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. 3 (27). 2026
July–September

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

2020 жылдан бері шығарылады
Том 7. 3 (27). 2026
Шілде-Қыркүйек

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Издается с 2020 г.
Том 7. 3 (27). 2026
Июль-Сентябрь

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

A Adalbek, S.Ye. Kerimkhulle, K.D. Kuliev, O. Mekesh, Zh.Zh. Khamitova
DEVELOPMENT OF RESULT-ORIENTED DATA SMOOTHING MODELS AND ANALYSIS OF THEIR EFFECTIVENESS 11
A.T. Iskakova, G.E. Kurmankulova, K. Gromaszek, K.O. Togzhanova, R.S. Algozhayeva
APPLICATION OF MACHINE LEARNING AND DEEP LEARNING FOR PREDICTIVE PLANNING AND RESOURCE EFFICIENCY IN PRODUCTION SYSTEMS 30
A.B. Kassymova, R.K. Uskenbayeva, Zh.A. Sydykov, D. Omirzak, S. Kalibekov
ARCHITECTURE OF AN INTELLIGENT CHATBOT FOR SMART CITY: AI-BASED RECEIVING, ANALYSIS AND ROUTING OF CITIZEN REQUESTS 46
O.M. Olzhayev, B.S. Omarov, A. Suliman
YOLO-CBAM: A LIGHTWEIGHT NEURAL NETWORK MODEL WITH INTEGRATED ATTENTION MECHANISM FOR ROAD DAMAGE ANALYSIS 66
K. Yermagambetova, A. Sarsekeyeva, Y. Akhlassov
OVERCOMING FRAGMENTATION IN THE CLASSIFICATION OF DIGITAL CULTURAL HERITAGE: AN INTEGRATIVE MULTIDIMENSIONAL MODEL 78

INFORMATION TECHNOLOGY

D. Abzhanova, S. Sharipova
ARCHITECTURE AND PRACTICAL IMPLEMENTATION OF A DIGITAL PLATFORM FOR MONITORING AND DATA EXCHANGE ON TRANSBOUNDARY WATER RESOURCES IN THE ZHAMBYL REGION 101
A. Aitanuly, O.A. Ussatova
ENHANCING TWO-FACTOR AUTHENTICATION FOR CLOUD HEALTHCARE SECURITY: THE ROLE OF DEEP LEARNING 122
O. Alseitov, O. Kuchanskyi
PHYSICS-INFORMED HYBRID LEARNING FOR HOURLY POWER-PLANT EMISSIONS: A MULTI-POLLUTANT BENCHMARK WITH UNCERTAINTY AND TRANSFER 141
A.M. Akhmetova, D.O. Baizhanova, Zh.N. Alisheva, B.A. Shaimov
ADAPTIVE OUTPUT ROUTER FOR NATURAL LANGUAGE MODELS CONSIDERING KEY PERFORMANCE INDICATORS 160
A.K. Aitim
AUTOMATIC AND SEMI-AUTOMATIC ANNOTATION OF KAZAKH TEXT CORPORA 180
Z.E. Balabayev, S.V. Biloshchytska
HANDWRITTEN TEXT RECOGNITION FOR CHAGATAI MANUSCRIPTS: COMPARING CRNN TRAINING FROM SCRATCH AND TRANSFORMER-BASED TRANSFER LEARNING 198
D.B. Baumuratova, T.K. Zhukabayeva, K.A. Abdrakhmanov, Z. Ahmad, A.A. Seitov
A NEURO-DIGITAL LEARNING MODEL FOR AI DRIVEN DIGITAL TRANSFORMATION IN HIGHER EDUCATION 219
T. Daoyun, G.A. Amirkhanova, Fu Yanwei, Y. Faruzkyzy
MULTIMODAL SEMANTIC MANIPULATION IN FAKE VIDEOS AND ITS INFLUENCE MECHANISM ON USERS' PERCEIVED TRUST 241
Zh. Doumchariyev, J. Tussupov, A. Akynbekova, M. Khassanova, M. Kassim
ENSEMBLE AND HYBRID LEARNING FOR DIGITAL TWIN MILK CLASSIFICATION 258
K.Z. Igibayev, A.Kh. Kozbakova, O. Auelbekov , T. Merembayev, K. Yessentayev
STUDY OF A DYNAMIC MIXING DEVICE IN A BIOREACTOR USING IOT TECHNOLOGIES 279
A.N. Ismukanova, G. Kochshanova, D.V. Pleskachev, Zh. Sarsenbayeva, P. Schmidt
AUTOMATED CLASSIFICATION OF SCIENTIFIC DATA USING ARTIFICIAL INTELLIGENCE: A COMPARATIVE ANALYSIS OF TRANSFORMER ARCHITECTURES AND HYBRID APPROACHES 306
T. Zh. Kamalov, Ali Abd. Almisreb, Mohammed A. Saleh
INVESTIGATION OF HYBRID QUANTUM-CLASSICAL ALGORITHMS FOR MACHINE LEARNING 323
B. Karimkyzy, K.G. Karasenko, N.A. Alisherov, N.M. Muratbek, A. Karim

KEY PUBLICATIONS OF 2021–2025 ON ARTIFICIAL INTELLIGENCE METHODS IN THE MANUFACTURING AND DIAGNOSTICS OF MEDICAL DEVICES FOR GLYCEMIC MONITORING: A THEMATIC ANALYTICAL REVIEW 338
S.K. Kumyzbaeva, K.E. Baglanbaev, K.S. Sarkyt, M.Zh. Konyrova
THREE-PHASE TRAINING OF INTELLIGENT ENERGY MANAGEMENT FOR A HYBRID ENERGY SYSTEM WITH A CLIMATE-AWARE STATE REPRESENTATION 357
N.E. Karabayev, S.K. Serikbayeva, E. Benkhelif, B. Tassuov
RESEARCH ON THE ROBUSTNESS OF ENSEMBLE MACHINE LEARNING MODELS TO ADVERSARIAL ATTACKS IN INTRUSION DETECTION SYSTEMS FOR IOT NETWORKS 369
M. Nurtas, G. Kapyshev
A METHODOLOGICAL FRAMEWORK FOR THE PREPROCESSING AND INTEGRATION OF GEOSPATIAL AND GEOPHYSICAL DATA FOR MINERAL PROSPECTIVITY PREDICTION USING MACHINE LEARNING 392
A. Neftissov, A. Biloshchytskyi, I. Kazambayev, A. Medetkhan, L. Kirichenko
ENERGY EFFICIENT MICROCLIMATE STABILIZATION FOR SEALED ARCHIVAL STORAGE ENCLOSURES 405
A.Zh. Omurtayeva, U.T. Makhazhanova, A.K. Adilova, Sh.K. Yelezhanova, J. Kultan
AN ADAPTIVE NEURAL-FUZZY MODEL FOR ASSESSING INVESTMENT RISKS IN THE AGRI-INDUSTRIAL COMPLEX: ODERNISATION OF THE DECISION-SUPPORT ARCHITECTURE THROUGH THE INTEGRATION OF MACHINE LEARNING 421
N. Shazhdekeyeva, A. Mukhanbetkaliyeva, A. Abdigaliyeva, A. Zhanekeshova, L. Nurgaliyeva
COMPARATIVE ANALYSIS OF MATHEMATICAL MODELS OF COLLABORATIVE FILTERING IN THE DEVELOPMENT OF A MOVIE RECOMMENDER SYSTEM 445
S. Sharmukhanbet, G. Turmukhanova, O. Findik, V. Makhatova, M. Rakhmetov
AN INTELLIGENT PIPELINE DEFECT DETECTION SYSTEM USING AN IN-PIPE ROBOT 461

INFORMATION SECURITY AND COMMUNICATIONTECHNOLOGIES

S.Z. Daruish, A.U. Utegenova, A.U. Kalizhanova, Zh.Zh. Zhumabekova
MACHINE LEARNING FOR POLLUTION CLASSIFICATION IN ROCKET BOOSTER IMPACT AREAS 480
S.N. Issabayeva, A.K. Rysbayeva, Zh.B. Ibrayeva, N.R. Mussabekov, N.Sh. Akimzhan
MACHINE LEARNING-BASED CYBERATTACK DETECTION IN OT-NETWORKS: A COMPARATIVE ANALYSIS OF NETWORK ANOMALY CLASSIFICATION METHODS 495
G. Shangytbayeva, A.N. Amanov, M.K. Issenova, G.E. Berikkhanova, K.O. Shakerkhan
INFORMATION SECURITY IN GOVERNMENT CLOUD SERVICES: RISK ASSESSMENT AND PROTECTION STRATEGIES 512
Ye. N. Tursynbek, N. Albanbay, D. Djenouri
OPTIMISATION OF COMMUNICATION COSTS IN TWO-TIER FEDERATED LEARNING FOR CYBERATTACK DETECTION IN IOT NETWORKS 529
A. Tastemirov, S.B. Mukhanov, S.T. Amanzholova, B. Aldosh
A COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS FOR ACTIVE DIRECTORY THREAT DETECTION: FEATURE-LEAKAGE CONTROL AND CROSS-SCENARIO VALIDATION ON WINDOWS SECURITY EVENT LOGS 545

МАЗМҰНЫ
ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР

А Адалбек, С.Е. Керімқұл, К.Д. Кулиев, О. Мекеш, Ж.Ж. Хамитова
ДЕРЕКТЕРДІҢ НӘТИЖЕ БАҒДАРЛАНҒАН ТЕГІСТЕУ МОДЕЛЬДЕРІН ӘЗІРЛЕУ ЖӘНЕ ОЛАРДЫҢ ТИІМДІЛІГІН ТАЛДАУ 11
А.Т. Искакова, Г.Е. Курманкулова, Конрад Громашек, К.О. Тогжанова, Р.С. Алгожаева
ӨНДІРІС ЖҮЙЕЛЕРІНДЕ БОЛЖАМДЫ ЖОСПАРЛАУ ЖӘНЕ РЕСУРСТАРДЫ ҮНЕМДЕУ ҮШІН МАШИНАЛЫҚ ЖӘНЕ ТЕРЕҢ ОҚЫТУДЫ ҚОЛДАНУ 30
А.Б. Касымова, Р.К. Ускенбаева, Ж.А. Сыдыков, Д. Өмірзақ, С. Калибеков
SMART CITY ҮШІН ИНТЕЛЛЕКТУАЛДЫ ЧАТ-БОТ АРХИТЕКТУРАСЫ: АЗАМАТТАРДЫҢ

ӨТІНІШТЕРІН ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ ҚАБЫЛДАУ, ТАЛДАУ ЖӘНЕ БАҒЫТТАУ 46
О.М. Олжаев, Б.С. Омаров, A.Suliman
YOLO-CBAM: ЖОЛДЫҢ ЗАҚЫМДАНУЫН ТАЛДАУҒА АРНАЛҒАН ИНТЕГРАЦИЯЛАНҒАН ЗЕЙІН МЕХАНИЗМІ БАР ЖЕҢІЛ САЛМАҚТЫ НЕЙРОНДЫҚ ЖЕЛІ МОДЕЛІ 66
К. Ермагамбетова, А. Сарсекеева, Е. Ахласов
ЦИФРЛЫҚ МӘДЕНИ МҰРАНЫ ЖІКТЕУДЕГІ ФРАГМЕНТАЦИЯНЫ ЕҢСЕРУ: ИНТЕГРАТИВТІ КӨПӨЛШЕМДІ МОДЕЛЬ 78

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР

Д.Е. Абжанова, С.Е. Шарипова
ЖАМБЫЛ ӨҢІРІНІҢ ТРАНСШЕКАРАЛЫҚ СУ РЕСУРСТАРЫН МОНИТОРИНГТЕУ ЖӘНЕ ДЕРЕКТЕР АЛМАСУҒА АРНАЛҒАН ЦИФРЛЫҚ ПЛАТФОРМАНЫҢ АРХИТЕКТУРАСЫ МЕН ПРАКТИКАЛЫҚ ІСКЕ АСЫРЫЛУЫ 101
А.Айтанұлы, О.А. Усатова
БҰЛТТЫ МЕДИЦИНАЛЫҚ ЖҮЙЕЛЕРДЕ ЕКІ ФАКТОРЛЫ АУТЕНТИФИКАЦИЯНЫ ЖАҚСАРТУ: ТЕРЕҢ ОҚЫТУ ӘДІСТЕРІНІҢ ҚОЛДАНУ МҮМКІНДІКТЕРІ 122
О. Альсеитов, О. Кучанский
ЭЛЕКТР СТАНЦИЯЛАРЫНЫҢ САҒАТТЫҚ ШЫҒАРЫНДЫЛАРЫН ФИЗИКАЛЫҚ ЗАҢДЫЛЫҚТАРДЫ ЕСКЕРЕТІН ГИБРИДТІ ОҚЫТУ: БЕЛГІСІЗДІК ПЕН ТАСЫМАЛДАУДЫ ЕСКЕРЕТІН КӨПЛАСТАУЫШТЫ БЕНЧМАРК 141
А.М. Ахметова, Д.О. Байжанова, Ж.Н. Алишева, Б.А. Шаймов
КРІ КӨРСЕТКІШТЕРІН ЕСКЕРЕ ОТЫРЫП, ТАБИҒИ ТІЛ МОДЕЛЬДЕРІНЕ АРНАЛҒАН ШЫҒЫС МАРШРУТТАУЫНЫҢ АДАПТИВТІ ЖҮЙЕСІ 160
Ә.Қ. Әйтім
ҚАЗАҚ МӘТІНДІК КОРПУСТАРЫН АВТОМАТТЫ ЖӘНЕ ЖАРТЫЛАЙ АВТОМАТТЫ АННОТАЦИЯЛАУ 180
Ж.Е. Балабаев, С.В. Белощицкая
ШАҒАТАЙ ҚОЛЖАЗБАЛАРЫНЫҢ МӘТІНІН ТАНУ: CRNN-ДІ НӨЛДЕН БАСТАП ОҚЫТУ МЕН ТРАНСФОРМЕР НЕГІЗІНДЕГІ ТРАНСФЕРЛІК ОҚЫТУДЫ САЛЫСТЫРУ 198
Д.Б. Баумуратова, Т.К. Жукабаева, К.А. Абдрахманов, З. Ахмад, А.А. Сеитов
ЖОҒАРЫ БІЛІМ БЕРУДЕ AI НЕГІЗДЕЛГЕН ЦИФРЛЫҚ ТРАНСФОРМАЦИЯҒА АРНАЛҒАН НЕЙРО-ЦИФРЛЫҚ ОҚЫТУ МОДЕЛІ 219
Т. Даоюн, Г. Амирханова, Фу Яньвэй, Е. Фарузқызы
ФЕЙК ВИДЕОЛАРДАҒЫ МУЛЬТИМОДАЛЬДЫ СЕМАНТИКАЛЫҚ МАНИПУЛЯЦИЯНЫҢ ПАЙДАЛАНУШЫЛАРДЫҢ ҚАБЫЛДАЙТЫН СЕНІМІНЕ ӘСЕР ЕТУ МЕХАНИЗМІ 241
Ж. Доумчариева, Д. Тусупов, А. Акынбекова, М. Хасанова, М. Кассим
ЦИФРЛЫҚ ЕГІЗДЕГІ СҮТ САПАСЫН АНСАМБЛЬДІК ЖӘНЕ ГИБРИДТІ ОҚЫТУ АРҚЫЛЫ ЖІКТЕУ 258
Қ.З. Игибаев, А.Х. Козбакова, О. Әуелбеков, Т. Мерембаев, Қ. Есентаев
БИОРЕАКТОРДА IoT ТЕХНОЛОГИЯЛАРЫН ПАЙДАЛАНА ОТЫРЫП ДИНАМИКАЛЫҚ АРАЛАСТЫРУ ҚҰРЫЛҒЫСЫН ЗЕРТТЕУ 279
А.Н. Исмуканова, Г.Кощанова, Д.В. Плескачев, Ж.Сарсенбаева, П. Шмидт
ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕ ҒЫЛЫМИ ДЕРЕКТЕРДІ АВТОМАТТЫ ТҮРДЕ ЖІКТЕУ: ТРАНСФОРМЕР АРХИТЕКТУРАЛАРЫ МЕН ГИБРИДТІ ТӘСІЛДЕРДІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ 306
Т.Ж. Камалов, Али Абд. Альмисреб, Мохаммед А. Салех
МАШИНАЛЫҚ ОҚЫТУҒА АРНАЛҒАН ГИБРИДТІ КВАНТТЫҚ-КЛАССИКАЛЫҚ АЛГОРИТМДЕРІН ЗЕРТТЕУ 323
Б. Кәрімқызы, К.Г. Карасенко, Н.А. Әлішеров, Н.М. Мұратбек, А. Кәрім
2021–2025 ЖЫЛДАР АРАЛЫҒЫНДАҒЫ ГЛИКЕМИЯЛЫҚ МОНИТОРИНГКЕ АРНАЛҒАН МЕДИЦИНАЛЫҚ ҚҰРЫЛҒЫЛАРДЫ ӨНДІРУ ЖӘНЕ ДИАГНОСТИКАЛАУДАҒЫ ЖАСАНДЫ ИНТЕЛЛЕКТ ӘДІСТЕРІ БОЙЫНША НЕГІЗГІ ЖАРИЯЛАНЫМДАР 338
С. К. Кумызбаева, К.Е. Багланбаев, К.С. Саркыт, М. Конырова
КҮЙ-ЖАЙДЫ КЛИМАТҚА СЕЗІМТАЛ ҰСЫНУМЕН ГИБРИДТІ ЭНЕРГОЖҮЙЕНІҢ ИНТЕЛЛЕКТУАЛДЫ ЭНЕРГОМЕНЕДЖМЕНТІН ҮШФАЗАЛЫ ОҚЫТУ 357
Н.Е. Қарабаев, С.К. Серикбаева, Е. Бенкхелифа, Б. Тасуов

IOT-ЖЕЛІЛЕРІНДЕГІ БАСЫП КІРУДІ АНЫҚТАУЖҮЙЕЛЕРІНДЕ МАШИНАЛЫҚ ОҚЫТУДЫҢ АНСАМБЛЬДІ КҮЛГІЛЕРІНІҢ БӘСЕКЕЛЕС ШАБУЫЛДАРҒА ТӨЗІМДІЛІГІН ЗЕРТТЕУ 369
М. Нұртас, Ғ. Капышев
МАШИНАЛЫҚ ОҚЫТУДЫ ҚОЛДАНА ОТЫРЫП, ПАЙДАЛЫ ҚАЗБАЛАРДЫҢ ПЕРСПЕКТИВАЛЫЛЫҒЫН БОЛЖАУ ҮШІН ГЕОКЕҢІСТІКТІК ЖӘНЕ ГЕОФИЗИКАЛЫҚ ДЕРЕКТЕРДІ АЛДЫН АЛА ӨҢДЕУ ЖӘНЕ БІРІКТІРУДІҢ ӘДІСТЕМЕЛІК НЕГІЗІ 392
А. Нефтиссов, А. Билощицкий, И. Казамбаев, А. Медетхан, Л. Кириченко
ГЕРМЕТИКАЛЫҚ АРХИВТІК САҚТАУ ҚАПТАМАЛАРЫНДА МИКРОКЛИМАТТЫ ЭНЕРГИЯ ТИІМДІЛІГІН ЕСКЕРЕ ОТЫРЫП ТҰРАҚТАНДЫРУ 405
Ә.Ж. Омуртаева, У.Т. Махажанова, А.К. Адилова, Ш.К. Ележанова, Я. Култан
АГРОӨНЕРКӘСІПТІК КЕШЕНДЕГІ ИНВЕСТИЦИЯЛЫҚ ТӘУЕКЕЛДЕРДІ БАҒАЛАУҒА АРНАЛҒАН БЕЙІМДЕЛГІШ НЕЙРОНДЫҚ-БҰЛДЫР МОДЕЛЬ: МАШИНАЛЫҚ ОҚЫТУДЫ ИНТЕГРАЦИЯЛАУ АРҚЫЛЫ ШЕШІМ ҚАБЫЛДАУДЫ ҚОЛДАУ СӘУЛЕТІН ЖАҢҒЫРТУ 421
Н. Шаждекеева, А. Мұханбетқалиева, А. Әбдіғалиева, А. Жанекешова, Л. Нұрғалиева
ФИЛЬМДЕРДІ ҰСЫНУ ЖҮЙЕСІН ҚҰРУ КЕЗІНДЕГІ КОЛЛАБОРАТИВТІК СҮЗГІЛЕУДІҢ МАТЕМАТИКАЛЫҚ МОДЕЛЬДЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ 445
С. Шармұханбет, Г. Тұрмұханова, О. Финдик, В. Махатова, М. Рахметов
ҚҰБЫРІШІЛІК РОБОТТЫ ПАЙДАЛАНУ АРҚЫЛЫ ҚҰБЫР АҚАУЛАРЫН АНЫҚТАУДЫҢ ЗИЯТКЕРЛІК ЖҮЙЕСІ 461

АҚПАРАТТЫҚ ҚАУІПСІЗДІК ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАРҒА АРНАЛҒАН

С.З. Дәруіш, А.У. Утегенова, А.У. Калижанова, Ж.Ж. Жумабекова
ЗЫМЫРАН ТАСЫМАЛДАУШЫЛАРДЫҢ ҚҰЛАУ АЙМАҚТАРЫНЫҢ ЛАСТАНУЫН МАШИНАЛЫҚ ОҚЫТУМЕН ЖІКТЕУ 480
С.Н. Исабаева, А.К. Рысбаева, Ж.Б. Ибраева, Н.Р. Муссабеков, Н.Ш. Акимжан
МАШИНАЛЫҚ ОҚЫТУ НЕГІЗІНДЕ ОТ-ЖЕЛІЛЕРІНДЕГІ КИБЕРШАБУЫЛДАРДЫ АНЫҚТАУ: ЖЕЛІЛІК АНОМАЛИЯЛАРДЫ ЖІКТЕУ ӘДІСТЕРІНІҢ САЛЫСТЫРМАЛЫ ТАЛДАУЫ 495
Г.А. Шангытбаева, А.Н. Аманов, М.К. Исенова, Г.Е. Берикханова, Қ.О. Шакерхан
МЕМЛЕКЕТТІК ОРГАНДАРДЫҢ БҰЛТТЫ СЕРВИСТЕРІНДЕГІ АҚПАРАТТЫҚ ҚАУІПСІЗДІК: ТӘУЕКЕЛДЕРДІ БАҒАЛАУ ЖӘНЕ ҚОРҒАУ СТРАТЕГИЯЛАРЫ 512
Е.Н. Турсынбек, Н. Албанбай, Д. Дженоури
IOT ЖЕЛІЛЕРІНДЕ КИБЕРШАБУЫЛДАРДЫ АНЫҚТАУ ҮШІН ЕКІ ДЕҢГЕЙЛІ ФЕДЕРАЦИЯЛАНҒАН ОҚЫТУДАҒЫ БАЙЛАНЫС ШЫҒЫНДАРЫН ОҢТАЙЛАНДЫРУ 529
А. Тастемиров, С.Т. Аманжолова, Б. Алдош, С.Б. Муханов
ACTIVE DIRECTORY ЖҮЙЕСІНДЕ ҚАУІПТЕРДІ АНЫҚТАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН САЛЫСТЫРМАЛЫ БАҒАЛАУ: WINDOWS ҚАУІПСІЗДІК ОҚИҒАЛАРЫ ЖУРНАЛДАРЫНДА БЕЛГІ ЖЫЛЫСТАУЫН БАҚЫЛАУ ЖӘНЕ КРОСС-СЦЕНАРЛЫҚ ВАЛИДАЦИЯ 545

СОДЕРЖАНИЕ
ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИСОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

А Адалбек, С.Е. Керимкулов, К.Д. Кулиев, О. Мекеш, Ж.Ж. Хамитова
РАЗРАБОТКА РЕЗУЛЬТАТИВНО-ОРИЕНТИРОВАННЫХ МОДЕЛЕЙ СГЛАЖИВАНИЯ ДАННЫХ И АНАЛИЗ ИХ ЭФФЕКТИВНОСТИ 11
А.Т. Искакова, Г.Е. Курманкулова, К. Громашек, К.О. Тогжанова, Р.С. Алгожаева
ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО И ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗНОГО ПЛАНИРОВАНИЯ И РЕСУРСОСБЕРЕЖЕНИЯ В ПРОИЗВОДСТВЕННЫХ СИСТЕМАХ 30
А.Б. Касымова, Р.К. Ускенбаева, Ж.А.Сыдыков, Д. Омирзак, С. Калибеков
АРХИТЕКТУРА ИНТЕЛЛЕКТУАЛЬНОГО ЧАТ-БОТА ДЛЯ SMART CITY: ПРИЁМ, АНАЛИЗ И МАРШРУТИЗАЦИЯ ОБРАЩЕНИЙ ГРАЖДАН НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА 46
О.М. Олжаев, Б.С. Омаров, A.Suliman
YOLO-CBAM: ЛЕГКОВЕСНАЯ НЕЙРОСЕТЕВАЯ МОДЕЛЬ С ИНТЕГРИРОВАННЫМ МЕХАНИЗМОМ ВНИМАНИЯ ДЛЯ АНАЛИЗА ДОРОЖНЫХ ПОВРЕЖДЕНИЙ 66
К. Ермагамбетова, А. Сарсекеева, Е. Ахласов
ПРЕОДОЛЕНИЕ ФРАГМЕНТАЦИИ В КЛАССИФИКАЦИИ ЦИФРОВОГО КУЛЬТУРНОГО НАСЛЕДИЯ:

ИНТЕГРАТИВНАЯ МНОГОМЕРНАЯ МОДЕЛЬ 78

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Д.Е. Абжанова, С.Е. Шарипова
АРХИТЕКТУРА И ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ЦИФРОВОЙ ПЛАТФОРМЫ МОНИТОРИНГА И ОБМЕНА ДАННЫМИ О ТРАНСГРАНИЧНЫХ ВОДНЫХ РЕСУРСАХ ЖАМБЫЛСКОГО РЕГИОНА 101
А. Айтанулы, О.А. Усатова
СОВЕРШЕНСТВОВАНИЕ ДВУХФАКТОРНОЙ АУТЕНТИФИКАЦИИ ДЛЯ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ ОБЛАЧНЫХ МЕДИЦИНСКИХ СИСТЕМ: ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ 122
О. Альсеитов, О. Кучанский
ФИЗИЧЕСКИ ИНФОРМИРОВАННОЕ ГИБРИДНОЕ ОБУЧЕНИЕ ДЛЯ ПОЧАСОВОГО ПРОГНОЗИРОВАНИЯ ВЫБРОСОВ ЭЛЕКТРОСТАНЦИЙ: МНОГОКОМПОНЕНТНЫЙ БЕНЧМАРК С ОЦЕНКОЙ НЕОПРЕДЕЛЁННОСТИ И ПЕРЕНОСИМОСТЬЮ МОДЕЛЕЙ 141
А.М. Ахметова, Д.О. Байжанова, Ж.Н. Алишева, Б.А. Шаймов
АДАПТИВНЫЙ МАРШРУТИЗАТОР ВЫВОДА С УЧЕТОМ КЛЮЧЕВЫХ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ ДЛЯ МОДЕЛЕЙ ЕСТЕСТВЕННОГО ЯЗЫКА 160
Ә.Қ. Әйтім
АВТОМАТИЧЕСКАЯ И ПОЛУАВТОМАТИЧЕСКАЯ АННОТАЦИЯ КАЗАХСКИХ ТЕКСТОВЫХ КОРПУСОВ 180
Ж.Е. Балабаев, С.В. Белощицкая
РАСПОЗНАВАНИЕ РУКОПИСНОГО ТЕКСТА ШАГАТАЙСКИХ РУКОПИСЕЙ: СРАВНЕНИЕ ОБУЧЕНИЯ CRNN С НУЛЯ И ТРАНСФЕРНОГО ОБУЧЕНИЯ НА БАЗЕ ТРАНСФОРМЕРОВ 198
Д.Б. Баумуратова, Т.К. Жукабаева, К.А. Абдрахманов, З. Ахмад, А.А. Сеитов
НЕЙРО-ЦИФРОВАЯ МОДЕЛЬ ОБУЧЕНИЯ ДЛЯ ЦИФРОВОЙ ТРАНСФОРМАЦИИ ВЫСШЕГО ОБРАЗОВАНИЯ С ПОМОЩЬЮ AI 219
Т. Даоюн, Г. Амирханова, Фу Яньвэй, Е. Фарузкызы
МЕХАНИЗМ ВЛИЯНИЯ МУЛЬТИМОДАЛЬНОЙ СЕМАНТИЧЕСКОЙ МАНИПУЛЯЦИИ В ФЕЙКОВЫХ ВИДЕО НА ВОСПРИНИМАЕМОЕ ДОВЕРИЕ ПОЛЬЗОВАТЕЛЕЙ 241
Ж. Доумчариева, Д. Тусупов, А. Акынбекова, М. Хасанова, М. Кассим
КЛАССИФИКАЦИЯ КАЧЕСТВА МОЛОКА В ЦИФРОВОМ ДВОЙНИКЕ АНСАМБЛЕВЫМИ И ГИБРИДНЫМИ МЕТОДАМИ 258
К.З. Игибаев, А.Х. Козбакова, О. Ауелбеков, Т. Мерембаев, К. Есентаев
ИССЛЕДОВАНИЕ УСТРОЙСТВА ДИНАМИЧЕСКОГО ПЕРЕМЕШИВАНИЯ В БИОРЕАКТОРЕ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ IoT 279
А. Н. Исмуканова, Г.Кощанова, Д.В. Плескачев, Ж.Сарсенбаева, П. Шмидт
АВТОМАТИЗИРОВАННАЯ КЛАССИФИКАЦИЯ НАУЧНЫХ ДАННЫХ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТРАНСФОРМЕРНЫХ АРХИТЕКТУР И ГИБРИДНЫХ МЕТОДОВ 306
Т.Ж. Камалов, Али Абд. Альмисреб, Мохаммед А. Салех
ИССЛЕДОВАНИЕ ГИБРИДНЫХ КВАНТОВО-КЛАССИЧЕСКИХ АЛГОРИТМОВ ДЛЯ МАШИННОГО ОБУЧЕНИЯ 323
Б. Каримкызы, К.Г. Карасенко, Н.А. Алишеров, Н.М. Муратбек, А. Карим
КЛЮЧЕВЫЕ ПУБЛИКАЦИИ 2021–2025 ГОДОВ ПО МЕТОДАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРОИЗВОДСТВЕ И ДИАГНОСТИКЕ МЕДИЦИНСКИХ ИЗДЕЛИЙ ДЛЯ ГЛИКЕМИЧЕСКОГО МОНИТОРИНГА 338
С.К. Кумызбаева, К.Е. Багланбаев, К.С. Сарқыт, М.Ж. Конырова
ТРЁХФАЗОВОЕ ОБУЧЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО ЭНЕРГОМЕНЕДЖМЕНТА ГИБРИДНОЙ ЭНЕРГОСИСТЕМЫ С КЛИМАТО-ОСВЕДОМЛЁННЫМ ПРЕДСТАВЛЕНИЕМ СОСТОЯНИЯ 357
Н.Е. Қарабаев, С.К. Серикбаева, Е. Бенкхелифа, Б. Тасуов
ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ АНСАМБЛЕВЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ К СОСТЯЗАТЕЛЬНЫМ АТАКАМ В СИСТЕМАХ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ IOT-СЕТЕЙ 369
М. Нуртас, Г. Капышев
МЕТОДОЛОГИЧЕСКАЯ ОСНОВА ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ И ИНТЕГРАЦИИ ГЕОПРОСТРАНСТВЕННЫХ И ГЕОФИЗИЧЕСКИХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ПЕРСПЕКТИВНО-

СТИ МИНЕРАЛЬНЫХ РЕСУРСОВ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ 392
А. Нефтиссов, А. Билощицкий, И. Казамбаев, А. Медетхан, Л. Кириченко
ЭНЕРГОЭФФЕКТИВНАЯ СТАБИЛИЗАЦИЯ МИКРОКЛИМАТА В ГЕРМЕТИЧНЫХ АРХИВНЫХ ХРАНИЛИЩАХ 405
А.Ж. Омуртаева, У.Т. Махажанова, А.К. Адилова, Ш.К. Ележанова, Я.Култан
АДАПТИВНАЯ НЕЙРО-НЕЧЕТКАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ ИНВЕСТИЦИОННЫХ РИСКОВ В АГРОПРОМЫШЛЕННОМ КОМПЛЕКСЕ: МОДЕРНИЗАЦИЯ АРХИТЕКТУРЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПОСРЕДСТВОМ ИНТЕГРАЦИИ МАШИННОГО ОБУЧЕНИЯ 421
Н. Шаждекеева, А. Муханбеткалиева, А. Абдигалиева, А. Жанекешова, Л. Нургалиева
СРАВНИТЕЛЬНЫЙ АНАЛИЗ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ПРИ ПОСТРОЕНИИ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ФИЛЬМОВ 445
С. Шармуханбет, Г. Турмуханова, О.Финдик, В.Махатова, М. Рахметов
ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ОБНАРУЖЕНИЯ ДЕФЕКТОВ ТРУБОПРОВОДОВ С ИСПОЛЬЗОВАНИЕМ ВНУТРИТРУБНОГО РОБОТА 461

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

С.З. Дәруіш, А.У. Утегенова, А.У. Калижанова, Ж.Ж. Жумабекова
МАШИННОЕ ОБУЧЕНИЕ ДЛЯ КЛАССИФИКАЦИИ ЗАГРЯЗНЕНИЯ РАЙОНОВ ПАДЕНИЯ РАКЕТ-НОСИТЕЛЕЙ 480
С.Н. Исабаева, А.К. Рысбаева, Ж.Б. Ибраева, Н.Р. Муссабеков, Н.Ш. Акимжан
ОБНАРУЖЕНИЕ КИБЕРАТАК В ОТ-СЕТЯХ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ СЕТЕВЫХ АНОМАЛИЙ 495
Г.А. Шангытбаева, А.Н. Аманов, М.К. Исенова, Г.Е. Берикханова, Қ.О. Шакерхан
ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ В ОБЛАЧНЫХ СЕРВИСАХ ГОСУДАРСТВЕННЫХ ОРГАНОВ: ОЦЕНКА РИСКОВ И СТРАТЕГИИ ЗАЩИТЫ 512
Е.Н. Турсынбек, Н. Албанбай, Д. Дженоури
ОПТИМИЗАЦИЯ КОММУНИКАЦИОННЫХ ЗАТРАТ В ДВУХУРОВНЕВОМ ФЕДЕРАТИВНОМ ОБУЧЕНИИ ДЛЯ ОБНАРУЖЕНИЯ КИБЕРАТАК В IOT-СЕТЯХ 529
А. Тастемиров, С.Т. Аманжолова, Б. Алдош, С.Б. Муханов
СРАВНИТЕЛЬНАЯ ОЦЕНКА МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ УГРОЗ В ACTIVE DIRECTORY: КОНТРОЛЬ УТЕЧКИ ПРИЗНАКОВ И КРОСС-СЦЕНАРНАЯ ВАЛИДАЦИЯ НА ЖУРНАЛАХ БЕЗОПАСНОСТИ WINDOWS 545

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 7. Is.3. Number 27 (2026). Pp. 545–563
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2026.27.3.030

A COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS FOR ACTIVE DIRECTORY THREAT DETECTION: FEATURE-LEAKAGE CONTROL AND CROSS-SCENARIO VALIDATION ON WINDOWS SECURITY EVENT LOGS

A. Tastemirov*, S.B. Mukhanov, S.T. Amanzholova, B. Aldosh
Astana IT University, Astana, Kazakhstan.
E-mail: 255146@astanait.edu.kz

Azamat Tastemirov — student/researcher, School of Cybersecurity, Astana IT University, Astana, Kazakhstan
E-mail: 255146@astanait.edu.kz, https://orcid.org/0009-0009-6215-3081;
S.B. Mukhanov — PhD, Assistant Professor, School of Cybersecurity, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-8761-4272;
S.T. Amanzholova — Candidate of Technical Sciences, Associate Professor, School of Cybersecurity, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-6779-9393;
Balziya Aldosh — student/researcher, School of Cybersecurity, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-7358-5490.

© A. Tastemirov, S.B. Mukhanov, S.T. Amanzholova, B. Aldosh

Abstract. Windows Active Directory (AD) is the backbone of authentication and authorization in most enterprise networks and is therefore a priority target for credential abuse, lateral movement, and privilege escalation. Windows Security Event Logs record detailed structured telemetry of these actions, but rule-based detection remains brittle: it generates excessive false positives and overlooks patterns that span multiple log fields. In this work five supervised machine-learning model families — Logistic Regression, linear Support Vector Machine, Random Forest, Gradient Boosting, and XGBoost — are compared on 23,268 AD-related Windows events drawn from ten public adversary-emulation captures (OTRF Security-Datasets, 2024). All models share a common preprocessing pipeline that explicitly audits every log field for session-level leakage and discards 108 of the 235 raw columns before encoding. Each model is evaluated under three complementary protocols: a 60/40 stratified holdout, stratified 5-fold cross-validation, and leave-one-scenario-out (LOSO), in

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

which the model is tested on an entire adversary-emulation capture that was unseen during training. Under the standard protocols all five families look strong (CV5 ROC-AUC > 0.85 and PR-AUC > 0.95), yet under LOSO the same models collapse to ROC-AUC in the range 0.15–0.45 and Matthews correlation near zero. This protocol gap, obtained on identical features and hyperparameters, quantifies the dataset-specific signal that any single-split evaluation hides. We discuss its implications for dataset bias, interpretability, inference cost, SOC deployment, and the sequence-level attack chains that an event-level classifier cannot capture by construction.

Keywords: Active Directory, Windows Security Logs, threat detection, intrusion detection, machine learning, data leakage, leave-one-scenario-out, XGBoost, Random Forest, SIEM

For citation: Tastemirov A., Amanzholova S.T., Aldosh B., Mukhanov S.B. (2026). A comparative evaluation of machine learning models for active directory threat detection: feature-leakage control and cross-scenario validation on windows security event logs // International journal of information and communication technologies. Vol. 7. No. 27. Pp. 545–563. https://doi.org/10.54309/IJICT.2026.27.3.030. (In Eng.).

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

ACTIVE DIRECTORY ЖҮЙЕСІНДЕ ҚАУІПТЕРДІ АНЫҚТАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН САЛЫСТЫРМАЛЫ БАҒАЛАУ: WINDOWS ҚАУІПСІЗДІК ОҚИҒАЛАРЫ ЖУРНАЛДАРЫНДА БЕЛГІ ЖЫЛЫСТАУЫН БАҚЫЛАУ ЖӘНЕ КРОСС-СЦЕНАРЛЫҚ ВАЛИДАЦИЯ

А. Тастемиров*, С.Т. Аманжолова, Б. Алдош, С.Б. Муханов
Astana IT University, Астана, Қазақстан.
E-mail: 255146@astanait.edu.kz

Тастемиров Азамат — студент/зерттеуші, Киберқауіпсіздік мектебі, Astana IT University, Астана, Қазақстан
E-mail: 255146@astanait.edu.kz, https://orcid.org/0009-0009-6215-3081
Муханов С.Б. — PhD, ассистент-профессор, Киберқауіпсіздік мектебі, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-8761-4272;
Аманжолова С.Т. — техника ғылымдарының кандидаты, қауымдастырылған профессор, Киберқауіпсіздік мектебі, Astana IT University, Астана, Қазақстан https://orcid.org/0000-0002-6779-9393;
Алдош Балзия — студент/зерттеуші, Киберқауіпсіздік мектебі, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-7358-5490.

© А. Тастемиров, С.Б. Муханов, С.Т. Аманжолова, Б. Алдош

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

Аннотация. Active Directory (AD) көптеген корпоративтік желілерде аутентификация және авторизация үдерістерінің негізі болып табылады және сол себепті тіркелгі деректерін қолдану, желі ішінде көлденең қозғалыс және артықшылықтарды арттыру шабуылдарының басым мақсатына айналады. Windows қауіпсіздік оқиғалары журналдары осы әрекеттер туралы егжей-тегжейлі құрылымдалған телеметрияны тіркейді, бірақ ережеге негізделген анықтау әлсіз болып қала береді: ол жалған оң нәтижелерді шамадан тыс көп шығарады және бірнеше журнал өрісін қамтитын заңдылықтарды таба алмайды.Бұл жұмыста қадағаланатын машиналық оқытудың бес отбасы — логистикалық регрессия, сызықтық тірек векторлар әдісі (SVM), кездейсоқ орман, градиенттік бустинг және XGBoost — OTRF Security-Datasets жобасынан алынған он ашық шабуыл имитациясы жазбаларынан құралған 23 268 AD-ға қатысты Windows оқиғасында салыстырылады. Барлық модельдер бірыңғай алдын ала өңдеу құбыры бойынша оқытылады; онда әрбір журнал өрісі сессия деңгейіндегі ағуға тексеріліп, 235 бастапқы бағанның 108-і кодтауға дейін алынып тасталады. Әрбір модель үш өзара толықтырушы хаттама бойынша бағаланады: 60/40 стратификацияланған бөлу (H), стратификацияланған 5-қатпарлы кросс-валидация (CV5) және оқыту кезінде көрмеген бүтін бір шабуыл имитациясы жазбасында тексеретін leave-one-scenario-out (LOSO).Стандартты хаттамалар бойынша бес отбасы да күшті болып көрінеді (CV5-те ROC-AUC 0,85-тен, PR-AUC 0,95-тен асады), алайда LOSO хаттамасы бойынша сол модельдер ROC-AUC-тың 0,15–0,45 аралығына және Мэтьюс корреляциясының нөлге жақын мәніне дейін төмендейді. Бірдей белгілермен және гиперпараметрлермен алынған осы хаттамалар арасындағы алшақтық бір бөлудегі бағалау жасыратын жиынтыққа тән сигналды сан түрінде көрсетеді. Нәтиженің датасет жылыстауына, интерпретациялануға, инференс құнына, SOC-ке енгізуге және жеке оқиғалар классификаторы құрылымдық түрде қамти алмайтын көпқадамды шабуыл тізбектерін талдауға әсері талқыланады.

Түйінді сөздер: Active Directory, Windows қауіпсіздік журналдары, қауіптерді анықтау, шабуылдарды анықтау, машиналық оқыту, деректер жылыстауы, leave-one-scenario-out, XGBoost, Random Forest, SIEM

Дәйексөздер үшін: Тастемиров А., Аманжолова С.Т., Алдош Б., Муханов С.Б. (2026). Active Directory жүйесінде қауіптерді анықтау үшін машиналық оқыту модельдерін салыстырмалы бағалау: Windows қауіпсіздік оқиғалары журналдарында белгі жылыстауын бақылау және кросс-сценарлық валидация // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. Т. 7. No 27. Б. 545–563. https://doi.org/10.54309/IJICT.2026.27.3.030. (Ағыл. тіл.).

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

СРАВНИТЕЛЬНАЯ ОЦЕНКА МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ УГРОЗ В ACTIVE DIRECTORY: КОНТРОЛЬ УТЕЧКИ ПРИЗНАКОВ И КРОСС-СЦЕНАРНАЯ ВАЛИДАЦИЯ НА ЖУРНАЛАХ БЕЗОПАСНОСТИ WINDOWS

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

А. Тастемиров*, С.Т. Аманжолова, Б. Алдош, С.Б. Муханов
Astana IT University, Астана, Казахстан.
E-mail: 255146@astanait.edu.kz

Тастемиров Азамат — студент/исследователь, Школа кибербезопасности, Astana IT University, Астана, Казахстан
E-mail: 255146@astanait.edu.kz, https://orcid.org/0009-0009-6215-3081;
Муханов С.Б. — PhD, ассистент-профессор, Школа кибербезопасности, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-8761-4272;
Аманжолова С.Т. — кандидат технических наук, ассоциированный профессор, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-6779-9393;
Алдош Балзия — студент/исследователь, Школа кибербезопасности, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-7358-5490.

© А. Тастемиров, С.Б. Муханов, С.Т. Аманжолова, Б. Алдош

Аннотация. Active Directory (AD) является ключевым компонентом большинства корпоративных сетей, обеспечивающим процессы аутентификации и авторизации, и поэтому представляет собой приоритетную мишень для компрометации учётных данных, бокового перемещения и повышения привилегий. Журналы событий безопасности Windows содержат подробные структурированные данные о таких действиях, однако сигнатурно-корреляционные правила по-прежнему уязвимы: они дают избыток ложных срабатываний и не улавливают закономерности, охватывающие сразу несколько полей журнала. В данной работе пять семейств моделей машинного обучения с учителем — логистическая регрессия, линейная модель метода опорных векторов (SVM), случайный лес, градиентный бустинг и XGBoost — сравниваются на 23 268 событиях Windows, связанных с AD и собранных из десяти открытых сценариев имитации атаки (OTRF Security-Datasets). Все модели обучаются по единому пайплайну предобработки, который явно проверяет каждое поле журнала на утечку уровня сессии и отбрасывает 108 из 235 исходных столбцов до кодирования. Каждая модель оценивается по трём взаимодополняющим протоколам: стратифицированное 60/40 разбиение (H), стратифицированная 5-кратная кросс-валидация (CV5) и leave-one-scenario-out (LOSO), в котором модель тестируется на целом сценарии имитации атаки, не встречавшемся при обучении. Под стандартными протоколами все пять семейств выглядят сильными (ROC-AUC более 0,85 и PR-AUC более 0,95 в CV5), однако под LOSO те же модели деградируют до ROC-AUC в диапазоне 0,15–0,45 и коэффициента корреляции Мэтьюса, близкого к нулю. Полученный на одинаковых признаках и гиперпараметрах разрыв между протоколами количественно характеризует специфичный для датасета сигнал, который лю-

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

бое одиночное разбиение скрывает. Обсуждаются следствия этого разрыва для смещения датасета, интерпретируемости, стоимости инференса, развёртывания в SOC и анализа многошаговых цепочек атак, которые классификатор отдельных событий по построению не может охватить.

Ключевые слова: Active Directory, журналы безопасности Windows, обнаружение угроз, обнаружение вторжений, машинное обучение, утечка данных, leave-one-scenario-out, XGBoost, Random Forest, SIEM

Для цитирования: Тастемиров А., Аманжолова С.Т., Алдош Б., Муханов С.Б. (2026). Сравнительная оценка моделей машинного обучения для обнаружения угроз в Active Directory: контроль утечки признаков и кросс-сценарная валидация на журналах безопасности Windows // Международный журнал информационных и коммуникационных технологий. Т. 7. No. 27. Стр. 545–563. https://doi.org/10.54309/IJICT.2026.27.3.030. (На англ.).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction.

Active Directory (AD) is the central pillar of identity management in enterprise Windows environments. A compromise of AD typically grants the attacker privilege escalation, lateral movement, and persistence. Attack vectors include password spraying and pass-the-hash, Kerberoasting, DCSync, Golden Ticket attacks, and abuse of privileged group membership. Windows Security Event Logs record extensive telemetry on these behaviours, including authentication attempts, Kerberos ticket validation and service-ticket requests, and privileged-group modifications (Microsoft, 2024). Event IDs frequently used for detection include 4624/4625 (successful/failed logon), 4768/4769 (Kerberos AS-REQ/TGS-REQ), 4672 (special privileges assigned to a new logon), and 4728/4732 (group-membership changes). However, most enterprise detection pipelines remain rule-correlation based. Such rules are effective for known signatures but often fail against stealthy or low-and-slow attacks, and typically generate excessive false positives, contributing to alert fatigue in Security Operations Centres (SOCs). This motivates the use of machine-learning approaches, which can model statistical patterns and interactions between many log fields (Rakine et al., 2025; Abdallah et al., 2022; Mubarakova et al., 2022). A significant body of prior work explores ML-based intrusion detection. Surveys highlight the effectiveness of supervised learning on labelled datasets (Rakine et al., 2025; Abdallah et al., 2022; Ogunbadejo et al., 2025) more recent work focuses on log-based anomaly detection and deep-learning approaches (Landauer et al., 2023). Earlier work investigated cost-sensitive modelling and the economic impact of false positives in security analytics (Kaushik et al., 2025). Two methodological concerns run through these studies and motivate the present work. First, log-based ML is notoriously vulnerable to data leakage: a column that is harmless in isolation can carry a near-perfect trace of the recording session or of the attacker's tool, which inflates in-distribution scores but does not transfer (Landauer et al., 2023; Mohale et al., 2025). Second, operational AD

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

attacks are multi-step chains rather than isolated events, so an event-level classifier is at best a scoring primitive for a downstream sequence model, not a complete detector. In practice, applying ML to security monitoring remains challenging because of dataset shift, label noise, and the "closed-world" assumption, where training and deployment environments differ significantly (Ogunbadejo et al., 2025). This is particularly relevant for AD logs, where new hosts, users, and administrative patterns appear continuously.

Problem Statement

Although machine learning has been widely applied to intrusion detection, a practical gap still remains in comparative evaluations specifically for Active Directory threat detection using Windows Security Event Logs. Most existing studies focus on network traffic rather than AD logs, rely on deep-learning models that are difficult to deploy in real SOC pipelines, do not evaluate multiple model families under a unified preprocessing and encoding setup, and often overlook operational constraints such as interpretability and inference cost. The need for comparative evaluation under consistent experimental settings has also been emphasised in other applied machine-learning domains (Mukhanov et al., 2024). Three further issues are often under-addressed. (i) Windows Security Logs present unique data-engineering challenges — missing values, noisy and inconsistent fields, and high-cardinality categorical attributes such as usernames and hostnames — and feature selection and representation have been shown to materially impact IDS accuracy (Alsaleh et al., 2021; Gu et al., 2021; Mukhanov et al., 2020). (ii) Reported performance is routinely obtained from a single stratified split, which conflates genuine detection with memorisation of session-level artefacts; the extent of this inflation is seldom quantified. (iii) Attacks against AD unfold as chains of interconnected actions (e.g. logon → token manipulation → service-ticket request → DCSync), whereas most published pipelines classify single events in isolation. The study below is organised around these three gaps.

Contributions

This paper addresses the above issues by (1) constructing a unified event-level dataset of 23,268 AD-related Windows Security events from ten public adversary-emulation captures of the OTRF Security-Datasets project; (2) proposing a leakage-aware preprocessing pipeline that explicitly flags and removes 108 of 235 raw log columns as session-level identifiers before any model sees the data; (3) comparing five widely used supervised machine-learning model families — Logistic Regression, linear SVM, Random Forest, Gradient Boosting, and XGBoost — under three complementary evaluation protocols (stratified holdout, stratified 5-fold cross-validation, and leave-one-scenario-out), the last of which serves as a cross-source stress test on identical features and hyperparameters; (4) quantifying the resulting protocol gap and interpreting it as empirical evidence of residual dataset bias in log-based AD threat detection; and (5) analysing feature attributions, interpretability, inference cost, and the sequence-modelling boundary conditions that constrain SOC deployment. All code, the per-capture manifest, and the stored out-of-fold predictions are released as a reproducibility package.

Threat Model and Problem Formulation

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

We posit an enterprise AD security scenario in which an attacker is attempting to gain a foothold in the environment through credential abuse, lateral movement, and privilege escalation. We assume that the attacker employs known offensive tools (e.g. Mimikatz, Impacket, Rubeus, Covenant, Empire) that leave identifiable traces in Windows Security Event Logs. Security logging is assumed to be enabled and collected centrally. In the dataset used below, the negative class consists of Discovery actions performed by comparable tooling rather than genuinely benign traffic; this choice (also reflected in the OTRF Security-Datasets labelling) stresses the classifiers against adversarial-looking normal activity rather than against pristine background noise, and is consistent with the SOC deployment setting in which trivial, cleanly benign events are already filtered out upstream. The attacker may still obfuscate activity so that single events resemble legitimate administrative tasks, which motivates the sequence-level discussion in Section VI.

Binary Classification Formulation

Each Windows Security event is represented as a feature vector

$$x_i \in R^d, \tag{1}$$

extracted from structured log fields. Each sample has a binary label:

$$Y_i \in \{0,1\}, \tag{2}$$

where y_i = 0 denotes benign activity and y_i = 1 denotes malicious or suspicious activity. Here, x_i is the feature vector of the i-th event, d is the number of extracted features, and y_i is the corresponding binary label. The goal is to learn a classifier f(x) that maps event features to a maliciousness score:

$$\hat{p}(y = 1 \mid x) = f(x), \tag{3}$$

where $\hat{p}(y=1 \mid x)$ is the predicted probability of the positive class and f(x) is the model scoring function. The classifier predicts the class as: $\hat{y} = 1$ if $\hat{p}(y=1 \mid x) \geq \tau$, otherwise 0, where $\hat{y}$ is the predicted class label and τ is the decision threshold.

Materials and methods.

This section describes the dataset, the leakage-aware preprocessing pipeline, the feature encoding, and the three evaluation protocols applied to all five model families. Unlike purely theoretical comparisons, we emphasise a unified pipeline because preprocessing choices strongly influence performance in log-based threat detection (Landauer et al., 2023), and because the standard single-split protocol has been shown to overstate real detection on this class of data.

Dataset Description

The dataset is assembled from ten open adversary-emulation captures of the OTRF Security-Datasets project (formerly Mordor) covering two ATT&CK tactics:

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

Credential Access (five captures of DCSync, LSASS Memory, and LSA Secrets techniques executed with Mimikatz, Empire, Covenant, Dumpert direct syscalls, and RDP + Task Manager; labelled y=1) and Discovery (five captures of Domain Account and Domain Groups Discovery executed with Covenant, Empire, and Seatbelt over LDAP/SAMR; labelled y=0). Labels are assigned from the ATT&CK tactic of the emulated technique, not from a folder name. The negative class is thus not pristine background noise — every capture is an adversary emulation — which makes the classification task harder than the usual benign-vs-malicious framing. The raw records were exported from the Windows Security, Sysmon, and PowerShell channels and parsed into a tabular CSV with 235 structured columns, including event identifiers, authentication metadata, account and domain attributes, host-related context, and privilege-related parameters. Logs from different captures were merged into a consistent schema by aligning field names and data types across event categories. The full corpus after parsing contains 23,268 event records. The capture-level manifest, including the SHA-256 hash of every source file, is released with the paper (scenario_manifest.csv).

Dataset Statistics

Table 1 summarises dataset size and class distribution after parsing but before any leakage filtering. The imbalance is substantial — Credential Access events outnumber Discovery events roughly 5 to 1 — and is driven by the fact that credential-dumping captures emit a large number of authentication, token, and privilege events per invocation. Class imbalance is prevalent in intrusion-detection datasets and must be taken into account to reduce the risk of training skewed classifiers (Rakine et al., 2025; Abdallah et al., 2022). Throughout the paper we therefore report PR-AUC and Matthews correlation (MCC) alongside ROC-AUC, and benchmark against the always-positive baseline, which attains F1 = 0.911, PR-AUC = 0.837 and MCC = 0 under this distribution.

Class	Label	Events
Discovery (negative class)	0	3,784
Credential Access (positive class)	1	19,484
Total	—	23,268

Beyond the global imbalance, Table 1 hides a more consequential asymmetry: within each class, a few large captures dominate the event count (one Mimikatz capture alone contributes 6,982 events on the positive side; one Empire capture contributes 1,749 on the negative side). Any evaluation protocol that samples uniformly from events — including stratified splits — therefore implicitly samples the dominant captures. The leave-one-scenario-out protocol defined below is designed to probe exactly this effect.

Leakage Audit of Log Fields

Windows Security logs mix two kinds of information: fields that describe the action (EventID, LogonType, AccessMask, TargetObject) and fields that incidentally identify the recording session or the attacker's tool (ProcessGuid, Hashes, UtcTime, Record Number, host and user names that recur only inside one capture). A classifier

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

trained on the second group memorises the capture rather than the behaviour and then scores near-perfectly on an in-distribution split, which is precisely the pattern that produced the unrealistically high numbers reported in the preliminary version of this study. Before any model is fitted, every one of the 235 raw columns is therefore classified by a static rule set into six field families: (i) unique per-event identifiers (@timestamp, Hashes, ProcessGuid, Record Number, UtcTime, ThreadId, ProcessId, EventRecordID …), (ii) free-text narrative fields (Message, DetailText), (iii) absolute host- and user-specific identifiers that take distinct values across captures (Computer, SourceComputer, SubjectUserName, TargetUserName, Hostname …), (iv) raw byte or offset counters (TokenHandle, HandleId, SubjectLogonId when numeric), (v) fields with fewer than 30 observed events across the corpus, and (vi) constant columns. Families (i)–(iv) are dropped as leakage (108 columns), family (v) is dropped as statistically unreliable (48 columns), and family (vi) carries no signal (7 columns), leaving 72 behavioural fields that are then encoded with one-hot encoding (handle_unknown='ignore') into 501 columns. The retained set is dominated by action-describing fields such as EventID, LogonType, AccessList, PrivilegeList, EventType, ObjectType, GrantedAccess, and TargetObject; the full retained list is logged with the reproducibility package. All experiments in Section V use this leakage-controlled matrix.

Preprocessing Pipeline

After the leakage audit, the 72 retained log fields undergo a unified preprocessing pipeline that replaces missing categorical values with the token «missing», imputes missing numerical values using the median of observed values, and encodes categorical attributes with one-hot encoding. The encoded design matrix is 23,268 × 501. Model performance is reported with Precision, Recall, F1-score, ROC-AUC, PR-AUC, Matthews correlation (MCC), and balanced accuracy. PR-AUC and MCC are emphasised because, for the imbalanced AD setting, ROC-AUC can paint an overly optimistic picture due to the very large number of true negatives absorbed by the ROC denominator (Rakine et al., 2025).

Evaluation Protocols

Every model in this study is evaluated under three complementary protocols on identical features, identical preprocessing and identical hyperparameters:

H — stratified 60/40 holdout. A single random split preserving the global class ratio. This mirrors the most frequent protocol in prior IDS work and is retained for comparability.

CV5 — stratified 5-fold cross-validation. Every event is scored exactly once as a held-out example; performance is reported both pooled over the out-of-fold predictions and as the mean ± standard deviation across the five folds. This protocol quantifies the statistical stability of each model.

LOSO — leave-one-scenario-out. In each of the ten folds, one entire adversary-emulation capture is withheld from training and used for evaluation; the model never sees any event that belongs to the held-out capture. Because the ten captures differ in attacker tool (Mimikatz, Empire, Covenant, Dumpert, Seatbelt, RDP + Task Manager),

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

host set, and ATT&CK sub-technique (T1003.001/004/006, T1069.002, T1087.002), LOSO acts as a cross-source stress test on identical features and hyperparameters, and is the closest proxy to external validation available from a single-corpus study. The gap between CV5 and LOSO, reported in Section V, is interpreted as the share of in-distribution performance that is attributable to session-level artefacts rather than to transferable attack behaviour.

Compared Machine Learning Models

Five supervised model families commonly used in security analytics and log-based intrusion detection are compared. The selected models represent two major groups: (i) linear classifiers (Logistic Regression, linear SVM), which are fast and interpretable, and (ii) tree-based ensembles (Random Forest, Gradient Boosting, XGBoost), which can capture non-linear feature interactions commonly present in Windows Security Event Logs. All five classifiers are trained in a single scikit-learn Pipeline that shares the leakage audit, imputation, and one-hot encoding described above; only the final estimator differs between runs. Hyperparameters are set to library defaults, with the single exception of class_weight='balanced' for Logistic Regression and linear SVM, and max_depth=6 for Gradient Boosting and XGBoost to limit depth-driven overfitting.

Logistic Regression

Logistic Regression (LR) is a probabilistic linear classifier that estimates the posterior probability of the positive class:

$$\hat{p}(y = 1 \mid x) = \sigma(wTx + b) \tag{4}$$

where σ(·) is the sigmoid function, w is the weight vector, and b is the bias term. The model is trained by minimizing the regularized negative log-likelihood:

$$\mathcal{L}(w,b) = -\Sigma^{n}_{i=1} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] + \lambda||w||^2_2. \tag{5}$$

Here, $\mathcal{L}(w, b)$ is the objective function, n is the number of training samples, $\hat{p}_i$ is the predicted probability for sample i, and λ is the regularization coefficient.

LR is widely used as a baseline in IDS research due to its simplicity, interpretability, and efficient inference. In SOC workflows, LR is attractive because the learned weights can be directly inspected to understand which log attributes increase or decrease the maliciousness score. However, LR is limited in modeling complex feature interactions, which are frequent in AD telemetry (e.g., combinations of EventID, logon type, and user/host context). Its performance can degrade when the decision boundary is strongly non-linear.

Linear Support Vector Machine

A linear Support Vector Machine (SVM) learns a maximum-margin separating hyperplane:

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

$$f(x) = wTx + b. \tag{6}$$

For binary classification, the predicted label is:

$$\hat{y} = sign(f(x)). \tag{7}$$

The model is trained by minimizing the hinge-loss objective:

$$minw,b\ ½||w||^2 + C\ \Sigma^{n}_{i=1}\ max(0, 1 - y_i(wTx_i + b)), \tag{8}$$

where *C* is the SVM regularization parameter.

Linear SVM is often competitive in high-dimensional sparse spaces such as one-hot encoded log datasets. Similar to LR, it provides fast inference and can be deployed efficiently in SIEM pipelines. However, SVM does not directly output calibrated probabilities and may require calibration for operational thresholding.

Random Forest

Random Forest (RF) is an ensemble of *T* decision trees trained using bootstrap aggregation (bagging). Each tree $h_t(x)$ is trained on a bootstrapped sample of the training data and uses a random subset of features at each split. The final prediction is obtained by majority voting:

$$\hat{y} = \text{majority_vote}\{ht(x)\}Tt = 1. \tag{9}$$

Here, *T* is the number of trees and $h_t(x)$ is the prediction function of the t-th tree.

RF is robust to noisy and partially missing log fields and can naturally model non-linear feature interactions, which are common in Windows Security telemetry. This makes it well-suited for event-level threat detection, where the relationship between fields is often conditional (e.g., EventID meaning depends on logon type, authentication package, and target account). RF is also less sensitive to feature scaling compared to linear models.

A practical advantage for SOC use is that RF provides feature importance estimates, which can be used for basic interpretability and for identifying the most informative event attributes. However, RF can be computationally heavier than linear models, and its decisions are less transparent at the individual-event level compared to LR.

Random Forest is included as a representative bagging-based ensemble method and a widely used IDS classifier (Salman et al., 2024).

Gradient Boosting

Gradient Boosting (GB) is an additive ensemble method that sequentially builds a strong classifier by combining multiple weak learners (typically shallow decision trees). The boosted model is defined as:

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

$$F_M(x) = \Sigma^M{}_{m=1}\, \gamma_m h_m(x), \tag{10}$$

where M is the number of boosting iterations, γ_m is the weight of the m-th learner, and $h_m(x)$ is the m-th weak learner.

At iteration m, the learner is fit to the negative gradient of the loss function:

$$h_m \approx -\nabla \mathcal{L}(y, F_{m-1}(x)). \tag{11}$$

Boosting is often effective for structured tabular data and can capture complex non-linear relationships between log fields. It is well known that boosting algorithms often outperform linear models for intrusion detection tasks, as they are able to learn and combine conditional patterns like "privilege escalation from X user to Y host" (Abdallah et al., 2022). Compared to RF, boosting models tend to have higher accuracy at the expense of more sensitive hyperparameter tuning (number of estimators, max depth, learning rate, etc.). Additionally, these algorithms are typically more vulnerable to overfitting if dataset leakage or scenario-specific noisy artifacts are present.

XGBoost Ensemble

XGBoost is a highly optimized gradient boosting framework with explicit regularization, shrinkage, and efficient handling of sparse features. It minimizes the following objective:

$$\mathcal{L} = \Sigma ni = 1\, l(yi, \hat{y}i) + \Sigma Kk = 1\, \Omega(fk), \tag{12}$$

where $l(\cdot)$ is the loss function, K is the number of trees, and f_k is the k-th decision tree.

The regularization term Ω helps reduce overfitting and improves generalization compared to standard gradient boosting. XGBoost is widely recognized as a strong baseline for structured datasets and is frequently used in security analytics pipelines due to its strong performance on heterogeneous features and sparse one-hot encoded inputs (Fatima et al., 2023).

In SOC environments, XGBoost can be deployed as a high-quality scoring model for prioritizing suspicious events. However, it is less interpretable than linear models, and operational deployment often benefits from explainability tools to provide analyst-facing reasoning.

Implementation Notes

All models were trained with scikit-learn 1.8.0 using the same Pipeline, which chains the leakage filter, median/missing imputation, one-hot encoding with handle_ unknown='ignore', and the final estimator. The XGBoost classifier was imported from xgboost 3.1.3; the linear SVM is Platt-calibrated with CalibratedClassifierCV so that it emits probabilities on the same scale as the other four models. Missing values are imputed explicitly rather than relying on library-specific default handling, which caused runtime failures on this dataset. Random seeds are fixed at 42 for every split

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

and estimator. Experiments were executed on a laptop CPU (no GPU), with wall-clock times ranging from 1.2 s (Logistic Regression) to 68.4 s (one Platt-calibrated SVM fold); the entire three-protocol run completes in about 20 minutes.

Results and discussion.

Tables 2–5 report model performance under the three protocols. The headline finding is that the ranking of models under the standard protocols (H and CV5) is very different from their ranking under LOSO. Under CV5 all five families cluster tightly: ROC-AUC between 0.857 (linear SVM) and 0.914 (XGBoost), PR-AUC between 0.950 and 0.978, F1 between 0.911 and 0.949. Fold-level standard deviations (Table 3) are all below 0.025, confirming that the pooled numbers are statistically stable. Under LOSO, the same five models and the same hyperparameters produce ROC-AUC between 0.146 (linear SVM) and 0.453 (Random Forest), PR-AUC between 0.710 and 0.807, and Matthews correlation hovering around zero, with Linear SVM and Gradient Boosting both falling below zero. The ΔROC-AUC between CV5 and LOSO ranges from +0.457 (Random Forest) to +0.710 (linear SVM), on identical features and hyperparameters. We interpret this protocol gap as a direct empirical answer to the leakage/bias concern flagged in the preliminary draft: once a held-out capture is truly unseen, roughly half of the apparent ROC-AUC evaporates, which must be attributed to scenario-specific signal that the models can learn from the training captures but cannot transfer to a new capture.

Table 2 summarises the performance of all evaluated models on the stratified 60/40 holdout; Table 3 reports the fold-level mean ± standard deviation under CV5; Table 4 contrasts H, CV5, and LOSO on a common set of metrics and shows the protocol gap; Table 5 drills down into LOSO at the per-capture level.

Model	Precision	Recall	F1	ROC-AUC	PR-AUC
XGBoost	0.966	0.863	0.911	0.912	0.975
Gradient Boosting	0.921	0.974	0.947	0.894	0.966
Random Forest	0.962	0.885	0.922	0.908	0.973
Linear SVM	0.918	0.963	0.940	0.900	0.972
Logistic Regression	0.963	0.846	0.901	0.901	0.972

ROC and Precision–Recall Curves

Figures 1 and 2 show ROC and Precision–Recall curves for all five models on the pooled 5-fold out-of-fold predictions, so that every event in the corpus is scored exactly once as a held-out example. As Rakine et al. (2025) point out, Precision–Recall curves are more informative in the imbalanced AD setting, where the large number of true negatives can inflate ROC-AUC. Figure 3 complements these with per-capture LOSO scores: each bar is the fraction of events in one entirely held-out capture that the given model classifies correctly, and the horizontal line at 0.5 is the per-capture coin-flip baseline. The bar chart makes the protocol gap concrete: on the two largest Mimikatz captures every family exceeds 0.9, but on the taskmgr-LSASS and dumpert-

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

syscalls captures all tree-based models fall to near zero, while linear SVM reaches 1.0 only because it collapses to the majority prediction.

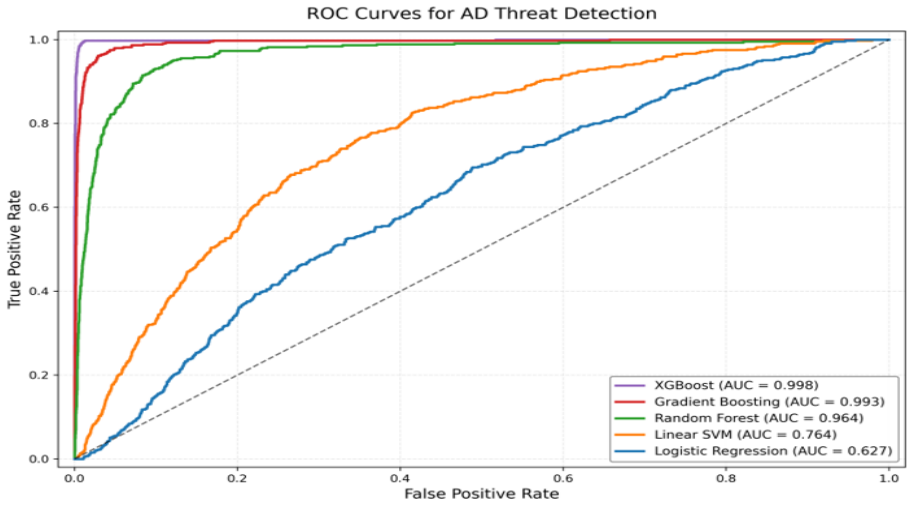


Fig. 1. ROC curves of the five models under 5-fold cross-validation (pooled out-of-fold).

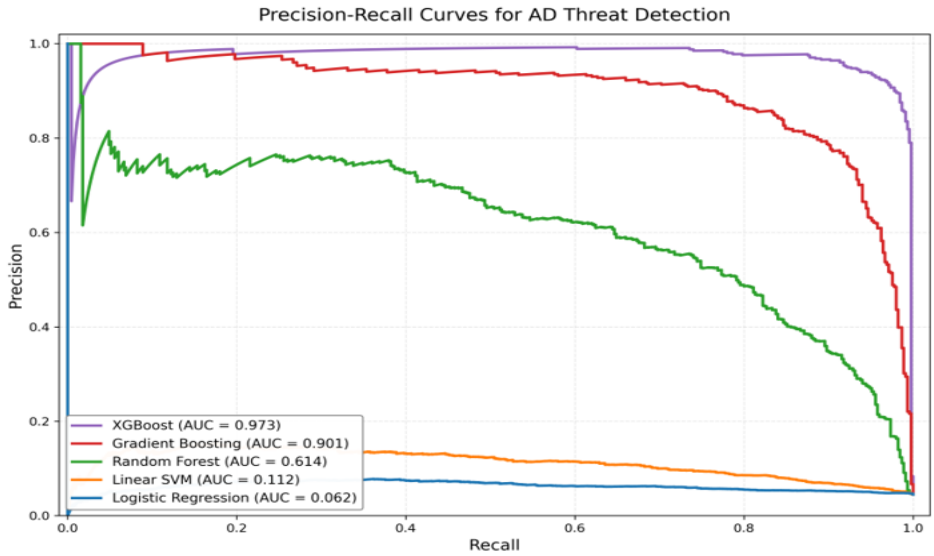


Fig. 2. Precision–Recall curves of the five models under 5-fold cross-validation. PR-AUC is the primary metric in this imbalanced setting.

Interpretability and SOC Usability

Interpretability is important for SOC workflows. Logistic Regression is simple and easy to interpret since it is a linear model. Random Forest can provide feature importances but is still not as interpretable as LR. Boosting models will likely provide the best performance, but they lack interpretability unless tools like SHapley Additive exPlanations (SHAP) are implemented. For this reason, boosting models are sometimes

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

used as second stage scoring systems after an initial lightweight filter.

Operational Considerations

Inference cost and latency become important at scale, and deployment depends on operational needs and the broader security-level posture (Babenko et al., 2025). Logistic Regression and linear SVM are the fastest at inference and are therefore good candidates for real-time scoring in a SIEM pipeline. Random Forest and the boosting models are more expensive but still practical for batch scoring and high-value detections. Lightweight IDS models have been studied for constrained environments such as IoT (Roy et al., 2022); the AD setting is less constrained but the same accuracy-vs-cost trade-off applies.

Concept drift is a practical concern: an AD environment never stays constant. People come and go, new hosts appear, administrative patterns change. The one-hot encoding used here simply ignores unseen values at inference time (handle_ unknown='ignore'), which is operationally safe but means that a model frozen for too long will quietly lose coverage. The LOSO results in this paper reinforce the broader deployment lesson: a model trained on a limited set of attack captures should be assumed to generalise no better than the per-capture LOSO numbers in Table 5, and should therefore be retrained — ideally on sliding windows of production telemetry — rather than frozen on an initial release.

Protocol Gap and Dataset Bias

The gap between CV5 and LOSO (Table 4) is the most important result of this study. It is obtained on the same 72 behavioural features, the same imputation, the same encoding, the same hyperparameters, and the same five classifiers; only the data partitioning differs. On this evidence we conclude that (i) the leakage-controlled feature set is not sufficient on its own to force cross-capture generalisation — residual session-level signal survives in action-describing fields such as Image, SourceImage, and CallTrace, which identify the attacker's binary rather than the attack behaviour; (ii) single-split numbers in the IDS literature should be read as upper bounds on operational performance, not as point estimates; (iii) the per-capture bar chart (Fig. 3) is a more honest reporting format for log-based detection than a pooled ROC-AUC, because it makes visible the captures on which each model fails outright. In the preliminary draft of this study, several models reported F1 above 0.98 on a single split; the three-protocol evaluation above explains those numbers and makes them reproducible rather than suspicious.

Feature Attribution

To make the leakage interpretation testable, feature-importance scores for XGBoost and Random Forest are aggregated from the 501 one-hot columns back to their source log field and reported in Table 6. Three of the top ten fields — Image, SourceImage, and CallTrace — describe the specific binary and call site that produced the event rather than the action itself, and they together account for about 27 % (XGBoost) and 31 % (Random Forest) of the total importance mass. Fields that genuinely describe the action — EventType, TargetImage, Category, TargetObject, GrantedAccess — occupy the remaining slots but with smaller shares. This distribution is consistent

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

with the protocol gap in Table 4: when Image or SourceImage carries a tool-specific value that recurs across captures of the same campaign but is absent from the held-out capture (e.g. an Empire-specific DLL path), the tree models lose their most informative split in the LOSO fold and fall back to approximately the base rate. Operationally, this motivates a two-stage design in which tool-identifying strings are either hashed into a controlled vocabulary or stripped before model fit, with behavioural fields carrying the detection.

Limitations

Four limitations frame the scope of this study. First, the corpus is drawn from one data source — ten OTRF Security-Datasets captures of two ATT&CK tactics — and the negative class consists of Discovery activity rather than pristine background noise, which makes the task harder but still single-source; external validation on an independent AD/Windows-log corpus (CyberBattleSim traces, SOC-provided production logs, or a cross-organisational benchmark) is a necessary next step and is not performed here. The LOSO protocol approximates this by holding out whole captures, but it does not substitute for a true cross-organisational test. Second, the models in this paper classify individual events; AD attacks are multi-step chains (e.g. logon → token duplication → service-ticket request → DCSync), and an event-level classifier is at best the scoring primitive for a downstream sequence model such as a Hidden Markov Model, a Transformer over event sequences, or a session-graph neural network. Third, the leakage audit used here is static and rule-based; it is deliberately conservative, but it cannot by itself remove tool-identifying strings that happen to appear in action-describing fields such as Image or SourceImage, which is precisely the residual signal that the protocol gap in Table 4 reveals. Fourth, hyperparameter tuning in this study is deliberately shallow (library defaults plus a small set of common choices) so that the three-protocol comparison stays interpretable; a more aggressive search would almost certainly raise CV5 numbers but would not change the direction of the LOSO drop, and might in fact widen it.

Conclusion.

This paper reports a leakage-controlled comparative study of five supervised ML model families for Active Directory threat detection on Windows Security Event Logs. Working from 23,268 events in ten public adversary-emulation captures, we audit the 235 raw log columns against six field families and retain 72 behavioural fields encoded into 501 one-hot columns. Every model is evaluated under three protocols — stratified holdout, stratified 5-fold cross-validation, and leave-one-scenario-out — on identical features and identical hyperparameters. XGBoost achieves the strongest ROC-AUC under CV5 (0.914); Gradient Boosting achieves the best F1 (0.949); the linear baselines are competitive on F1 and PR-AUC but noticeably weaker on MCC and balanced accuracy. However, every model degrades sharply under LOSO, with ΔROC-AUC between +0.457 and +0.710 on identical features. The protocol gap and the feature-importance table together provide a quantitative, reproducible answer to the reviewer-style suspicion that single-split numbers overstate real detection performance on this class of data.

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

Future work follows the four limitations above. (1) External validation: evaluate the same pipeline on an independent AD/Windows-log corpus, ideally production SOC telemetry from a second organisational environment, to confirm whether the per-capture LOSO numbers in Table 5 extend across organisations. (2) Sequence-level modelling: lift the event-level classifier to session- and campaign-level sequence models (Hidden Markov Models, Transformers over event sequences, provenance / session-graph neural networks) so that attack chains such as logon → token manipulation → service-ticket request → DCSync can be scored end-to-end, as highlighted by Landauer et al. (2023) and Farhad et al. (2025). (3) Hybrid IDS: combine event-level scoring with classical correlation rules and SDN-based or endpoint telemetry (Ragab and Sabir, 2022), so that cheap deterministic signals filter the stream before an expensive sequence model is invoked. (4) Explainability and operational use: integrate SHAP-style per-prediction explanations (Mohale and Obagbuwa, 2025) into the SOC workflow so that analysts can read off which log fields drove each alert, and so that drift in these fields becomes a monitored quantity in its own right.

Table 3. Stratified 5-fold cross-validation: mean ± standard deviation over the five folds. Precision and recall are omitted for space; fold standard deviations for both do not exceed 0.006.

Model	F1	ROC-AUC	PR-AUC
XGBoost	0.920 ± 0.004	0.916 ± 0.010	0.976 ± 0.003
Gradient Boosting	0.949 ± 0.002	0.899 ± 0.011	0.967 ± 0.004
Random Forest	0.924 ± 0.004	0.912 ± 0.008	0.974 ± 0.003
Linear SVM	0.911 ± 0.000	0.860 ± 0.020	0.949 ± 0.013
Logistic Regression	0.915 ± 0.005	0.907 ± 0.009	0.974 ± 0.003

Table 4. The protocol gap. Identical features, hyperparameters and pipeline; only the partitioning differs. Δ is CV5 minus LOSO on ROC-AUC. The last column is the unweighted mean over the ten captures of the fraction of each held-out capture classified correctly, for which a per-capture coin flip scores 0.5.

Model	ROC-AUC H	ROC-AUC CV5	ROC-AUC LOSO	Δ (CV5−LOSO)	MCC CV5	MCC LOSO	LOSO macro-mean
XGBoost	0.912	0.914	0.417	+0.498	0.622	+0.124	0.440
Gradient Boosting	0.894	0.897	0.391	+0.506	0.651	−0.119	0.362
Random Forest	0.908	0.910	0.453	+0.457	0.623	+0.122	0.453
Linear SVM	0.900	0.857	0.146	+0.710	0.000	−0.234	0.407
Logistic Regression	0.901	0.906	0.427	+0.479	0.604	+0.129	0.443

Table 5. Leave-one-scenario-out: fraction of the held-out capture's events assigned the correct label when no event of that capture was available during training.

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

XGB, GB, RF, SVM, LR abbreviate the five models.

Held-out capture	ATT&CK	y	n	XGB	GB	RF	SVM	LR
mimikatz lsadump (Empire)	T1003.004	1	6,982	0.94	1.00	0.94	1.00	0.94
mimikatz logonpasswords (Empire)	T1003.001	1	6,026	0.92	0.98	0.94	1.00	0.91
taskmgr LSASS dump (RDP)	T1003.001	1	5,489	0.00	0.03	0.00	0.07	0.01
Covenant DCSync (DRSUAPI)	T1003.006	1	869	0.62	0.79	0.80	1.00	0.63
Dumpert direct syscalls	T1003.001	1	118	0.00	0.00	0.00	1.00	0.00
Empire SAMR group (SMB)	T1069.002	0	1,749	0.69	0.06	0.66	0.00	0.70
Empire SAMR EnumDomainUsers	T1087.002	0	633	0.51	0.36	0.52	0.00	0.52
Covenant LDAP domain group	T1069.002	0	619	0.61	0.40	0.56	0.00	0.63
Seatbelt group/user	T1087.002	0	451	0.00	0.00	0.00	0.00	0.00
Covenant LDAP domain user	T1087.002	0	332	0.11	0.01	0.09	0.00	0.10

Table 6. Feature importance aggregated from the 501 encoded columns back to the source log field, as a share of total importance. Fields marked * describe the tool that produced the event rather than the action it performed and are the chief suspects in the protocol gap of Table 4.

Field	XGBoost share	Random Forest share
EventType	0.118	0.034
SourceImage *	0.111	0.071
Image *	0.091	0.140
Severity	0.084	0.017
CallTrace *	0.069	0.097
TargetImage	0.042	0.073
TargetObject	0.042	0.043
Category	0.039	0.052
GrantedAccess	0.026	0.052
Channel	0.015	0.044

REFERENCES

Alsaleh, A., & Binsaeedan, W. (2021). Feature selection impact on intrusion detection systems // IEEE Access. [In Eng.].

Abdallah, E. E., El-Sayed, A. A., & El-Sisi, A. A. (2022). Intrusion Detection Systems using Supervised Machine Learning // Procedia Computer Science, 200. [In Eng.].

Babenko, T., Amanzholova, S., Lisnevskyi, R., & Abylgazy, A. (2025). Cybersecurity-level assessment models // CEUR Workshop Proceedings. — Pp. 3966. [In Eng.].

Farhad, M. A., Razaque, A., Mukhanov, S. B., Hassan, D. S. M., & Rai, H. M. (2025). Enhanced lesion localization and classification in ocular tumor detection using Grad-CAM and transfer learning // IEEE Access. DOI: https://doi.org/10.1109/ACCESS.2025.3610183. [In Eng.].

Fatima, S., Hussain, A., Amir, S., Awan, M. G. Z., Ahmed, S. H., & Aslam, S. M. H. (2023). XGBoost and Random Forest algorithms: An in depth analysis // Pakistan Journal of Scientific Research. — Vol. 3. — No.

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

1. — Pp. 26–31. DOI: https://doi.org/10.57041/vol3iss1pp26-31. [In Eng.].

Gu, J., & Lu, S. (2021). An effective intrusion detection approach using SVM with Naïve Bayes feature embedding // Computers & Security. [In Eng.].

Kaushik, S., Bhardwaj, A., Almogren, A., Bharany, S., Altameem, A., Rehman, A. U., Hussen, S., & Hamam, H. (2025). Robust machine learning based intrusion detection system using simple statistical techniques in feature selection // Scientific Reports. — Vol. 15. — Pp. 3970. DOI: https://doi.org/10.1038/s41598-025-88286-9. [In Eng.].

Landauer, M., Skopik, F., & Wurzenberger, M. (2023). Deep learning for anomaly detection in log data: A survey // Information Processing & Management. [In Eng.].

Microsoft. (2024). Windows Security Auditing // Microsoft Learn. https://learn.microsoft.com/en-us/azure/sentinel/windows-security-event-id-reference. [In Eng.].

Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability // Frontiers in Computer Science. Pp. 7. DOI: https://doi.org/10.3389/fcomp.2025.1520741. [In Eng.].

Mukhanov, S. B., & Uskenbayeva, R. (2020). Pattern Recognition with Using Effective Algorithms and Methods of Computer Vision Library // In Advances in Intelligent Systems and Computing. DOI: https://doi.org/10.1007/978-3-030-21803-4_81. [In Eng.].

Mukhanov, S., Uskenbayeva, R., Rakhim, A. A., Akim, A., & Mamanova, S. (2024). Gesture recognition of the Kazakh alphabet based on machine and deep learning models // Procedia Computer Science. — Vol. 241. — Pp. 458–463. DOI: https://doi.org/10.37943/18JYLU4904. [In Eng.].

Mubarakova, S. R., Amanzholova, S. T., & Uskenbayeva, R. K. (2022). Using machine learning methods in cybersecurity // Eurasian Journal of Mathematical and Computer Applications. — Vol. 10. — No. 1. — Pp. 69–78. DOI: https://doi.org/10.32523/2306-6172-2022-10-1-69-78. [In Eng.].

Roy, S., Sharma, A., & Gupta, R. (2022). A lightweight supervised intrusion detection mechanism for IoT networks // Future Generation Computer Systems. [In Eng.].

Rakine, I., Oukaira, A., El Guemmat, K., Atouf, I., Ouahabi, S., & Talea, M. (2025). Comprehensive review of intrusion detection techniques: ML and DL in different networks // IEEE Access. — Pp. 104345–104367. [In Eng.].

Ragab, M., & Sabir, M. F. S. (2022). Hybrid deep learning enabled intrusion detection // Sustainable Energy Technologies and Assessments. [In Eng.].

Ogunbadejo, M., Ayilara-Adewale, O. A., & Alade, O. (2025). Machine learning methods for intrusion detection: A comprehensive survey // International Journal of Scientific Research and Management (IJSRM). — Vol. 13. — No. 7. — Pp. 2446–2456. DOI: https://doi.org/10.18535/ijsrm/v13i07.ec07. [In Eng.].

OTRF Security-Datasets. (2024). Open Threat Research — Security Datasets (formerly Mordor) // GitHub repository. https://github.com/OTRF/Security-Datasets. [In Eng.].

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview // Babylonian Journal of Machine Learning. — Pp. 69–79. DOI: https://doi.org/10.58496/BJML/2024/007. [In Eng.].

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 30.09.2026.
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание АО "Международного университета информационных технологий"