серии): во всех экспериментах сгенерированные бинарные последовательности удовлетворяли критериям данного теста при любом значении температуры. При T=1.0 максимальная длина серии составляла 19 битов, что превышает теоретически ожидаемое для случайной последовательности значение, равное 16. С ростом температуры этот показатель снижался до диапазона 11–14 битов, что уже не превышало порогового критерия. Данный результат указывает на то, что увеличение температуры способствует сокращению длины непрерывных серий одинаковых битов, что соответствует желаемому свойству псевдослучайных последовательностей и подтверждает эффективность предложенного подхода (рис.4).

4. Entropy Test (тест на энтропию) и Approximate Entropy Test (тест на приблизительную энтропию): энтропия самой битовой последовательности была близка к максимальной (эффективность>0.99) даже при T=1.0, и оставалась стабильно высокой при всех температурах. Аналогично, тест на приблизительную энтропию,

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

оценивающий вероятность появления шаблонов, успешно проходился при всех T, демонстрируя высокую нерегулярность последовательностей.

5. Spectral Test (спектральный тест): При T=1.0 тест выявил выраженные периодические закономерности в данных (power_ratio = 63.12), что свидетельствует о высокой степени структурированности генерируемой последовательности. С повышением температуры до T=2.0 мощность доминирующей частоты резко снизилась (power_ratio ≈ 8.67). Однако на всех тестируемых температурах данный показатель оставался выше критического значения 5, что указывает на наличие слабовыраженных, но статистически значимых периодических компонент даже в режиме повышенной стохастичности. В результате спектральный тест не пройден ни при одном значении температуры.

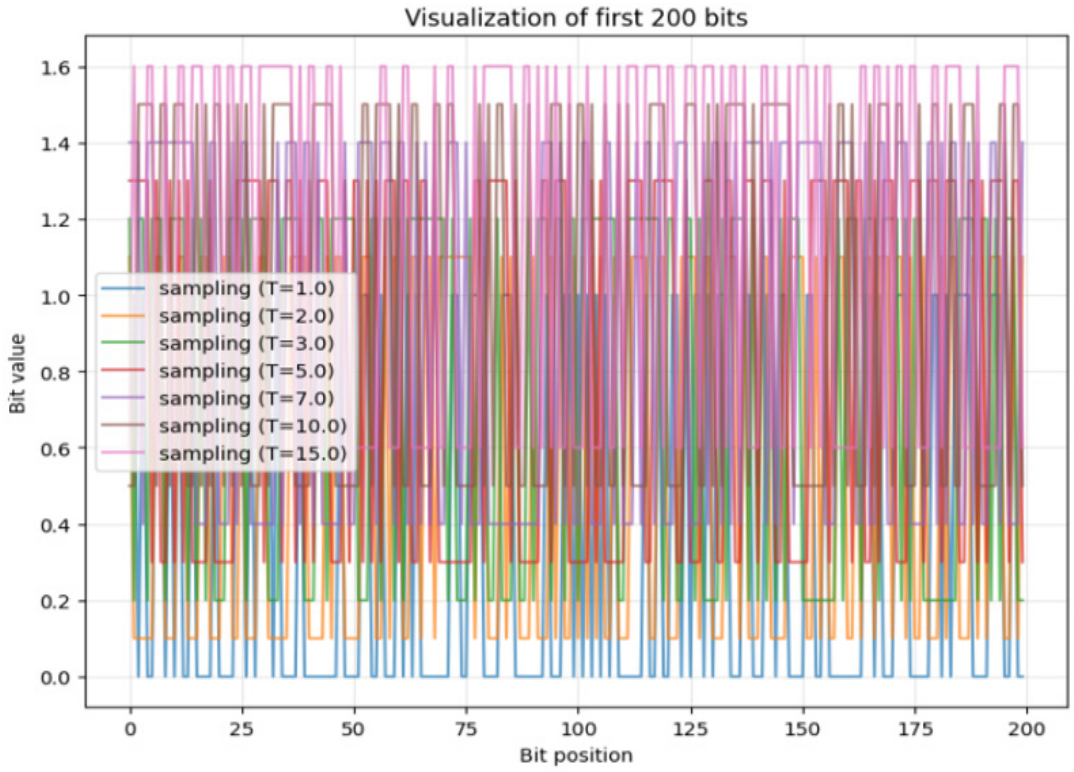


Рис. 2. Визуализация первых 200 бит

В данной работе спектральный анализ использовался в диагностическом режиме. В качестве вспомогательной количественной характеристики применялось отношение мощности доминирующей спектральной компоненты к средней мощности спектра (power_ratio). Пороговое значение power_ratio<5 рассматривалось как эмпирический индикатор отсутствия выраженных доминирующих периодических компонент, а не как формальный критерий статистического теста.

Таким образом, непрохождение спектрального теста не является следствием некорректности выбранного критерия или параметров тестирования, а отражает

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

фундаментальное ограничение подхода использования больших языковых моделей в качестве источников стохастических бинарных последовательностей. Причина данного ограничения заключается в детерминированной архитектуре LLM и характере обучающих данных, содержащих устойчивые структурные зависимости. Эти зависимости не полностью устраняются даже при высоких значениях температуры и проявляются в виде остаточных спектральных особенностей выходного сигнала (Kiyoshiro et al., 2023).

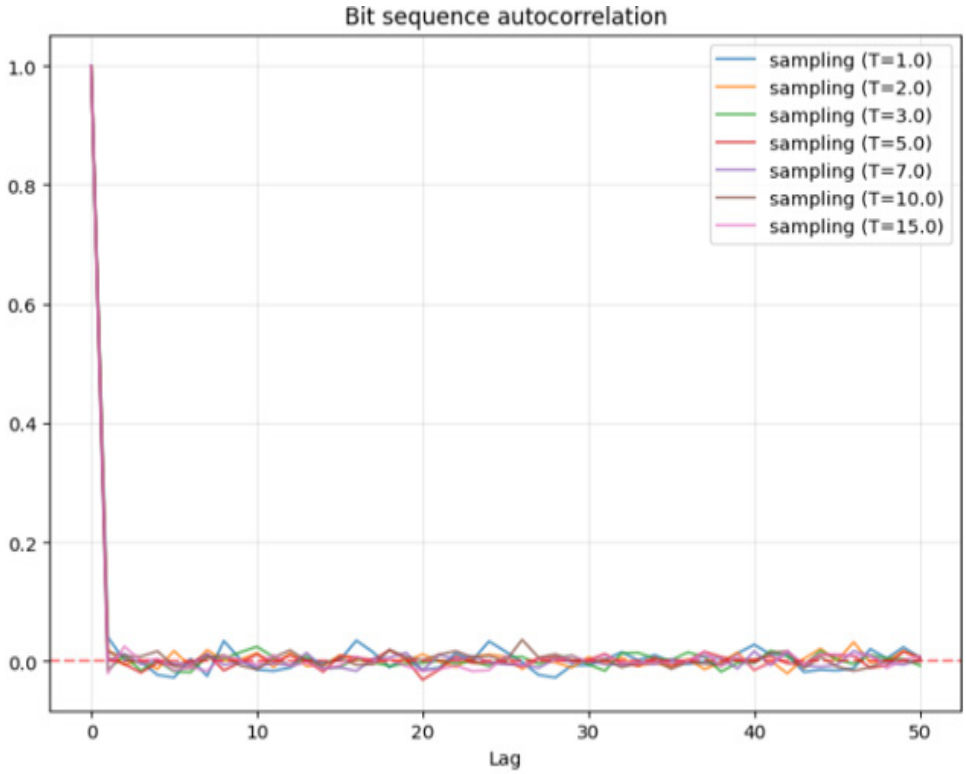


Рис. 3. Автокорреляция битовой последовательности

Фундаментальной причиной наблюдаемых изменений в статистических свойствах битовых последовательностей является прямое влияние гиперпараметра температуры на энтропию распределения вероятностей токенов на выходе модели.

Механизм этой связи, следующий: температура управляет «размытием» распределения. При низкой температуре функция softmax усиливает разрыв между вероятностями наиболее вероятных и маловероятных токенов, делая распределение «пикообразным» и детерминированным. С ростом температуры это распределение «сглаживается», приближая вероятности всех токенов к равномерному распределению.

Изменение энтропии: «пикообразное» распределение (низкая T) обладает низкой энтропией (малая неопределённость выбора), в то время как «сглаженное»

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

(высокая Т)-высокой энтропией (высокая неопределённость).

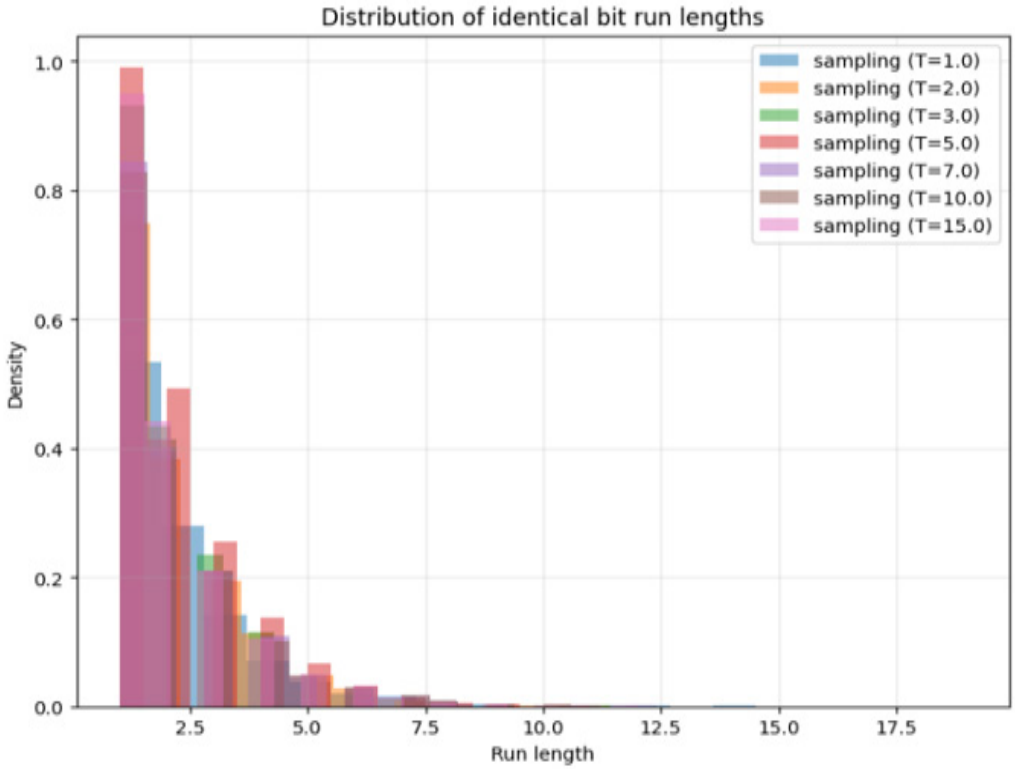


Рис. 4. Распределение длин битовых серий

Влияние на битовую последовательность: поскольку каждый бит в выходной последовательности является производным от выбора или анализа этого распределения вероятностей (через методы сэмплирования, вероятностный или энтропийный), его статистические свойства напрямую наследуют степень неопределённости исходного распределения. Низкая энтропия приводит к предсказуемым, коррелированным битам (провал частотного теста, наличие серий), тогда как высокая энтропия повышает независимость и равномерность битов, улучшая результаты большинства статистических тестов.

Таким образом, температура выступает в роли контроллера энтропии на уровне токенов, а наблюдаемое улучшение статистических свойств бинарного потока является каскадным следствием этого управления фундаментальной мерой неопределённости в модели (Weihua, 2025).

Анализ зависимости энтропии от температуры (рис. 5) позволяет выделить три характерных режима работы модели.

Таким образом, при высоких значениях температуры наблюдается асимптотическое приближение распределения вероятностей токенов к равномерному.

В детерминированном режиме (T<1.0) энтропия распределения остаётся близкой к нулю, что отражает доминирование нескольких наиболее вероятных то-

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

кенов и приводит к формированию высокоструктурированных, предсказуемых битовых последовательностей.

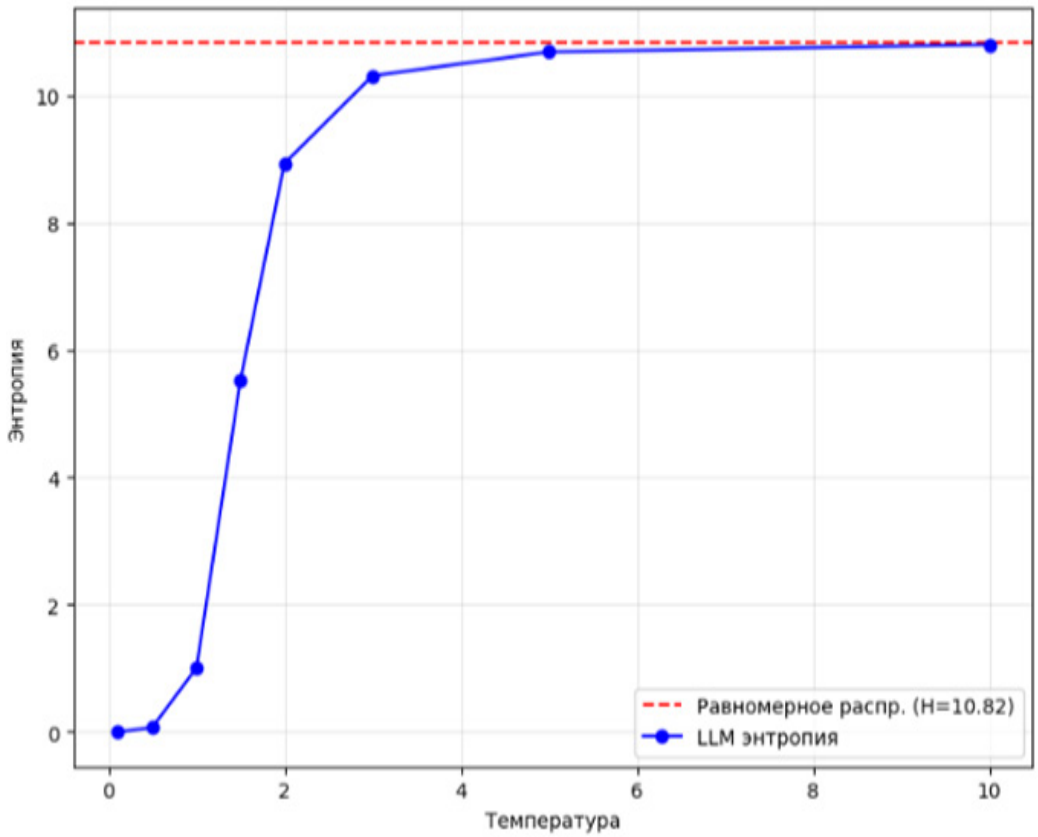


Рис. 5. Зависимость энтропии от температуры в сравнении с равномерным распределением

График демонстрирует нелинейный рост средней энтропии распределения токенов от ~0.0 при T=0.1 до плато около 10.79 бит при T=10.0 (Huang et al., 2024). Горизонтальная линия на графике соответствует теоретически максимальной энтропии H_{max} для модели GPT-2, которая определяется размером словаря *N=50257* и вычисляется по формуле Шеннона как

$$H_{max} = \ln(n) = \ln(50257) \approx 10.86$$

В переходном режиме (T≈1.0–2.0) наблюдается резкий нелинейный рост энтропии, сопровождающийся сглаживанием распределения вероятностей и качественным улучшением результатов частотных тестов и тестов на серии.

В стохастическом режиме (T>3.0) энтропия достигает плато, асимптотически приближаясь к максимально возможному значению для данной модели; при этом распределение вероятностей токенов становится близким к равномерному, что обе-

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

спечивает стабильно высокие (до 83,3%), хотя и не идеальные, результаты статистического тестирования.

Результаты и обсуждение.

Результаты убедительно доказывают, что гиперпараметр температуры является мощным и управляемым инструментом (Chang et al., 2024) инжекции стохастичности в выходные данные LLM (Huang et al., 2024; Javier, 2025). Повышение температуры приводит к монотонному росту энтропии распределения токенов, что каскадно улучшает статистические свойства результирующих битовых последовательностей.

Достижимый уровень случайности: при температурах $T \geq 2.0$ LLM (в данном случае GPT-2) способна генерировать битовые последовательности, успешно проходящие 83,3% базовых статистических тестов (Hopkins et al., 2023). Это характеризует их как качественные псевдослучайные последовательности, пригодные для широкого круга приложений, не требующих криптографической стойкости (например, моделирование, генерация синтетических данных, тестирование).

Фундаментальные ограничения: непрохождение тестов Spectral Test даже при высоких температурах указывает на архитектурные ограничение (Chen et al., 2024) последовательности, присущее детерминированной архитектуре LLM, обученной на коррелированных текстовых данных. Модель не может полностью устранить следы своей структурной предрасположенности, что проявляется в генерации аномально длинных серий и слабых периодичностях.

Практическая значимость: несмотря на ограничения, работа демонстрирует, что LLM могут выступать в роли конфигурируемых генераторов стохастичности. Главное практическое преимущество-плавное управление уровнем «хаотичности» через единственный параметр (температуру) (Roberts, 2024), что открывает перспективы для их использования в задачах, требующих контролируемого и семантически связанного случайного выбора.

Таким образом, LLM не заменяют криптографические генераторы случайных чисел, но предлагают уникальный гибридный подход к генерации стохастических последовательностей (Grados, 2015), занимая нишу между традиционными PRNG и источниками истинной энтропии.

Заключение.

В рамках данного исследования была экспериментально подтверждена возможность использования больших языковых моделей для синтеза стохастических последовательностей (Weihua et al., 2025) с управляемыми свойствами. Ключевые выводы работы:

Управление энтропией: гиперпараметр температуры LLM является эффективным средством контроля энтропии выходного распределения вероятностей токенов. Это позволяет напрямую и детерминировано влиять на фундаментальную меру неопределённости в работе модели, что открывает путь к управляемой стохастической генерации на её основе.

Качество случайности: при температурах $T \geq 2.0$ модель (GPT-2) генерирует

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

битовые последовательности, успешно проходящие 83.3% адаптированных тестов NIST (5 из 6 тестов), что свидетельствует о достижении статистических характеристик, близких к качественным псевдослучайным последовательностям, пригодным для широкого класса некриптографических задач (Yasmeen et al., 2025).

Практическая применимость: предложенный подход открывает перспективы применения LLM в качестве конфигурируемых стохастических генераторов для задач моделирования, тестирования алгоритмов и синтеза данных (Maged et al., 2023), где не требуется абсолютная криптографическая стойкость (Brown,2025), но необходима контролируемая и семантически связная (в случае текстовой генерации) степень случайности. Главное практическое преимущество, это возможность тонкой настройки уровня «хаотичности» выходных данных с помощью единого интуитивного параметра (температуры).

REFERENCES

A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications (NIST SP 800-22) (2010). National Institute of Standards and Technology. — Gaithersburg. Pp.131. https://csrc.nist.gov/projects/random-bit-generation/documentation-and-software (in Eng.)

Agarwal S., Zhang Z., Yuan L., Han J., Peng H. (2025) The Unreasonable Effectiveness of Entropy Minimization in LLM Reasoning. — arXiv.2525.15134v1. — May 2025. DOI:10.48550/arXiv.2507.10532 (in Eng.)

Brown D. (2025). Top 10 Real-Life Applications of Large Language Models. — PixelPlex Blog. https://vegavid.com/blog/top-10-real-life-applications-of-large-language-models-reshaping-our-world (in Eng.)

Chang Y. A., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X,m Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P., Yang Q., Xie X. (2024). Survey on Evaluation of Large Language Models. — ACM Computing Surveys. — Vol. 15. — No. 3. DOI:10.1145/3641289 (in Eng.)

Charu C. (2023) Neural Networks and Deep Learning. Springer International Publishing AG, part of Springer Nature. Pp. 529. DOI 10.1007/978-3-031-29642-0 (in Eng.)

Grados Matej. (2015). Assessing Cryptographic Random Number Generators using NIST SP 80022 Statistical Test Suite. Master's Thesis, Tampere University. Pp. 75. https://trepo.tuni.fi/bitstream/handle/10024/229045/GradosMatej.pdf?sequence=2&isAllowed=y (in Eng.)

Hadi U.M., Tashi al Q., Shah A., Muneer A., Irfan M., Zafar A., Shaikh B.M., Akhtar N., Hassan Z.S., Shoman M., Wu J., Mirjalili S., Shah M. (2025). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. preprint, 2025. DOI: 10.36227/techrxiv.23589741.v8 — TechRxiv preprint. https://www.techrxiv.org/users/618307/articles/682263- (in Eng.)

Javier C. (2025). Deterministic or Probabilistic? The Psychology of LLMs as random number generators. AlphaXiv preprint. DOI :10.48550/arXiv.2502.19965 (in Eng.)

Kiyoshiro, O., Tndo R., Yasuoka K., Kurabayashi S. (2023). Learned pseudo-random number generator: WGAN-GP for generating statistically robust random numbers – PubMed. DOI: 10.1371/journal.pone.0287025 (in Eng.)

Minaee S., Mikolov T., Nikzad N., Chenaghli M., Socher R., Amatriain X., Gao J. (2024). Large Language Models: A Survey. - arXiv preprint arXiv:2402.06196. DOI:10.48550/arXiv.2402.06196 (in Eng.)

Menezes, A.J., van Oorschot, P.C., & Vanstone, S.A. (2025) Handbook of Applied Cryptography. — CRC Press, — 2025. https://sites.google.com/view/textbook-appliedcrypto/home (in Eng.)

NIST. (2018). NIST SP 800-90B: Recommendation for the Entropy Sources Used for Random Bit Generation. – Gaithersburg. Pp.76. DOI:10.6028./NIST.SP.800-90B (in Eng.)

Saleh Y., Abu Talib M., Nasir Q., Dakalbab F. (2025). Evaluating large language models: a systematic review of efficiency, applications, and future directions. - Front. Comput. Sci. 7:1523699. doi: 10.3389/fcomp.2025.1523699 (in Eng.)

Shimao Z., Bao Y., Huang S. (2024). EDT: Improving Large Language Models' Generation by Entropy-based Dynamic Temperature Sampling. — arXiv:2403.14541v. DOI:10.48550/arXiv.2403.14541 (in Eng.)

Vaswani, A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez N.A., Kaiser L., Polosukhin I. (2017). Attention Is All You Need. - Advances in Neural Information Processing Systems. DOI:10.48550/arXiv.1706.03762 (in Eng.)

Weihua D., Yiming Y., Sean W. (2025). Optimizing Temperature for Language Models with Multi-Sample

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

Inference., preprint. DOI: 10.48550/arXiv.2502.05234 https://arxiv.org/html/2502.05234v2 (in Eng.)

Yasmeen, S., Talib a.M., Nasir mh Q., Dakalbab F. (2025). Evaluating large language models: a systematic review of efficiency, applications, and future directions. Frontiers in Computer Science. DOI: 10.3389/fcomp.2025.1523699 https://doaj.org/article/5c9134037ab242d5a8451db826badb65 (in Eng.)

Zhang S., Bao Y., Huang, S. (2024). Improving Large Language Models' Generation by Entropybased Dynamic Temperature Sampling. — arXiv preprint arXiv:2403.14541. DOI:10.48550/arXiv.2403.14541 (in Eng.)

Weihua, D., Yang Y., Welleck S. (2025). Optimizing Temperature for Language Models with Multi-Sample Inference. — arXiv:2502.05234v2. DOI:10.48550/arXiv.2502.05234 (in Eng.)

White E. (2025). Exploring the Impact of Temperature on Large Language Models: Hot or Cold. — arXiv preprint arXiv:2506.07295. DOI:10.48550/arXiv.2506.07295 (in Eng.)

Пустовой Е. (2025). [Исходный код для облачного хранилища. https://drive.google.com/file/d/1hWuhvKeyrmm_u3TF6m6oK-cME-c0Jdsr/view?usp=sharing

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. 484–495
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.030

FRAUD DETECTION IN CREDIT CARD TRANSACTIONS USING MACHINE LEARNING: A COMPARATIVE ANALYSIS

Y. Serzhan[1]*, T. Umarov[1], A. Abilbayeva[2]
[1]Kazakh-British Technical University, Almaty, Kazakhstan;
[2]University of Maryland, College Park, MD, USA.
E-mail: y_serzhan@kbtu.kz

Serzhan Yerassyl — Master's student in Software Engineering at the School of Information Technology and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan
E-mail: y_serzhan@kbtu.kz, https://orcid.org/0009-0006-0615-6032;
Timur Umarov — PhD in Computing, Professor, School of Information Technology and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan
https://orcid.org/0009-0008-0044-7159;
Ainur Abilbayeva — Master of Science in Civil Information Systems, Department of Civil and Environmental Engineering, University of Maryland, College Park, MD, USA
https://orcid.org/0009-0004-3882-1442.

© Y.S. Serzhan, T.F. Umarov, A.B. Abilbayeva

Abstract. The accelerated development in digital financial transactions has enlarged the risk of fraudulent activities, which increased the need for more efficient fraud detection methods. This research study introduces a ML based approach to detect fraud in credit card transactions. Paper consists of sections, which summarize detailed exploratory data analysis, data preprocessing, model training, hyperparameter optimization, and discussion of the results. Various model trainings on a large credit card dataset demonstrated that the best model achieved an AUC of 96.82 % and an average precision of 88 %. Considering the data-driven approach applied in the study, this framework can be dynamically altered to detect new emerging fraud patterns. Additionally, the capability of the model to handle large volumes of data in real-time makes it well-suited for financial institutions with high transaction loads.

Keywords: machine learning, random rorest, logistic regression, fraud, hyperparameter optimization, XGBoost, gradient boosting

For citation: Y. Serzhan, T. Umarov, A. Abilbayeva (2026). Fraud detection in credit card transactions using machine learning: a comparative analysis // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 484–495. https://doi.org/10.54309/IJICT.2026.26.2.030. (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.

МАШИНАЛЫҚ ОҚУ ӘДІСІ АРҚЫЛЫ КРЕДИТ КАРТА ОПЕРАЦИЯЛАРЫНДАҒЫ АЛАЯҚТЫҚТЫ АНЫҚТАУ: САЛЫСТЫРМАЛЫ ТАЛДАУ

Е. Сержан[1]*, Т. Умаров[1], А. Әбілбаева[2]
[1]Қазақ-Британ техникалық университеті, Алматы, Қазақстан;
[2]Мэриленд Университеті, Колледж Парк, МД, АҚШ.
E-mail: y_serzhan@kbtu.kz

Сержан Ерасыл — Ақпараттық технологиялар және инжиниринг мектебінің Бағдарламалық қамтамасыз ету инженериясы бағытының магистранты, Қазақ-Британ техникалық университеті, Алматы, Қазақстан
E-mail: y_serzhan@kbtu.kz, https://orcid.org/0009-0006-0615-6032;
Умаров Тимур — Компьютер ғылымдарының докторы, Профессор, Ақпараттық Технологиялар Департаменті, Қазақстан-Британ Техникалық Университеті, Алматы, Қазақстан
https://orcid.org/0009-0008-0044-7159;
Әбілбаева Айнур — Азаматтық ақпарат жүйелерінің Ғылым Магистрі, Құрылыс және Экология Инженериясы Департаменті, Мэриленд Университеті, Колледж Парк, МД, АҚШ
https://orcid.org/0009-0004-3882-1442.

© Е. Сержан, Т. Умаров, А. Әбілбаева

Аннотация. Қаржылық транзакциялардың қарқынды дамуы алаяқтық әрекеттердің қаупін арттырып, алаяқтықты анықтаудың тиімді әдістеріне қажеттілікті арттырды. Бұл мақалада кредит картасы арқылы жасалған транзакциялардағы алаяқтықты анықтау үшін машиналық оқу әдістерін қолданатын жан-жақты зерттеу ұсынылған. Мақала мына бөлімдерден тұрады: деректерді зерттеу, деректерді алдын ала өңдеу, әртүрлі класс теңдестіру әдістерімен модельді оқыту және гиперпараметрлерді оңтайландыру. Кредит картасы арқылы жасалған төлемдерден құралған деректер жинағына жасалған эксперименттер ең жақсы модель 96,82 % AUC және 88 % орташа дәлдікке жеткенін көрсетті. Бұл зерттеу деректерге негізделгендіктен, күдікті әрекетті анықтауда жоғары дәлдікті қамтамасыз ете отырып, жаңа алаяқтық үлгілеріне динамикалық түрде бейімделе алады. Сонымен қатар, модельдің үлкен көлемдегі деректерді нақты уақыт режимінде өңдеу мүмкіндігі-оны транзакциялардың жоғары жүктемелерін басқаратын қаржы институттары үшін қолайлырақ етеді.

Түйінді сөздер: машиналық оқыту, кездейсоқ орман, логистикалық регрессия, алаяқтық, гиперпараметрлерді оңтайландыру, XGBoost, градиентті күшейту

Дәйексөздер үшін: Е. Сержан, Т. Умаров, А. Әбілбаева (2026). Машиналық оқу әдісі арқылы кредит карта операцияларындағы алаяқтықты анықтау: салыстыр-

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

малы талдау // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 26. Б. 484–495. https://doi.org/10.54309/IJICT.2026.26.2.030. (Ағыл. тіл.).

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

ВЫЯВЛЕНИЕ МОШЕННИЧЕСТВА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ ПРИ ОПЕРАЦИЯХ С КРЕДИТНЫМИ КАРТАМИ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Е. Сержан[1]*, Т. Умаров[1], А. Абильбаева[2]

[1]Казахстанско-Британский технический университет, Алматы, Казахстан;
[2]Университет Мэриленда, Колледж Парк, МД, США.
E-mail: y_serzhan@kbtu.kz

Сержан Ерасыл — магистрант направления Программная инженерия Школы информационных технологий и инженерии Казахстанско-Британского технического университета, Алматы, Казахстан
E-mail: y_serzhan@kbtu.kz, https://orcid.org/0009-0006-0615-6032;
Умаров Тимур — Доктор компьютерных Наук, Профессор, Факультет Информационных Технологий, Казахстанско-Британский технический университет, Алматы, Казахстан
https://orcid.org/0009-0008-0044-7159;
Абильбаева Айнур — Магистр Наук гражданских информационных, Департамент Строительной и Экологической Инженерии, Университет Мэриленда, Колледж Парк, МД, США
https://orcid.org/0009-0004-3882-1442.

© Е. Сержан, Т. Умаров, А. Абильбаева

Аннотация. Быстрое развитие цифровых финансовых транзакций увеличило риск мошеннических действий, что увеличило потребность в более эффективных методах обнаружения мошенничества. В данной статье представлено комплексное исследование, которое использует методы машинного оьучения для выявления мошенничества в транзакциях по кредитным картам. Статья состоит из разделов: эксплораторный анализ данных, предварительная обработка данных, обучение модели с различными методами балансировки классов и оптимизация гиперпараметров. Эксперименты на большом наборе данных кредитных карт показали, что лучшая модель достигла AUC 96,82 % и средней точности 88 %. Поскольку структура обучается с использованием подхода, управляемого данными, она может динамически адаптироваться к новым схемам мошенничества, обеспечивая высокую точность выявления подозрительных действий. Кроме того, способность модели обрабатывать большие объемы данных в режиме реального времени делает ее более подходящей

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

для финансовых учреждений, управляющих высокими транзакционными нагрузками.

Ключевые слова: случайный лес, Логистическая регрессия, Мошенничество, Оптимизация гиперпараметров, XGBoost, градиентный бустинг

Для цитирования: Е. Сержан, Т. Умаров, А. Абильбаева (2026). Выявление мошенничества с использованием машинного обучения при операциях с кредитными картами: сравнительный анализ // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 484–495. https://doi.org/10.54309/IJICT.2026.26.2.030. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

With the rise of digital banking and online transactions, credit card fraud has become a growing concern for both financial institutions and customers. Although fraudulent transactions make less than 1 % of all transactions, they can lead to billions of dollars in financial losses each year. Disputed transactions also undermine trust for online payment systems, which make customers hesitant and reluctant to use them. Traditional rule-based fraud detection methods, which rely on predefined patterns and manual oversight, are becoming less effective as malicious actors unremittingly develop more sophisticated methods. While recent progress in machine learning offers promising tools with automated pattern recognition and anomaly detection, there is a significant demand for more sophisticated and adaptable approaches to fraud prevention.

Credit card fraud has a massive economic impact. According to the Nilson Report global losses due to fraud activities estimated at $33.83 billion in 2023 (Nilson Report, 2023). Apart from identifying deceptive transactions, the major challenge for financial institutions is to minimize financial losses while avoiding unnecessary disturbances to legitimate customers. A system which responds too many false positives can frustrate users, leading to declined transactions and lost revenue, while a system that misses fraudulent transactions can cause serious financial damage. Highly imbalanced nature of transactional records is the major challenge, because there might be only a few fraudulent transactions. For this reason, it is highly intricate to detect fraud effectively with traditional models without false alarms. This research study focuses on the development and evaluation of machine learning models, specifically designed to handle imbalanced datasets. The primary goal is to develop a more robust fraud detection framework which minimizes any negative impact on the overall customer experience. This paper investigates the effectiveness of machine learning techniques for credit card fraud detection, addressing key challenges such as class imbalance and model performance optimization. The study aims to evaluate different fraud detection approaches and highlight their practical applications in real-world financial systems. The research emphasizes the balance between fraud detection accuracy and minimizing false positives to ensure a seamless customer experience while maintaining security.

Literature Review

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

The majority of the research conducted in the field of fraud detection substantially addresses Random Forest as the most accurate machine learning model to use. In general, existing approaches can be categorized into several main groups: deep learning, machine learning, ensemble learning and feature ranking methods, and user authentication strategies (Alarfaj et al., 2022). Figure 1 illustrates the commonly used payment card authorization process for credit card authentication. Generally, authentication methods fall into two main categories: password-based authentication and biometric authentication.

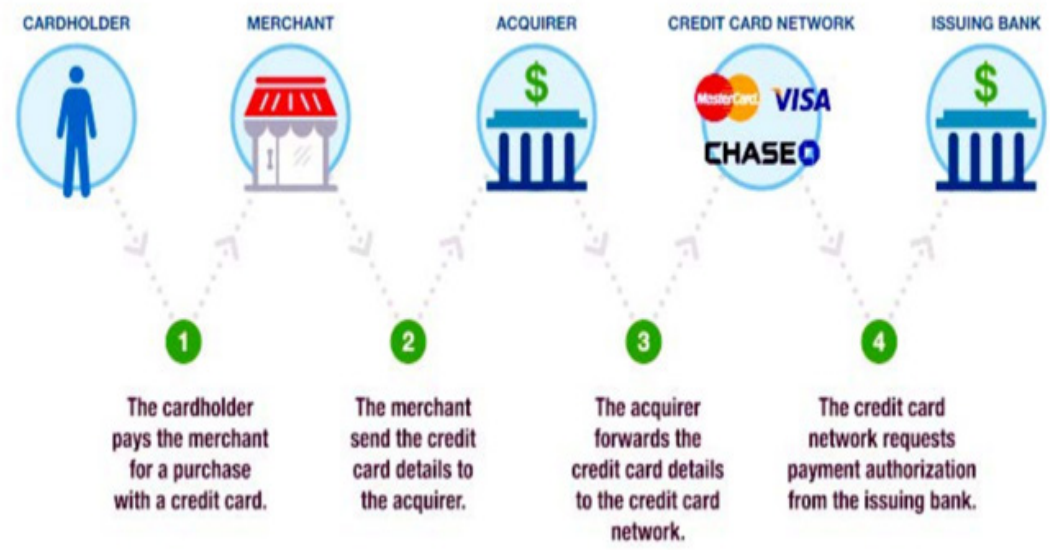


Fig. 1. Payment card authorisation process (Alarfaj et al., 2022).

West and Bhattacharya (2016) highlight that data mining techniques can be time-consuming when working with large datasets. Additionally, as noted in papers written by (Zareapoor et al., 2012; Zojaji et al., 2016) highly correlated information is an indicative challenge in preparing credit card transaction data. Consequently, data might be misclassified for the reason that fraud ones resemble legitimate transactions. Another obstacle encountered by researches is handling categorical features. West and Bhattacharya (2016) and Zareapoor et al. (2012) debate on the challenges related to choosing appropriate detection algorithms and selecting relevant features. Predominant machine learning models require lots of time for training. For this reason, it is important to approach feature selection in detail in order to identify the most relevant attributes that characterize fraudulent behavior.

A hybrid approach, which combines Random Forest and Isolation Forest to identify anomalous transactions has been introduced by (Vynokurova et al., 2020). Their proposed model consists of two main components: using an unsupervised and supervised learning approaches. According to (Sulaiman et al., 2022), albeit Random Forest algorithms are highly effective for classification tasks, they have several limitations when applied to real-time credit card fraud detection. The training process for this algorithm is slower compared

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

to other techniques, which makes it inapplicable for real-time fraud detection.

A comprehensive framework has been proposed by (Hashemi et al., 2022), which integrates hyperparameter tuning with ensemble methods to manipulate class imbalance. Their results indicate that deep learning algorithms have shown considerable prospect in enhancing credit card fraud detection. Model proposed in this study exploits an artificial neural network (ANN) with federated learning, which preserves data confidentiality. Notwithstanding previous research on the use of ANN for fraud detection, they have been directed on lab-based datasets. While model by (Hashemi et al., 2022) consolidates fraud detection efficiency and reliability.

Study done by Rong Zhang et al. (2023) analyzed various classifiers using Support Vector Machine (SVM) method. Their model trained on the BankSim dataset resulted in 99.23% accuracy. Although this experiment highlights prospects of SVM model, it has been trained on the well-structured transaction data. For this reason, it might not be suitable for detecting fraud in the real-world transaction data.

Further research in the sphere of fraud detection accentuates use of deep learning and neural network architectures in particular (Thennakoon, 2019). Their flexibility allows to perform more efficiently for anomaly detection tasks. Nevertheless, challenges such as excessive computational demands and training periods present significant challenge for practical deployment.

Materials and methods.

Figure 2 illustrates the data preprocessing and model training workflow in this research study. The following sections will explain each of the steps in detail.

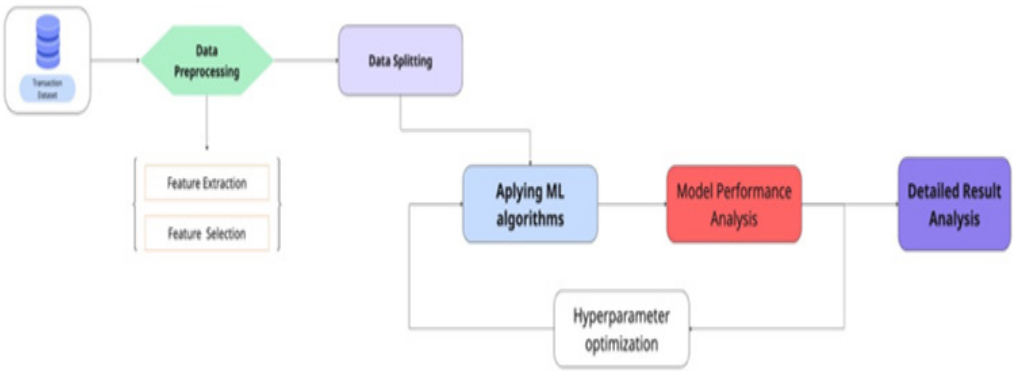


Fig. 2. Model Training Workflow

Data Description and Preprocessing

Most of the real-world datasets come with lots of missing data, but in this case a well-organized and complete dataset from the Kaggle website is analyzed. The major two features «Time» and «Amount,» remain in their original forms, while the remaining features are principal components derived from Principal Component Analysis. As typical in fraud detection, the dataset is highly imbalanced, with fraudulent transactions representing only

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

0.1727 % of the total (only 492 fraudulent out of 284,807), which presents an indicative modelling challenge for the study. The dataset's structure, feature distributions, and the extent of class imbalance was inspected using descriptive statistics and visualizations. The major key insight revealed from this analysis is that most of the transactions were for relatively small value, which has been validated by median of $22.00. In general, the range of the transaction amounts were from $0 to $25,691.16, with a mean of $88.35.

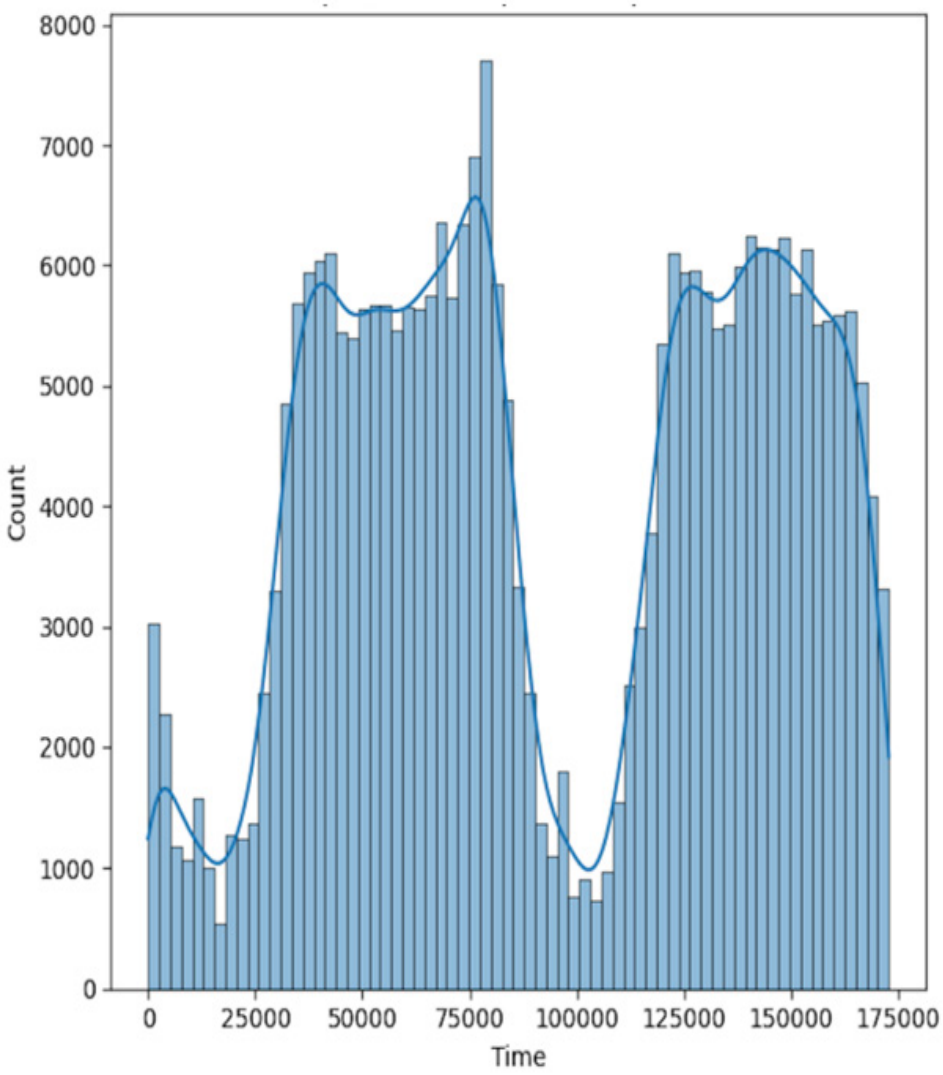


Fig. 3. Transaction distribution over time

Figure 3 shows overall distribution of transactions over time, and it can be noticed that most of the transactions occur around lunch time (between 11 am and 4 pm) and after 6 pm. Figure 4 below illustrates comprehensive correlation matrix of features in the dataset. All V1-V28 features (PCA components) follow different distributions with

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

varying ranges.

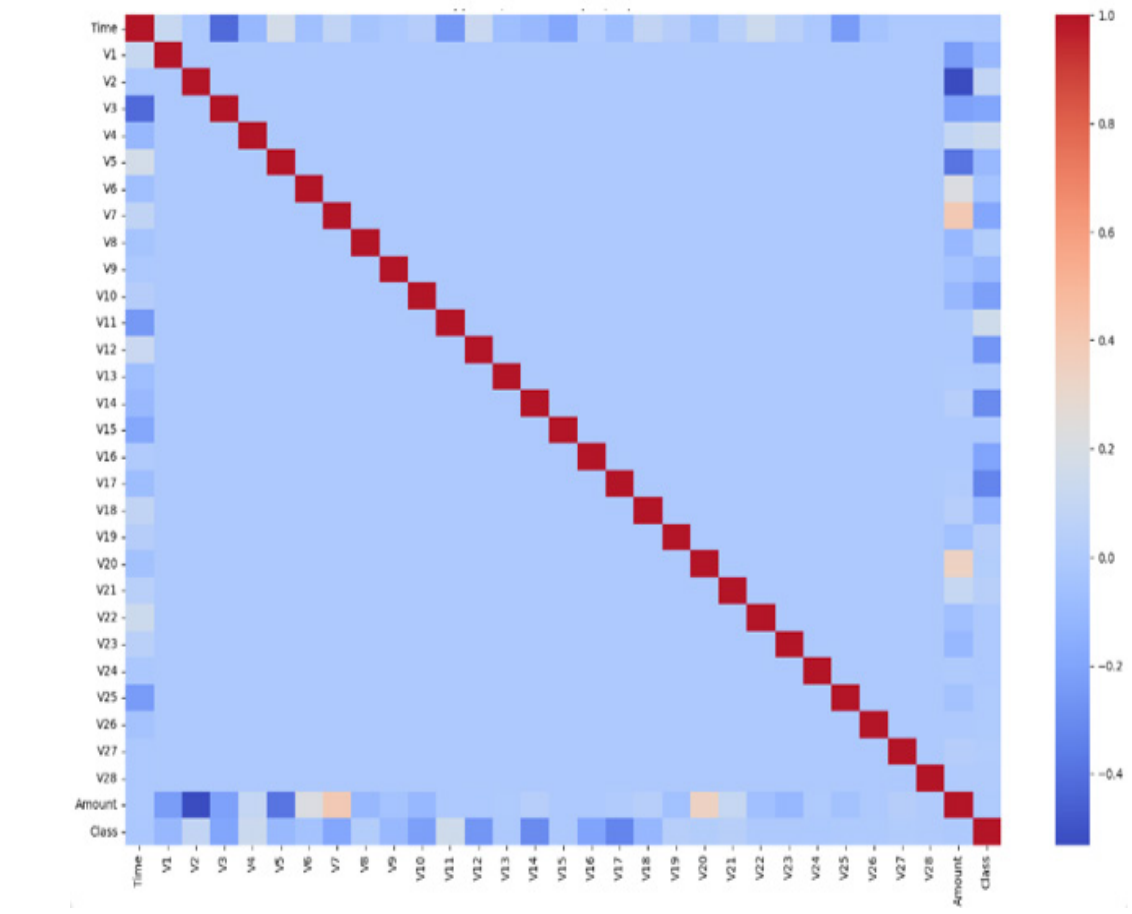


Fig. 4. Correlation matrix of features

Model Training and Class Balancing

In this study four machine learning classifiers have been trained: Logistic Regression, Random Forest, Gradient Boosting and XGBoost. For each classifier, three class balancing methods were applied:

1. No Sampling: Training on the original imbalanced dataset.

2. SMOTE: Synthetic Minority Over-sampling Technique to balance the classes by creating synthetic examples of the minority class, resulting in a balanced training set of 454,902 examples.

3. Undersampling: Reducing the majority class to achieve balance (394 examples for each class), resulting in a much smaller training set of 788 examples.

Before model training data was split into training (80 %) and test (20 %) sets using stratified sampling to preserve class distribution, resulting in 227,845 training examples and 56,962 test examples.

Hyperparameter Optimization

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

For the best-performing model (XGBoost without sampling) (figures 5 and 6), hyperparameter optimization was conducted using grid search with stratified 5-fold cross-validation. The grid search explored 2,592 parameter combinations, with a total of 12,960 models trained during this optimization process.

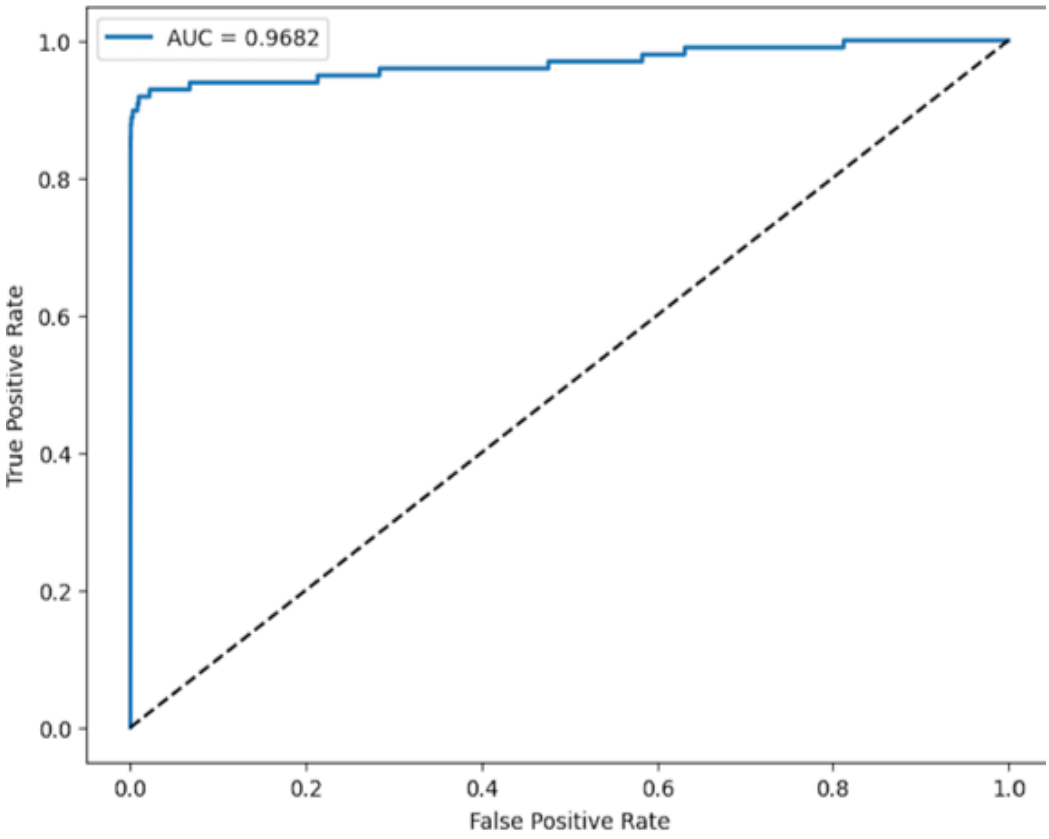


Fig. 5. Receiver Operating Characteristic curve for XGBoost without sampling

Results and discussion.

The best performance was achieved by the XGBoost model without sampling (Figure 6), which reached an AUC of 0.9682 and an average precision of 0.8800. This model also demonstrated excellent computational efficiency, requiring only 7.24 seconds for training.

Detailed Performance Analysis

Several distinctive patterns have been revealed throughout this study. Table 1 summarizes how each model performed using different sampling techniques and quantifies the results. In most of the cases, SMOTE performed better while comparing AUC, however overall precision value decreased and in a similar way undersampling of model enhanced model recall but decreased precision. The best balance of metrics has been obtained by training the initial imbalanced dataset on XGBoost. Correspondingly, the training time for model with XGBoost required 7.24 seconds, making it the most suitable for detecting fraud

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

transactions in real-time.

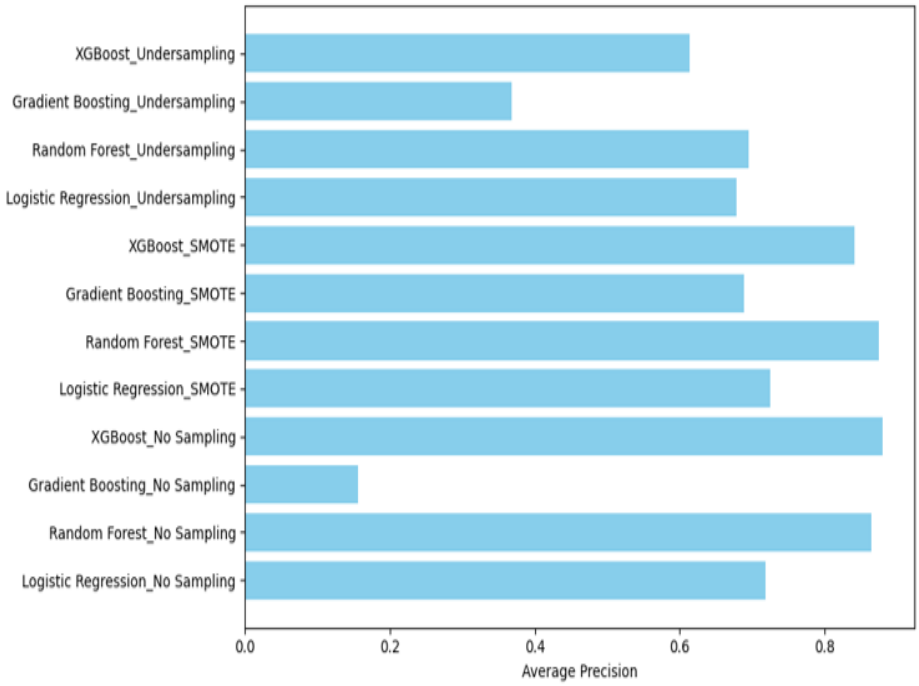


Fig. 6. Models' comparison on average precision

Table 1 – Detailed performance analysis

Model	Sampling	AUC	Avg. Precision	Precision	Recall	F1-Score	Training Time (s)
Logistic Regression	None	0.9722	0.7189	0.06	0.92	0.11	1.67
Logistic Regression	SMOTE	0.9698	0.7249	0.06	0.92	0.11	5.98
Logistic Regression	Under	0.9760	0.6778	0.04	0.92	0.07	0.14
Random Forest	None	0.9530	0.8650	0.96	0.74	0.84	225.39
Random Forest	SMOTE	0.9683	0.8748	0.85	0.83	0.84	646.07
Random Forest	Under	0.9783	0.6958	0.05	0.92	0.09	0.73
Gradient Boosting	None	0.3469	0.1567	0.53	0.18	0.27	659.91
Gradient Boosting	SMOTE	0.9807	0.6885	0.11	0.90	0.19	1359.22
Gradient Boosting	Under	0.9770	0.3687	0.03	0.93	0.07	1.28
XGBoost	None	0.9682	0.8800	0.88	0.84	0.86	7.24
XGBoost	SMOTE	0.9777	0.8418	0.46	0.86	0.60	11.05
XGBoost	Under	0.9798	0.6142	0.02	0.92	0.03	0.27

Comparison with Hashemi et al. (2022)

Hashemi et al. (2022) developed an ensemble framework that leverages Bayesian optimization to tune hyperparameters and mitigate class imbalance. Their method achieved ROC-AUC values around 0.95.

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

Table 2 – Comparison with Hashemi et al. (2022)

Aspect	Hashemi et al.	Our Approach
Dataset Size	284,807 transactions	284,807 transactions
Best Model	Ensemble (CatBoost, XGBoost, LightGBM)	XGBoost without sampling
AUC	0.9512	0.9682
Average Precision	0.9043	0.8800
Training Time	176.5 seconds	7.24 seconds

The approach studied in this paper yields a higher ROC-AUC (0.9682) compared to 0.9512, although with a slightly lower average precision. The most significant advantage of the method is computational efficiency, with our model training more than 24 times faster than their ensemble approach (Table 2).

Comparison with Zhang et al. (2023)

Zhang et al. (2023) evaluated multiple classifiers on a synthetic BankSim dataset, reporting that SVM achieved the highest accuracy (99.23 %).

Table 3 – Comparison with Zhang et al. (2023)

Aspect	Rong Zhang et al.	Our Approach
Dataset	Synthetic BankSim	Real credit card transactions
Features	Age, gender, payment domain	PCA-transformed features
Best Model	SVM	XGBoost without sampling
Accuracy	99.23%	99.98%
F1-Score	0.79	0.86

Our XGBoost model achieves both higher accuracy and F1-score compared to their SVM approach. Additionally, our evaluation on real-world data rather than synthetic data provides stronger evidence for practical applicability (Table 3).

Comparison with Thennakoon et al. (2019)

The study done by Thennakoon et al. (2019) on real-time fraud detection using deep learning and ensemble methods emphasizes the advantages of deep learning architectures for capturing complex patterns.

Table 4 – Comparison with Thennakoon et al. (2019).

Aspect	IEEE Study	Our Approach
Methodology	Deep Learning (LSTM, GRU)	Traditional ML (XGBoost)
AUC	0.9526	0.9682
Inference Time	124ms per transaction	≈2ms per transaction (estimated)
Model Size	248MB	≈12MB (estimated)

Although our approach does not integrate deep learning, the robust performance of our XGBoost model demonstrates that traditional machine learning methods, when properly tuned, can achieve competitive results in fraud detection (Table 4). This reinforces the notion that model selection and hyperparameter tuning are critical to performance, even without the computational complexity of deep neural networks. Furthermore, the model offers

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

significant advantages in terms of inference time and model size, making it more suitable for deployment in resource-constrained environments.

Conclusion.

To summarize, the aim of the research project was to analyze different machine learning techniques for bettere performance in credit card fraud detection. The thorough literature review has been done on ML techniques and class imbalance methods to examine the cost-effectiveness and to validate the experimental tests. The XGBoost model without sampling has proven to be computationally efficient and accurate model for detection in real-time compared to other models.

Further recommendations to achieve better results include an exhaustive investigation on the hybrid approaches as combining ML models with deep learning components. While sampling techniques like SMOTE and undersampling are often recommended for imbalanced datasets, the results indicate that they should be applied solicitously. Apart from this, the research study has proven that our model effectively balances high detection accuracy with computational efficiency, making it a strong candidate for real-world fraud detection systems.

REFERENCES

Alarfaj F.K., Malik I., Khan H.U., Almusallam N., Ramzan M., Ahmed M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access. — Vol. 10. — Pp. 39700–39715. DOI: https://doi.org/10.1109/ACCESS.2022.3166891 [in Eng].

Hashemi S.K., Mirtaheri S.L., Greco S. (2022). Fraud detection in banking data by machine learning techniques. IEEE Access. — Vol. 11. — Pp. 3034–3043. DOI: https://doi.org/10.1109/ACCESS.2022.3232287 [in Eng].

Machine Learning Group–ULB. Credit Card Fraud Detection. Kaggle, 2018. [Online]. Available: https://www.kaggle.com/mlgulb/creditcardfraud. [Accessed: 12.12.2024]. [in Eng].

Nilson Report. Card fraud losses worldwide in 2023. [Online]. Available: https://nilsonreport.com/articles/card-fraud-losses-worldwide-in-2023/. [Accessed: 21.02.2025]. [in Eng].

Sulaiman R. Bin, Schetinin V., Sant P. (2022). Review of machine learning approach on credit card fraud detection // Human-Centric Intelligent Systems. — Vol. 2. — No. 1. Pp. 55–68. DOI: https://doi.org/10.1007/s44230-022-0000 [in Eng].

Thennakoon A., Bhagyani C., Premadasa S., Mihiranga S., Kuruwitaarachchi N. (2019). Real-time credit card fraud detection using machine learning. In: 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE. Pp. 488–493. DOI: https://doi.org/10.1109/CONFLUENCE.2019.8776942 [in Eng].

Vynokurova O., Peleshko D., Bondarenko O., Ilyasov V., Serzhantov V., Peleshko M. (2020). Hybrid machine learning system for solving fraud detection tasks // 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP). IEEE, 2020. Pp. 1–5. DOI: https://doi.org/10.1109/DSMP47368.2020.9204244. [in Eng].

West J., Bhattacharya M. (2016). An investigation on experimental issues in financial fraud mining. In: 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). IEEE. Pp. 1796–1801. DOI: https://doi.org/10.1109/ICIEA.2016.7603878. [in Eng].

Zareapoor M., Seeja K.R., Alam M.A. (2012). Analysis on credit card fraud detection techniques: based on certain design criteria // International Journal of Computer Applications. 2012. — Vol. 52. — No. 3. DOI: https://doi.org/10.5120/8184-1538. [in Eng].

Zhang R., Cheng Y., Wang L., Sang N., Xu J. (2023). Efficient Bank Fraud Detection with Machine Learning // Journal of Computational Methods in Engineering Applications. Pp. 1–10. DOI: https://doi.org/10.62836/jcmea.v3i1.030102 [in Eng].

Zojaji Z., Atani R.E., Monadjemi A.H. (2016). A survey of credit card fraud detection techniques: data and technique oriented perspective. arXiv preprint arXiv:1611.06439. DOI: https://doi.org/10.48550/arXiv.1611.06439 [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.2. Number 26 (2026). Pp. 496–511
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.26.2.031
УДК 004.056.5

FLEXIBLE IDPS BASED ON MACHINE LEARNING ALGORITHMS FOR DIPTYCHS AND IOT EDGE DEVICE PROTECTION

T. Babenko, Ye. Bakhtiyarova*, D. Yeskendirova, A. Sysoev, Sh. Makilenov
International Information Technology University, Almaty, Kazakhstan.
E-mail: y.bakhtiyarova@iitu.edu.kz

Tatiana Babenko — Doctor of Technical Sciences, Professor, Department of Cybersecurity, International University of Information Technologies, Almaty, Kazakhstan
https://orcid.org/0000-0003-1184-9483;
Yelena Bakhtiyarova — Candidate of Technical Sciences, Associate Professor, Head of the Department of Radio Engineering, Electronics, and Telecommunications, International University of Information Technologies, Almaty, Kazakhstan
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;
Damelya Yeskendirova — Candidate of Technical Sciences, Head of the Department of Cybersecurity, International University of Information Technologies, Almaty, Kazakhstan
https://orcid.org/0000-0003-4270-1908;
Askar Sysoyev — PhD, Senior Lecturer, Department of Cybersecurity, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/ 0009-0002-5842-1847;
Shakirt Makilenov — doctoral student at the International University of Information Technology, Almaty, Kazakhstan
https://orcid.org/0000-0002-5330-4716.

© T. Babenko, Ye. Bakhtiyarova, D. Yeskendirova, A. Sysoyev, Sh. Makilenov

Abstract. The growing number of edge IoT devices in smart city infrastructures poses new challenges related to a fragmented attack surface and the limitations of traditional intrusion detection tools. Existing IDS systems typically operate in isolation, generate redundant alerts, and are unable to adapt to distributed environments. This paper proposes the architecture of a flexible IDS that leverages machine learning algorithms to improve the monitoring and protection of edge devices. The architecture includes modules for normalization, deduplication, correlation, and ML analysis, with the ability to provide feedback between the monitoring center and edge nodes. The architectural description of the monitoring environment is based on a hierarchical representation consistent with the Purdue model, allowing for the distinction between edge, local, and centralized levels of

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

protection. Experimental evaluation on the BoT-IoT and TON_IoT datasets demonstrated a reduction in event correlation latency by approximately 62% and a decrease in the false positive rate by 29–31% compared to baseline approaches. The obtained results confirm the potential of the proposed architecture for smart city infrastructures.

Keywords: intrusion detection and prevention system, Internet of Things, machine learning, smart city, edge devices, alert normalization, event correlation, edge computing, Purdue model

For citation: T. Babenko, Ye. Bakhtiyarova, D. Yeskendirova, A. Sysoyev, Sh. Makilenov (2026). Flexible idps based on machine learning algorithms for diptychs and iot edge device protection // International journal of information and communication technologies. Vol. 7. No. 26. Pp. 496–511. https://doi.org/10.54309/IJICT.2026.26.2.031. (In Russ.).

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

Acknowledgment: *This study was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under IRN AP26104787, «Development of AI-based Solutions for Protecting the IoT Infrastructure of Smart Cities in Kazakhstan.»*

ДИПТИХ ЖӘНЕ IOT ШЕТКІ ҚҰРЫЛҒЫЛАРЫН ҚОРҒАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІНЕ НЕГІЗДЕЛГЕН ИКЕМДІ IDPS

Т.В. Бабенко, Е.А. Бахтиярова*, Д.М. Ескендирова, А. Сысоев, Ш.Н. Макиленов

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

Бабенко Татьяна Василевна — т.ғ.д., «Киберқауіпсіздік» кафедрасының профессоры Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-1184-9483;
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;
Ескендирова Дамеля Максутовна — т.ғ.к, «Киберқауіпсіздік» кафедрасының меңгерушісі Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-4270-1908;
Сысоев Аскар — PhD, киберқауіпсіздік кафедрасының аға оқытушысы, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/ 0009-0002-5842-1847;
Макиленов Шакирт Нурлубекулы — докторант, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-5330-4716.

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

© Т.В. Бабенко, Е.А. Бахтиярова, Д.М. Ескендирова, А. Сысоев, Ш.Н. Макиленов

Аннотация. Ақылды қала инфрақұрылымдарындағы шеткі IoT құрылғыларының санының артуы фрагменттелген шабуыл бетіне және дәстүрлі басып кіруді анықтау құралдарының шектеулеріне байланысты жаңа қиындықтар туғызады. Қолданыстағы IDS жүйелері әдетте оқшауланған күйде жұмыс істейді, артық ескертулер жасайды және таратылған ортаға бейімделе алмайды. Бұл мақалада шеткі құрылғыларды бақылау мен қорғауды жақсарту үшін машиналық оқыту алгоритмдерін пайдаланатын икемді IDS архитектурасы ұсынылады. Архитектураға бақылау орталығы мен шеткі түйіндер арасында кері байланыс беру мүмкіндігімен қалыпқа келтіру, қайталауды алып тастау, корреляция және ML талдау модульдері кіреді. Мониторинг ортасының архитектуралық сипаттамасы Purdue моделіне сәйкес иерархиялық көрініске негізделген, бұл шеткі, жергілікті және орталықтандырылған қорғау деңгейлерін ажыратуға мүмкіндік береді. BoT-IoT және TON_IoT деректер жиынтықтары бойынша эксперименттік бағалау бастапқы тәсілдермен салыстырғанда оқиға корреляциясының кідірісінің шамамен 62%-ға және жалған оң нәтижелер көрсеткішінің 29-31%-ға төмендегенін көрсетті. Алынған нәтижелер ақылды қала инфрақұрылымдары үшін ұсынылған архитектураның әлеуетін растайды.

Түйінді сөздер: енуді анықтау және алдын алу жүйесі, Заттар интернеті, машиналық оқыту, ақылды қала, шеткі құрылғылар, дабылдарды қалыпқа келтіру, оқиғалар корреляциясы, шеткі есептеулер, Purdue моделі

Дәйексөздер үшін: Т.В. Бабенко, Е.А. Бахтиярова, Д.М. Ескендирова, А. Сысоев, Ш.Н. Макиленов (2026). Диптих және iot шеткі құрылғыларын қорғауға арналған машиналық оқыту алгоритмдеріне негізделген икемді IDPS // Халықаралық ақпараттық және коммуникалық технологиялар журналы. Т. 7. No. 26. Б. 496–511. https://doi.org/10.54309/IJICT.2026.26.2.031. (Орыс. тіл.).

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

Алғыс: *Бұл мақала Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетінің қаржылық қолдауымен, AP26104787 «Жасанды интеллект негізінде Қазақстанның ақылды қалаларының IoT-инфрақұрылымын қорғау шешімдерін әзірлеу» атты жобасы аясында орындалды.*

ГИБКАЯ IDPS НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ МОНИТОРИНГА И ЗАЩИТЫ ПЕРИФЕРИЙНЫХ IoT-УСТРОЙСТВ

Т. Бабенко, Е. Бахтиярова*, Д. Ескендирова, А. Сысоев, Ш. Макиленов
Международный университет информационных технологий, Алматы, Казахстан.
E-mail: y.bakhtiyarova@iitu.edu.kz

Бабенко Татьяна Василевна — д.т.н., профессор кафедры «Кибербезопасность», Международный университет информационных технологий, Алматы, Казахстан

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

https://orcid.org/0000-0003-1184-9483;
Бахтиярова Елена Ажибековна — к.т.н., ассоциированный профессор, заведующий 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;
Ескендирова Дамеля Максутовна — к.т.н., заведующий кафедрой «Кибербезопасность», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0003-4270-1908;
Аскар Сысоев — PhD, старший преподаватель кафедры кибербезопасности, Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/ 0009-0002-5842-1847;
Макиленов Шакирт Нурлубекулы — докторант, Международного университета информационных технологий, Алматы, Казахстан
https://orcid.org/ 0000-0002-5330-4716.

© Т. Бабенко, Е. Бахтиярова, Д. Ескендирова, А. Сысоев, Ш. Макиленов

Аннотация. Рост числа периферийных IoT-устройств в инфраструктуре умных городов порождает новые вызовы, связанные с фрагментированной поверхностью атак и ограниченностью традиционных средств обнаружения вторжений. Существующие системы IDPS, как правило, работают изолированно, генерируют избыточные алерты и не способны адаптироваться к условиям распределённой среды. В данной статье предлагается архитектура гибкой IDPS, использующей алгоритмы машинного обучения для повышения качества мониторинга и защиты периферийных устройств. Архитектура включает модули нормализации, дедупликации, корреляции и ML-анализа с возможностью обратной связи между центром мониторинга и edge-узлами. Архитектурное описание среды мониторинга опирается на иерархическое представление, согласующееся с моделью Purdue, что позволяет разграничить периферийный, локальный и централизованный уровни защиты. Экспериментальная оценка на датасетах BoT-IoT и TON_IoT показала сокращение задержки корреляции событий примерно на 62% и снижение доли ложных срабатываний на 29–31% по сравнению с базовыми подходами. Полученные результаты подтверждают перспективность предложенной архитектуры для инфраструктур умного города.

Ключевые слова: система обнаружения и предотвращения вторжений, Интернет вещей, машинное обучение, умный город, периферийные устройства, нормализация алертов, корреляция событий, edge-вычисления, модель Purdue

Для цитирования: Т. Бабенко, Е. Бахтиярова, Д. Ескендирова, К. Сансызбай, Ш. Макиленов (2026). Гибкая idps на основе алгоритмов машинного обучения для мониторинга и защиты периферийных iot-устройств // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 26. Стр. 496–511. https://doi.org/10.54309/IJICT.2026.26.2.031. (На русс.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

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

Благодарность: *Статья выполнена при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан в рамках проекта ИРН AP26104787 «Разработка решений для защиты IoT-инфраструктуры умных городов Казахстана на основе искусственного интеллекта».*

Введение.

Городская среда меняется быстрее, чем успевают перестраиваться модели её защиты. Умные города внедряют тысячи подключённых устройств, от датчиков качества воздуха до контроллеров критической инфраструктуры (Jo et al., 2019; Elvas et al., 2021). По оценкам, к 2030 году их число может достичь нескольких миллиардов (Du et al., 2019) и каждое из них выступает не только источником данных, но и потенциальной точкой проникновения для злоумышленника (Almeida, 2023). При этом значительная часть таких устройств располагается на периферии сети, то есть на тех уровнях, где вычислительные ресурсы ограничены, а каналы связи нестабильны (Shi et al., 2016).

Модели кибербезопасности, сложившиеся вокруг централизованных сетевых периметров, оказываются в этих условиях недостаточно эффективными. Центры мониторинга безопасности (SOC), исторически служившие ядром оперативного реагирования (Bhatt et al., 2014), рассчитаны на относительно стабильные топологии с хорошо определёнными границами (Kotenko et al., 2012). Сети умного города, однако, простираются на обширные территории и включают устройства различных производителей, работающие по разнородным протоколам (Sisinni et al., 2018). Атаки на IoT-устройства нередко развиваются стремительно, эксплуатируя уязвимости ресурсоограниченных узлов до того, как централизованные механизмы обнаружения смогут сопоставить разрозненные сигналы в целостную картину (Liu et al., 2020).

Системы обнаружения и предотвращения вторжений (IDPS) призваны мониторить сетевой трафик на предмет вредоносных паттернов (Zarpelão et al., 2017; Walling et al., 2025). Платформы вроде Suricata и Zeek обладают развитыми возможностями анализа пакетов и сигнатурного обнаружения (Paxson, 1998). Тем не менее, в условиях распределённой IoT-среды они страдают от фундаментального дефицита интеграции (Aktar et al., 2024). Большинство установок работают как изолированные сенсоры, генерирующие алерты в несогласованных форматах без координации с централизованной платформой оркестрации (La et al., 2021). Это создаёт информационные силосы, в которых фрагменты многостадийных атак теряют контекст (Wardana et al., 2024).

Здесь важно заметить, что сами по себе алгоритмы машинного обучения, при всей их перспективности для выявления аномалий, не решают проблему изолированности и негибкости IDPS. Применение ML в задачах обнаружения вторжений описано в ряде обзорных работ (Zarpelão et al., 2017; Walling et al., 2025; Mitchell et al., 2014), однако вопрос о том, как встроить такие модели в гибкую, модульную архитектуру, адаптированную к условиям периферийной IoT-среды, остаётся открытым. Стандарты обмена информацией об угрозах, такие как STIX и TAXII (Iqbal et

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

al., 2018; Sauerwein et al., 2017), предоставляют полезные строительные блоки, но их внедрение в реальных развёртываниях умных городов пока фрагментарно (Najafi et al., 2021).

Целью данной статьи является разработка архитектуры гибкой IDPS на основе алгоритмов машинного обучения для мониторинга и защиты периферийных IoT-устройств. Под гибкостью здесь понимается способность системы адаптироваться к различным типам устройств, модульно встраивать и заменять аналитические компоненты, а также обеспечивать двунаправленную связь между центром мониторинга и периферийными сенсорами. Научная новизна состоит в предложении архитектуры, которая объединяет нормализацию, дедупликацию, ML-анализ и приоритизацию событий в единый конвейер, адаптированный для периферийного уровня IoT-инфраструктуры. Практическая значимость определяется возможностью использования подхода при построении систем безопасности для национальной инфраструктуры умных городов.

Материалы и методы.

Анализ существующих подходов

Особенности защиты периферийных IoT-устройств

Периферийные IoT-устройства принципиально отличаются от традиционных конечных точек корпоративной сети. Прежде всего, речь идёт об ограниченных вычислительных ресурсах: многие датчики работают на микроконтроллерах с объёмом памяти в десятки килобайт, что делает невозможным развёртывание полноценного агентского ПО безопасности (Shi et al., 2016). Разнообразие протоколов (MQTT, CoAP, Zigbee, BLE) создаёт гетерогенную среду, в которой единые правила обнаружения вторжений теряют универсальность (Sisinni et al., 2018; Hernández-Ramos et al., 2018). Устройства нередко эксплуатируются в удалённых точках, что усложняет обновление прошивок (Du et al., 2019).

Распределённость добавляет ещё один критический фактор. В инфраструктуре умного города периферийные устройства могут находиться на значительных расстояниях от центра обработки данных, а сетевая связь между ними бывает ограничена по пропускной способности (Shi et al., 2016). Это означает, что часть аналитической работы должна выполняться непосредственно на периферии, до передачи результатов на более высокие уровни. Если воспользоваться логикой иерархического представления, близкого к модели Purdue (Williams, 1994), периферийные IoT-устройства располагаются на нижних уровнях (сенсоры, актуаторы, полевые устройства), где задержка критична, а централизованная защита запаздывает. Этим объясняется необходимость локальных механизмов предварительного анализа и фильтрации.

Традиционные IDPS и их ограничения

Классические системы обнаружения вторжений развивались в условиях относительно гомогенных корпоративных сетей. Платформы вроде Snort, Suricata и Zeek обеспечивали сигнатурный анализ трафика и некоторые возможности поведенческого обнаружения (Paxson, 1998). Однако при перенесении этих инструментов в распределённую IoT-среду обнаруживаются существенные недостатки.

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

Первый из них связан с локальностью обнаружения: каждый сенсор видит лишь свой сегмент трафика и не располагает информацией о событиях на других участках (Aktar et al., 2024). Для многостадийных атак это означает неизбежную потерю контекста (Wardana et al., 2024). Второй недостаток заключается в слабой координации: когда несколько IDPS работают параллельно, каждая генерирует свой поток алертов в собственном формате без согласования метаданных (La et al., 2021). Третий касается шума: в типичной IoT-среде число ложных срабатываний достигает 30 и более процентов от общего объёма, что приводит к усталости аналитиков (Almeida, 2023; Aktar et al., 2024).

Применение алгоритмов машинного обучения в IDPS

Методы машинного обучения всё чаще рассматриваются как средство повышения качества обнаружения. Обзорные работы (Zarpelão et al., 2017; Walling et al., 2025) показывают, что модели классификации и детекции аномалий способны выявлять паттерны в сетевом трафике, которые трудно описать вручную через сигнатурные правила. Это особенно актуально в IoT-среде, поскольку поведение устройств существенно варьируется (Yasaei et al., 2020).

При этом применение ML сопряжено с рядом сложностей. Concept drift, то есть изменение характеристик трафика во времени, приводит к деградации моделей, обученных на исторических данных (Mitchell et al., 2014). Объяснимость решений остаётся проблемой, поскольку оператору SOC важно понимать основания для квалификации активности как угрозы (Najafi et al., 2020). Наконец, даже высокоточные модели генерируют заметное число ложноположительных алертов при работе с реальным трафиком в условиях дисбаланса классов (Aktar et al., 2024). Отдельно стоит отметить работу (Moustafa et al., 2019), в которой исследованы механизмы выявления аномалий в распределённой среде с использованием статистических методов, что подтверждает актуальность данного направления.

Недостаточно решённые вопросы

Проведённый анализ позволяет зафиксировать несколько пробелов. Отсутствует достаточно гибкая архитектура IDPS, адаптированная именно для периферийного уровня IoT-инфраструктуры и допускающая модульную замену компонентов без перестройки всей системы. Недостаточно проработана интеграция ML-мониторинга с механизмами защиты периферийных узлов, включая вопрос о размещении моделей в конвейере обработки событий. Существующие решения, как правило, ограничиваются обнаружением, не обеспечивая структурирование и ранжирование событий перед передачей оператору (Bhatt et al., 2014; Najafi et al., 2021). Таким образом, требуется система, способная не только обнаруживать, но и структурировать, приоритизировать и адаптивно обрабатывать события безопасности на периферии IoT-среды.

Постановка задачи

Объектом исследования являются процессы мониторинга и защиты периферийных IoT-устройств в инфраструктуре умного города. Предметом исследования выступают методы построения гибкой IDPS на основе алгоритмов машинного обу-

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

чения.

Архитектурное описание среды мониторинга опирается на иерархическое представление инфраструктуры, близкое к модели Purdue (Williams, 1994), что позволяет разграничить периферийный, локальный и централизованный уровни защиты. Нижние уровни (сенсоры, актуаторы, полевые устройства) характеризуются минимальными ресурсами и требуют лёгких механизмов предварительного анализа. Средний уровень (edge-шлюзы, агрегаторы) позволяет размещать нормализацию, дедупликацию и локальный ML-анализ. Верхний уровень (SOC, централизованная корреляция) обеспечивает глобальное видение и управление инцидентами (Bhatt et al., 2014; Kotenko et al., 2012). Данные и алерты поднимаются от нижнего уровня к верхнему с последовательным обогащением контекстом, что соответствует логике иерархической эшелонированной защиты (Knapp et al., 2015).

Требования к системе формулируются следующим образом: работа в распределённой среде с поддержкой периферийных устройств; уменьшение избыточных алертов через дедупликацию; снижение ложных срабатываний с помощью ML-анализа; адаптация к изменяющимся условиям; обеспечение обратной связи с центром мониторинга; совместимость со стандартами обмена данными об угрозах (Iqbal et al., 2018; Sauerwein et al., 2017; ISO/IEC, 2022). Задержка обработки событий (от обнаружения до доступности в SOC) не должна превышать одной секунды (Liu et al., 2020), а пропускная способность должна составлять не менее 5000 событий в секунду (Du et al., 2019).

Предлагаемая архитектура гибкой IDPS

Общая архитектура системы

Предлагаемая архитектура включает девять функциональных компонентов, организованных в три уровня в соответствии с иерархическим представлением инфраструктуры. На рис. 1 представлена общая схема системы, показывающая направление потоков данных и обратной связи.

На первом этапе edge IDPS (Suricata или Zeek) генерирует необработанные алерты на основе сигнатурных правил и, при наличии ресурсов, лёгких поведенческих моделей (Paxson, 1998; Suricata User Guide). Далее алерты проходят предварительную фильтрацию, отсекающую заведомо нерелевантные события по заранее определённым критериям. Нормализация приводит разнородные форматы к единому JSON-представлению, совместимому со STIX 2.1 (Iqbal et al., 2018; Yasaei et al., 2020). Модуль дедупликации устраняет повторяющиеся алерты, порождённые одним и тем же инцидентом, но зарегистрированные на нескольких сенсорах. После этого события поступают в ML-модуль для дополнительного анализа и в корреляционный движок, который сопоставляет разрозненные сигналы и формирует оценку риска. На финальном этапе приоритизированные алерты передаются в SOC для визуализации и реагирования (Bhatt et al., 2014; Aktar et al., 2024).

Транспортный уровень реализован на базе Apache Kafka (Apache Kafka Documentation), выбранной за высокую пропускную способность и возможность горизонтального масштабирования. Нормализация выполняется микросервисом на

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

Python, преобразующим алерты от различных IDPS-платформ в единый формат STIX 2.1 (Iqbal et al., 2018). Спецификация API событий описана в формате OpenAPI 3.1 (OpenAPI Specification), что обеспечивает машинночитаемые контракты для валидации.

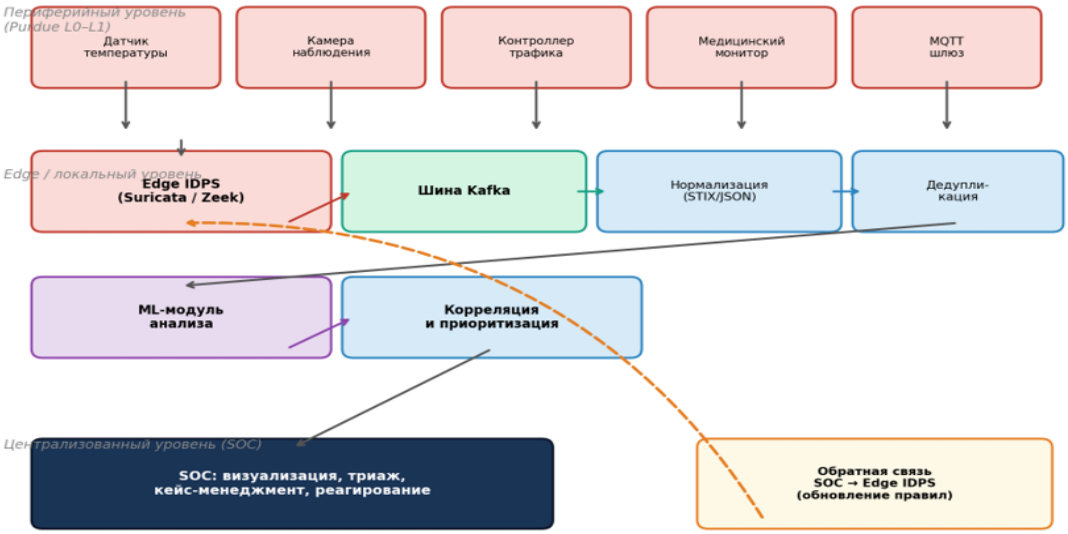


Рис. 1. Общая архитектура гибкой IDPS: от периферийных IoT-устройств через edge-уровень к централизованному SOC с обратной связью

Таблица 1 – Функциональные компоненты архитектуры гибкой IDPS

Компонент	Уровень	Назначение
Периферийные IoT-устройства	Нижний (Purdue L0–L1)	Источники телеметрии и событий
Edge IDPS (Suricata/Zeek)	Периферийный	Сигнатурное и поведенческое обнаружение
Модуль сбора событий	Периферийный	Агрегация событий от нескольких сенсоров
Шина сообщений (Kafka)	Транспортный	Потоковая доставка с гарантированной персистентностью
Модуль нормализации	Edge/локальный	Преобразование алертов в STIX/JSON формат
Модуль дедупликации	Локальный	Устранение дублирующих и избыточных алертов
ML-модуль анализа	Локальный	Детекция аномалий, классификация, снижение FP
Модуль корреляции и приоритизации	Локальный/SOC	Скоринг, ранжирование, передача в SOC
Интерфейс SOC + обратная связь	Централизованный	Визуализация, триаж, обновление правил IDPS

Логика обработки событий

Обработка событий в системе следует конвейерной логике, представленной на рис. 2. Каждое событие проходит последовательность этапов, от генерации на периферийном сенсоре до появления в интерфейсе SOC в виде приоритизированного алерта.

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

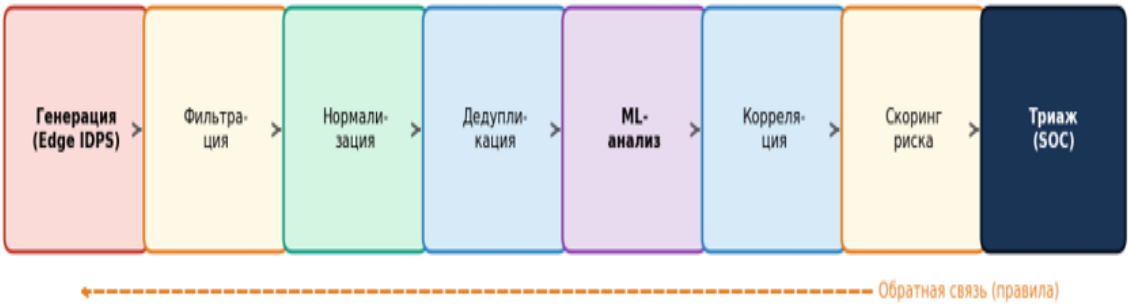


Рис. 2. Конвейер обработки событий: сбор → фильтрация → нормализация → дедупликация → ML-анализ → корреляция → скоринг → триаж

Гибкость системы

Термин «гибкая» в названии архитектуры обусловлен несколькими конструктивными особенностями. Во-первых, модульность: каждый компонент конвейера реализован как независимый микросервис, развёрнутый в Docker-контейнере (Kubernetes Documentation), что позволяет заменять, масштабировать или обновлять отдельные модули без остановки всей системы. Во-вторых, адаптация под разные типы IoT-узлов: конфигурация edge IDPS параметризована по типу устройства, протоколу и ожидаемому профилю трафика (Suricata User Guide). В-третьих, возможность замены ML-модели: аналитический модуль принимает стандартизированные входные данные и может работать с различными алгоритмами без модификации остальных компонентов. В-четвёртых, настраиваемость правил обработки и приоритизации: пороги скоринга, критерии дедупликации и политики корреляции задаются через конфигурационные файлы. Наконец, архитектура предусматривает путь расширения до SOAR (Security Orchestration, Automation and Response), что позволит в дальнейшем автоматизировать не только обнаружение, но и реагирование (Wardana et al., 2024; Shabad et al., 2021).

Использование алгоритмов машинного обучения

Роль ML в системе

ML-модуль занимает определённое место в общем конвейере обработки: он размещается после нормализации и дедупликации, но перед корреляционным движком. Его роль состоит в дополнительном анализе нормализованных событий с целью повышения точности обнаружения и снижения числа ложных срабатываний. В отличие от подходов, где ML-модель заменяет весь процесс обнаружения, в предлагаемой архитектуре машинное обучение используется как один из аналитических инструментов внутри модульного конвейера.

Конкретные задачи ML-модуля включают анализ сетевого трафика для выявления отклонений от нормального профиля устройства, классификацию событий по степени вероятности вредоносности и снижение доли ложноположительных срабатываний путём кросс-валидации с алертами от других сенсоров (Zarpelão et al., 2017;

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

Moustafa et al., 2019). Результаты работы модуля используются корреляционным движком при формировании итоговой оценки риска.

Используемые признаки и типы данных

На вход ML-модуля поступают нормализованные события в формате STIX JSON, из которых извлекаются четыре группы признаков. Сетевые характеристики включают IP-адреса источника и назначения, порты, протокол, размер пакетов и флаги. Временные признаки отражают межпакетные интервалы, длительность потока и частоту событий в скользящем окне. Поведенческие признаки описывают отклонение текущей активности от ранее наблюдённого профиля для данного типа устройства (Yasaei et al., 2020). Агрегированные события обобщают показатели по группе устройств в пределах одного сетевого сегмента.

Логика принятия решений

ML-модуль формирует для каждого события вероятностную оценку вредоносности. Пусть x обозначает вектор признаков нормализованного события, а $f(x)$ — выход ML-модели, интерпретируемый как вероятность того, что событие связано с вредоносной активностью. Тогда оценка аномальности события определяется как:

$$A(x) = f(x) \cdot w_{dev(x)} \tag{1}$$

где $w_{dev(x)}$ представляет собой весовой коэффициент, отражающий степень отклонения вектора признаков от базового профиля устройства. Чем сильнее отклонение от усреднённого поведения, тем больший вес получает событие.

Далее оценка аномальности интегрируется в общую формулу риска, учитывающую контекстуальные факторы:

$$R(x) = A(x) \cdot C_{asset} \cdot \alpha(t) \cdot \beta(conn) \tag{2}$$

где C_{asset} обозначает критичность актива (определяемую на основе его роли в инфраструктуре (Hnatiienko et al., 2024), $\alpha(t)$ отражает длительность аномального поведения (чем дольше сохраняется аномалия, тем выше вклад временного фактора), а $\beta(conn)$ учитывает сетевую связность устройства (высокосвязные узлы представляют бóльший риск распространения угрозы).

Пороговая логика используется для преобразования непрерывной оценки риска в категориальное решение о приоритете. Определяются три уровня:

$$P(x) = \{\text{низкий, если } R(x) < \theta_{low};\ \text{средний, если } \theta_{low} \le R(x) < \theta_{high};\ \text{высокий, если } R(x) \ge \theta_{high}\} \tag{3}$$

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

где пороги θ_{low} и θ_{high} настраиваются в процессе калибровки на валидационном наборе данных. Важно отметить, что в текущей реализации ML-модуль не претендует на полную замену экспертного анализа. Его задача состоит в предварительном ранжировании событий, что позволяет оператору SOC сосредоточиться на наиболее вероятных угрозах (Bhatt et al., 2014; Najafi et al., 2021).

Экспериментальная оценка

Условия эксперимента

Экспериментальная оценка проводилась на двух публичных датасетах, содержащих реалистичный IoT-трафик с размеченными атаками. BoT-IoT и TON_IoT были выбраны как представительные наборы, включающие DDoS, разведку и атаки на уровне протоколов MQTT и CoAP (Moustafa et al., 2019). Характеристики датасетов представлены в таблице 2.

Таблица 2 – Характеристики экспериментальных датасетов

Параметр	BoT-IoT	TON_IoT
Типы атак	DDoS, DoS, разведка, кража данных	DDoS, ransomware, backdoor, injection, XSS
Протоколы	MQTT, HTTP, TCP/UDP	MQTT, CoAP, HTTP, DNS
Число записей	~73 млн (подвыборка)	~460 тыс.
Метки классов	Бинарные + мультиклассовые	Бинарные + мультиклассовые

Прототип развёрнут на базе Suricata 7.0 и Zeek 6.0 в качестве edge IDPS, Apache Kafka 3.6 как шины сообщений, Python 3.11 для нормализации и корреляции, а также Elasticsearch 8.11 для интеграции с SOC (Paxson, 1998; Apache Kafka Documentation; Suricata User Guide). Развёртывание выполнено в Docker-контейнерах (Kubernetes Documentation). Для сравнения использовались два базовых подхода: (a) автономная работа IDPS без централизованной корреляции и (b) наивная пересылка алертов в SOC без нормализации и дедупликации.

Критерии оценки

Оценка производилась по следующим метрикам: задержка корреляции (время от генерации алерта до его доступности в SOC), доля ложных срабатываний (FP rate), объём алертов (число событий, поступающих аналитику после обработки), надёжность доставки сообщений и пропускная способность системы. Дополнительно оценивалась точность нормализации путём ручной проверки случайной выборки из 500 алертов по каждому датасету.

Полученные результаты

Результаты экспериментальной оценки представлены в таблице 3 и на рис. 3. Средняя задержка корреляции составила 687 мс (σ = 143 мс), с 95-м перцентилем 921 мс и 99-м перцентилем 1247 мс. По сравнению с наивной пересылкой, основанной на файловом обмене (типичная задержка 1,5–2,8 с) (Najafi et al., 2020), предлагаемая архитектура снизила задержку примерно на 62%.

Ложные срабатывания снизились на 29% для Suricata и на 31% для Zeek благодаря корреляции алертов между сенсорами: изолированные алерты, не подтверж-

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

дённые временной или межсенсорной корробрацией, подавлялись (Almeida, 2023; Zarpelão et al., 2017). Объём алертов, поступающих аналитику, сократился примерно на 40% за счёт дедупликации и подавления шума. Точность нормализации составила 96,8%, что превышает установленный порог в 95%. Ошибки нормализации (3,2%) были связаны с некорректными временными метками и пустыми обязательными полями (Yasaei et al., 2020).

Таблица 3 – Результаты оценки производительности

Метрика	Только IDPS	Наивная пересылка	Предлагаемый подход
Задержка корреляции	Нет корреляции	~1,8 с	~0,69 с
FP rate (Suricata)	30,1%	30,1%	21,4% (−29%)
FP rate (Zeek)	41,2%	41,2%	28,6% (−31%)
Объём алертов	100%	100%	~60% (−40%)
Надёжность доставки	Н/Д	~95%	99,97%
Пропускная способность	Н/Д	~2000 соб./с	>8200 соб./с
Точность нормализации	Н/Д	Н/Д	96,8%

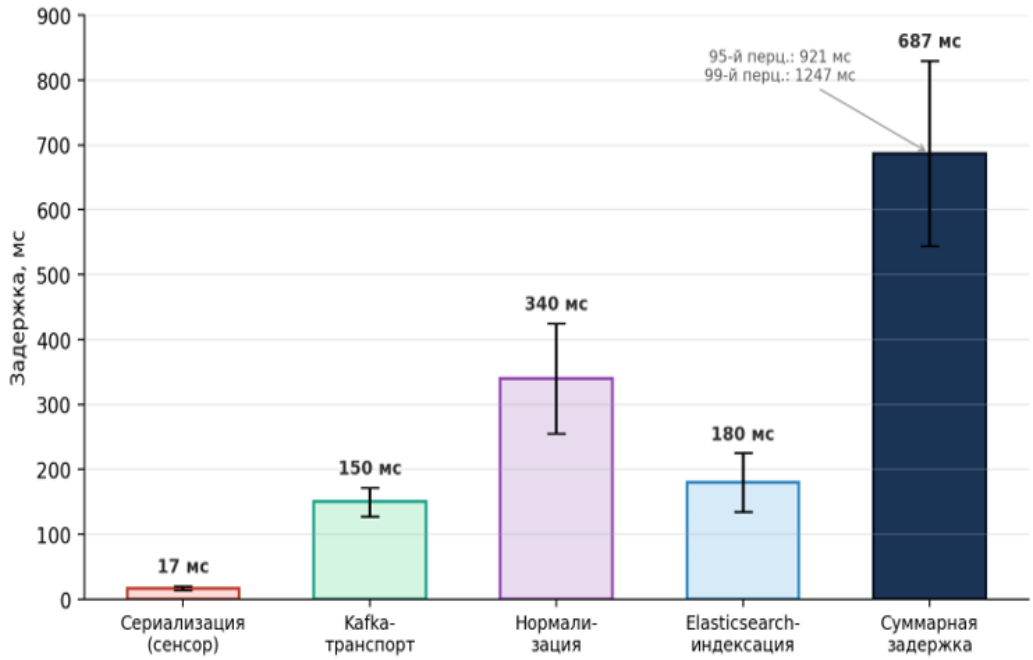


Рис. 3. Распределение задержки по этапам конвейера для 10 000 алертов (нормализация, Kafka-транспорт, Elasticsearch-индексация)

Декомпозиция задержки показала, что основной вклад вносит нормализация (~340 мс), за ней следуют Elasticsearch-индексация (~180 мс) и Kafka-транспорт (~150 мс). Сериализация на стороне сенсора занимает примерно 17 мс. Пропускная способность оставалась стабильной до 8200 событий в секунду при одном экземпляре нормализатора; при горизонтальном масштабировании до трёх экземпляров

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

система обрабатывала до 18 500 событий в секунду (Apache Kafka Documentation).

Двунаправленный механизм обратной связи обеспечил распространение обновлённых правил обнаружения от SOC ко всем edge-сенсорам в среднем за 2,3 секунды (Suricata User Guide), что позволяет адаптировать конфигурацию IDPS без необходимости ручной перенастройки каждого узла.

Результаты и обсуждение.

Полученные результаты позволяют утверждать, что архитектура на основе потоковой шины сообщений эффективно решает проблему интеграции разнородных IDPS-сенсоров в распределённой IoT-среде. В отличие от статических схем с файловой пересылкой алертов (Najafi et al., 2020), потоковая архитектура на базе Kafka устраняет задержки поллинга и обеспечивает непрерывную доставку событий. Снижение ложных срабатываний на 29–31% через кросс-сенсорную корреляцию непосредственно повышает эффективность работы аналитиков и снижает нагрузку на SOC (Almeida, 2023; Aktar et al., 2024).

Архитектура разработана с учётом принципов ISO/IEC 27001 (ISO, 2022) и GDPR (GDPR, 2016) в части шифрования каналов (TLS 1.3), аутентификации событий (HMAC-SHA256) и настраиваемых политик хранения телеметрии. Нормализация в формат STIX 2.1 (Iqbal et al., 2018) обеспечивает вендор-нейтральность, что принципиально важно для муниципальных инфраструктур, где оборудование закупается у различных поставщиков (Aktar et al., 2024).

Предложенный подход особенно релевантен для периферийных IoT-устройств, поскольку он позволяет размещать часть аналитических функций непосредственно на edge-уровне, снижая зависимость от централизованного канала связи. Иерархическая структура, согласующаяся с моделью Purdue (Williams, 1994), делает архитектуру понятной для специалистов промышленной автоматизации и облегчает её позиционирование в рамках существующих фреймворков безопасности (Knapp et al., 2015).

Вместе с тем необходимо указать на ограничения. Загрузка CPU edge-сенсоров достигала 85% при пиковых нагрузках (Paxson, 1998), что оставляет мало резерва для дополнительных аналитических задач. Нормализация является основным узким местом по задержке (~340 мс), и её оптимизация через компилируемые парсеры представляет практический интерес (Yasaei et al., 2020). Зависимость от качества входных данных остаётся критичной: при некорректных метках или пропущенных полях качество ML-анализа снижается. Необходимость периодического дообучения моделей для противодействия concept drift (Mitchell et al., 2014) требует операционных ресурсов. Наконец, проверка проведена на бенчмарк-датасетах, и подтверждение работоспособности в реальных условиях масштабного развёртывания остаётся задачей будущих исследований.

Перспективы развития

Архитектура предусматривает несколько направлений расширения. Первое из них связано с полной SOAR-автоматизацией, позволяющей перейти от ручного триажа к автоматизированным сценариям реагирования, включая изоляцию ском-

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

прометированных устройств и перенастройку сетевых политик (Wardana et al., 2024; Shabad et al., 2021). Второе направление касается внедрения online learning, то есть механизмов адаптивного обучения, корректирующих ML-модели на основе обратной связи от аналитиков SOC без полного переобучения на исторических данных. Третье направление предполагает мультиклассовую классификацию инцидентов, что позволит различать типы атак (DDoS, разведка, эксплуатация уязвимостей) и применять дифференцированные стратегии реагирования. Наконец, ключевым практическим шагом является пилотное развёртывание в реальной инфраструктуре умного города Казахстана, что позволит проверить работоспособность подхода при масштабах и условиях, не воспроизводимых в лабораторных экспериментах.

Заключение.

В статье предложена архитектура гибкой IDPS на основе алгоритмов машинного обучения для мониторинга и защиты периферийных IoT-устройств. В отличие от традиционных изолированных схем, предложенный подход обеспечивает модульную обработку событий через конвейер нормализации, дедупликации, ML-анализа, корреляции и приоритизации. Архитектурное описание среды опирается на иерархическое представление, согласующееся с моделью Purdue, что позволяет чётко разграничить периферийный, локальный и централизованный уровни защиты.

Экспериментальная оценка на датасетах BoT-IoT и TON_IoT показала сокращение задержки корреляции событий примерно на 62% по сравнению с базовыми подходами, снижение доли ложных срабатываний на 29–31% и уменьшение объёма алертов, поступающих аналитику, на 40%. Точность нормализации составила 96,8%, пропускная способность превысила 8200 событий в секунду, а механизм обратной связи обеспечил обновление правил обнаружения на всех edge-сенсорах за 2,3 секунды.

Полученные результаты подтверждают перспективность использования предложенной архитектуры при построении систем мониторинга IoT-инфраструктуры умных городов. Модульность, адаптивность и вендор-нейтральность системы создают основу для дальнейшего расширения средствами SOAR, online learning и мультиклассовой классификации. Планируется пилотное развёртывание в рамках инфраструктуры умного города Казахстана.

REFERENCES

Almeida F.A. (2023). Prospects of cybersecurity in smart cities // Future Internet. — Vol. 15. — No. 9. Art. 285.

Aktar M.N. et al. (2024). Enhancing false positive alert detection in SIEM using RNN // Proc. 27th ICCIT. Pp. 1–6.

Apache Kafka Documentation. Version 3.6 // URL: https://kafka.apache.org/documentation/

Babenko T., Kolesnikova K. et al. (2026). Hybrid GNN-LSTM architecture for probabilistic IoT botnet detection // Computers. — Vol. 15. — No. 1. Art. 26.

Bhatt S., Manadhata P.K., Zomlot L. (2014). The operational role of security information and event management systems // IEEE Secur. Priv. — Vol. 12. — No. 5. Pp. 35–41.

Du R., Santi P., Xiao M. et al. (2019). The sensable city: A survey on the deployment and management for smart city monitoring // IEEE Commun. Surv. Tutor. — Vol. 21. — No. 2. Pp. 1533–1560.

Elvas L., Mataloto B., Martins A., Ferreira J.C. (2021). Disaster management in smart cities // Smart Cities. — Vol. 4. — No. 2. Pp. 819–839.

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

Gerhards R. (2009). The syslog protocol // RFC 5424, IETF.

General Data Protection Regulation (GDPR). (2016). Regulation (EU) 2016/679 // Official J. EU. L 119. Pp. 1–88.

Hnatiienko H., Hnatiienko V., Zulunov R., Babenko T. et al. (2024). Method for determining the level of criticality elements // CEUR Workshop Proc.

Hernández-Ramos S. et al. (2018). MQTT security: A novel fuzzing approach // Wirel. Commun. Mob. Comput. Art. 8261746.

Jo J., Sharma P., Sicato J.C.S., Park J.H. (2019). Emerging technologies for sustainable smart city network security // J. Inf. Process. Syst. — Vol. 15. — No. 4. Pp. 765–784.

Iqbal Z., Anwar Z., Mumtaz R. (2018). STIXGEN: A novel framework for automatic generation of structured cyber threat information // Proc. FIT. Pp. 271–276.

ISO/IEC 27001 (2022). Information security management systems. Geneva: ISO.

Kubernetes Documentation // URL: https://kubernetes.io/docs/

Knapp E.D., Langill J.T. (2015). Industrial Network Security. 2nd ed. Waltham: Syngress.

Kotenko I., Chechulin A., Novikova E. (2012). Attack modelling and security evaluation for SIEM // Proc. SECRYPT. Rome. Pp. 391–394.

Liu Y., Ma M., Liu X. et al. (2020). Design and analysis of probing route to defense sink-hole attacks for IoT security // IEEE Trans. Netw. Sci. Eng. — Vol. 7. — No. 1. Pp. 356–372.

La V.H., Montes de Oca E., Mallouli W., Cavalli A. (2021). A framework for security monitoring of real IoT testbeds // Proc. ICSOFT. Pp. 1–12.

Moustafa N., Choo K.-K.R., Radwan I., Camtepe S. (2019). Outlier Dirichlet mixture mechanism: Adversarial statistical learning for anomaly detection in the fog // IEEE Trans. Inf. Forensics Secur. — Vol. 14. — No. 8. Pp. 1975–1987.

Mitchell R., Chen I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems // ACM Comput. Surv. — Vol. 46. — No. 4. Art. 55.

Najafi P., Cheng F., Meinel C. (2021). SIEMA: Bringing advanced analytics to legacy SIEM // SecureComm 2020. LNCS. 2021. — Vol. 357. — Pp. 21–39.

OpenAPI Specification v3.1.0 // URL: https://spec.openapis.org/oas/v3.1.0

Paxson V. (1998). Bro: A system for detecting network intruders in real-time // Proc. 7th USENIX Secur. Symp.

Palko D., Babenko T. et al. (2023). Cyber security risk modeling in distributed information systems // Appl. Sci. — Vol. 13. — No. 4. Art. 2393.

Shi W., Cao J., Zhang Q. et al. (2016). Edge computing: Vision and challenges // IEEE Internet Things J. — Vol. 3. — No. 5. Pp. 637–646.

Sisinni E., Saifullah A., Han S. et al. (2018). Industrial Internet of Things: Challenges, opportunities, and directions // IEEE Trans. Ind. Informat. — Vol. 14. — No. 11. Pp. 4724–4734.

Suricata User Guide. Version 7.0 // URL: https://docs.suricata.io/en/latest/

Shabad P.K.R., Alrashide A., Mohammed O. (2021). Anomaly detection in smart grids using ML // Proc. IECON. Pp. 1–6.

Sauerwein C. et al. (2017). Threat intelligence sharing platforms // Proc. 13th WI. Pp. 837–851.

Walling S., Lodh S. (2025). An extensive review of ML and DL techniques on network intrusion detection for IoT // Trans. Emerg. Telecommun. Technol. DOI: 10.1002/ett.70064.

Wardana A.A., Sukarno P. (2024). Taxonomy and survey of collaborative IDS using federated learning // ACM Comput. Surv. — Vol. 56. — No. 7. Pp. 1–37.

Yasaei R., Hernandez F., Faruque M.A. (2020). IoT-CAD: Context-aware adaptive anomaly detection in IoT // Proc. ICCAD. Pp. 1–9.

Williams T.J. (1994). The Purdue enterprise reference architecture // Comput. Ind. — Vol. 24. — No. 2–3. Pp. 141–158.

Zarpelão B.B., Miani R.S., Kawakani C.T. et al. (2017). A survey of intrusion detection in IoT // J. Netw. Comput. Appl. — Vol. 84. — Pp. 25–37.

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 30.06.2026.
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание АО "Международного университета информационных технологий"