ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№1
2026

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

1 (357)

JANUARY – MARCH 2026

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Chief Editor:

MUTANOV Galimkair Mutanovich, doctor of technical sciences, professor, academician of NAS RK, (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/1423665

EDITORIAL BOARD:

KALIMOLDAYEV Maksat Nuradilovich, (Deputy Editor-in-Chief), Doctor of Physical and Mathematical Sciences, Professor, Academician of NAS RK, Advisor to the General Director of the Institute of Information and Computing Technologies of the CS MES RK, Head of the Laboratory (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=56153126500, https://www.webofscience.com/wos/author/record/2428551

MAMYRBAEV Orken Zhumazhanovich, (Academic Secretary), PhD in Information Systems, Deputy Director for Science of the Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=55967630400, https://www.webofscience.com/wos/author/record/1774027

BAIGUNCHEKOV Zhumadil Zhanabaevich, Doctor of Technical Sciences, Professor, Academician of NAS RK, Institute of Cybernetics and Information Technologies, Department of Applied Mechanics and Engineering Graphics, Satbayev University (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=6506823633, https://www.webofscience.com/wos/author/record/1923423

WOICIK Waldemar, Doctor of Technical Sciences (Phys.-Math.), Professor of the Lublin University of Technology (Lublin, Poland), https://www.scopus.com/authid/detail.uri?authorId=7005121594, https://www.webofscience.com/wos/author/record/678586

SMOLARJ Andrej, Associate Professor Faculty of Electronics, Lublin polytechnic university (Lublin, Poland), https://www.scopus.com/authid/detail.uri?authorId=56249263000, https://www.webofscience.com/wos/author/record/1268523

KEILAN Alimkhan, Doctor of Technical Sciences, Professor (Doctor of science (Japan)), chief researcher of Institute of Information and Computational Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=8701101900, https://www.webofscience.com/wos/author/record/1436451

KHAIROVA Nina, Doctor of Technical Sciences, Professor, Chief Researcher of the Institute of Information and Computational Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=37461441200, https://www.webofscience.com/wos/author/record/1768515

OTMAN Mohamed, PhD, Professor of Computer Science Department of Communication Technology and Networks, Putra University Malaysia (Selangor, Malaysia), https://www.scopus.com/authid/detail.uri?authorId=56036884700, https://www.webofscience.com/wos/author/record/747649

NYSANBAYEVA Saule Yerkebulanovna, Doctor of Technical Sciences, Associate Professor, Senior Researcher of the Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=55453992600, https://www.webofscience.com/wos/author/record/3802041

USATOVA Olga Alexandrovna, PhD, Associate Professor, Chief Scientific Secretary of the Institute of Information and Computing Technologies of the National Academy of Sciences of the Republic of Kazakhstan (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=57204581062, https://www.webofscience.com/wos/author/record/JCO-3058-2023

KAPALOVA Nursulu Aldazharovna, Candidate of Technical Sciences, Head of the Laboratory cybersecurity, Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=57191242124,

KOVALYOV Alexander Mikhailovich, Doctor of Physical and Mathematical Sciences, Academician of the National Academy of Sciences of Ukraine, Institute of Applied Mathematics and Mechanics (Donetsk, Ukraine), https://www.scopus.com/authid/detail.uri?authorId=7202799321, https://www.webofscience.com/wos/author/record/38481396

MIKHALEVICH Alexander Alexandrovich, Doctor of Technical Sciences, Professor, Academician of the National Academy of Sciences of Belarus (Minsk, Belarus), https://www.scopus.com/authid/detail.uri?authorId=7004159952, https://www.webofscience.com/wos/author/record/46249977

TIGHINEANU Ion Mihailovich, Doctor of Physical and Mathematical Sciences, Academician, President of the Academy of Sciences of Moldova, Technical University of Moldova (Chisinau, Moldova), https://www.scopus.com/authid/detail.uri?authorId=7006315935, https://www.webofscience.com/wos/author/record/524462

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

© «Central Asian Academic Research Center» LLP, 2026

БАС РЕДАКТОР:

МҰТАНОВ Ғалымқайыр Мұтанұлы, техника ғылымдарының докторы, профессор, ҚР ҰҒА академигі, (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/1423665

РЕДАКЦИЯ АЛҚАСЫ:

ҚАЛИМОЛДАЕВ Мақсат Нұрәділұлы, (бас редактордың орынбасары), физика-математика ғылымдарының докторы, профессор, ҚР ҰҒА академигі, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты» бас директорының кеңесшісі, зертхана меңгерушісі (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=56153126500, https://www.webofscience.com/wos/author/record/2428551

МАМЫРБАЕВ Өркен Жұмажанұлы (ғалым хатшы), Ақпараттық жүйелер саласындағы техника ғылымдарының (PhD) докторы, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты» директорының ғылым жөніндегі орынбасары (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=55967630400, https://www.webofscience.com/wos/author/record/1774027

БАЙГУНЧЕКОВ Жұмаділ Жаңабайұлы, техника ғылымдарының докторы, профессор, ҚР ҰҒА академигі, Кибернетика және ақпараттық технологиялар институты, Қолданбалы механика және инженерлік графика кафедрасы, Сәтбаев университеті (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=6506823633, https://www.webofscience.com/wos/author/record/1923423

ВОЙЧИК Вальдемар, техника ғылымдарының докторы (физ-мат), Люблин технологиялық университетінің профессоры (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=7005121594, https://www.webofscience.com/wos/author/record/678586

СМОЛАРЖ Анджей, Люблин политехникалық университетінің электроника факультетінің доценті (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=56249263000, https://www.webofscience.com/wos/author/record/1268523

КЕЙЛАН Әлімхан, техника ғылымдарының докторы, профессор (ғылым докторы (Жапония)), ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институтының» бас ғылыми қызметкері (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=8701101900, https://www.webofscience.com/wos/author/record/1436451

ХАЙРОВА Нина, техника ғылымдарының докторы, профессор, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институтының» бас ғылыми қызметкері (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=37461441200, https://www.webofscience.com/wos/author/record/1768515

ОТМАН Мохаммед, PhD, Информатика, Коммуникациялық технологиялар және желілер кафедрасының профессоры, Путра университеті Малайзия (Селангор, Малайзия), https://www.scopus.com/authid/detail.uri?authorId=56036884700, https://www.webofscience.com/wos/author/record/747649

НЫСАНБАЕВА Сауле Еркебұланқызы, техника ғылымдарының докторы, доцент, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институтының» аға ғылыми қызметкері (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=55453992600, https://www.webofscience.com/wos/author/record/3802041

УСАТОВА Ольга Александровна, PhD, қауымдастырылған профессор, ҚР ҒЖБМ "Ақпараттық және есептеу технологиялары институтының" бас ғалым хатшысы (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=57204581062, https://www.webofscience.com/wos/author/record/JCO-3058-2023

КАПАЛОВА Нұрсұлу Алдажарқызы, техника ғылымдарының кандидаты, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты», Киберқауіпсіздік зертханасының меңгерушісі (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=57191242124,

КОВАЛЕВ Александр Михайлович, физика-математика ғылымдарының докторы, Украина Ұлттық Ғылым академиясының академигі, Қолданбалы математика және механика институты (Донецк, Украина), https://www.scopus.com/authid/detail.uri?authorId=7202799321, https://www.webofscience.com/wos/author/record/38481396

МИХАЛЕВИЧ Александр Александрович, техника ғылымдарының докторы, профессор, Беларусь Ұлттық Ғылым академиясының академигі (Минск, Беларусь), https://www.scopus.com/authid/detail.uri?authorId=7004159952, https://www.webofscience.com/wos/author/record/46249977

ТИГИНЯНУ Ион Михайлович, физика-математика ғылымдарының докторы, академик, Молдова Ғылым Академиясының президенті, Молдова техникалық университеті (Кишинев, Молдова), https://www.scopus.com/authid/detail.uri?authorId=7006315935, https://www.webofscience.com/wos/author/record/524462

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).
Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.
Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*
Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
http://www.physico-mathematical.kz/index.php/en/

© «Орталық Азия академиялық ғылыми орталығы» ЖШС, 2026

Главный редактор:

МУТАНОВ Галимкаир Мутанович, доктор технических наук, профессор, академик НАН РК, (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/1423665

Редакционная коллегия:

КАЛИМОЛДАЕВ Максат Нурадилович, (заместитель главного редактора), доктор физико-математических наук, профессор, академик НАН РК, советник генерального директора «Института информационных и вычислительных технологий» КН МНВО РК, заведующий лабораторией (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=56153126500, https://www.webofscience.com/wos/author/record/2428551

МАМЫРБАЕВ Оркен Жумажанович, (ученый секретарь), доктор философии (PhD) по специальности «Информационные системы», заместитель директора по науке РГП «Институт информационных и вычислительных технологий» Комитета науки МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=55967630400, https://www.webofscience.com/wos/author/record/1774027

БАЙГУНЧЕКОВ Жумадил Жанабаевич, доктор технических наук, профессор, академик НАН РК, Институт кибернетики и информационных технологий, кафедра прикладной механики и инженерной графики, Университет Сатпаева (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6506823633, https://www.webofscience.com/wos/author/record/1923423

ВОЙЧИК Вальдемар, доктор технических наук (физ.-мат.), профессор Люблинского технологического университета (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=7005121594, https://www.webofscience.com/wos/author/record/678586

СМОЛАРЖ Анджей, доцент факультета электроники Люблинского политехнического университета (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=56249263000, https://www.webofscience.com/wos/author/record/1268523

КЕЙЛАН Алимхан, доктор технических наук, профессор (Doctor of science (Japan)), главный научный сотрудник РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=8701101900, https://www.webofscience.com/wos/author/record/1436451

ХАЙРОВА Нина, доктор технических наук, профессор, главный научный сотрудник РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=37461441200, https://www.webofscience.com/wos/author/record/1768515

ОТМАН Мохамед, доктор философии, профессор компьютерных наук, Департамент коммуникационных технологий и сетей, Университет Путра Малайзия (Селангор, Малайзия), https://www.scopus.com/authid/detail.uri?authorId=56036884700, https://www.webofscience.com/wos/author/record/747649

НЫСАНБАЕВА Сауле Еркебулановна, доктор технических наук, доцент, старший научный сотрудник РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=55453992600, https://www.webofscience.com/wos/author/record/3802041

УСАТОВА Ольга Александровна, PhD, ассоциированный профессор, Главный ученый секретарь «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=57204581062, https://www.webofscience.com/wos/author/record/JCO-3058-2023

КАПАЛОВА Нурсулу Алдажаровна, кандидат технических наук, заведующий лабораторией кибербезопасности РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=57191242124,

КОВАЛЕВ Александр Михайлович, доктор физико-математических наук, академик НАН Украины, Институт прикладной математики и механики (Донецк, Украина), https://www.scopus.com/authid/detail.uri?authorId=7202799321, https://www.webofscience.com/wos/author/record/38481396

МИХАЛЕВИЧ Александр Александрович, доктор технических наук, профессор, академик НАН Беларуси (Минск, Беларусь), https://www.scopus.com/authid/detail.uri?authorId=7004159952, https://www.webofscience.com/wos/author/record/46249977

ТИГИНЯНУ Ион Михайлович, доктор физико-математических наук, академик, президент Академии наук Молдовы, Технический университет Молдовы (Кишинев, Молдова), https://www.scopus.com/authid/detail.uri?authorId=7006315935, https://www.webofscience.com/wos/author/record/524462

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

© ТОО «Центрально-азиатский академический научный центр», 2026

CONTENTS

COMPUTER SCIENCE

Akhmetova S.T., Yunussova A.A., Alisheva S.S., Olzhataeva B.T., Mussirepova E.B.
Social network data mining for automated offensive language detection.....13

Amanov A.N., Kazbekova G.N., Zhunissov N.M., Abibullayeva A.A., Aben A.B.
Artificial intelligence-based intrusion detection for DDOS attacks in Software Defined Networking.....30

Amanzholova S.T., Ussatova O.A., Mutanov G.M., Mukhanov S.B., Aitmukash D.
Backend architecture of a hybrid blockchain-based academic credential verification system.....52

Amirkhanova G.A., Nurgazy T.N., Amirkhanov B.S., Tokhtassyn M.M., Nurgazy N.N.
Developing a predictive digital twin for a food product based on Edge ML and IoT sensors.....73

Bekarystankyzy A., Ussen D., Kassenkhan A., Chinibayev Y.
Cold-start in educational recommender systems: classical and LLM-Era strategies.....91

Bimoldina Zh., Mussiraliyeva Sh., Bagitova K., Tereikovska L.
Detection of cyber-propaganda content using machine learning and semantic models....106

Chezhimbayeva K.S.
Forecasting key 5G network KPIs using MLP and LSTM neural network models.....129

Dauitbayeva A.O., Konyrbaev N.B., Abildayeva Zh.T., Yessirkepova A.U., Karim N.A.
Development of an application to optimize the process of employment of graduates.....148

Dzhsupbekova G., Othman M., Ordabayeva G.
Comparative analysis of artificial intelligence algorithms to detect network attacks.....167

Issakhov A., Orazmoldayev N., Zharkynbek Y., Abylkassymova A.
Numerical modeling of the spread of viral infection by airborne droplets in confined spaces.....182

Kantureeva M., Omarova G.S., Duisen Z.D., Shekerbek A.A., Tulebayev Y.B.
Application of machine learning methods in forecasting and optimizing the processes of evacuation of people in high-rise buildings.....202

Khusain B., Telmanov M., Khusain A.B., Brodskiy A.R., Sass A.S.
Digital twin of an integrated emission purification and decarbonization system for thermal units.....218

Kulakayeva A., Ashurov A., Zhumazhanov B., Daineko Ye., Zylgara A.
Algorithm for determining the initial orbital parameters of KazeEOSat-1 for deorbiting.....236

Mimenbayeva A.B., Turebayeva R.D., Ospanova T.T., Aruova A.B., Naizagarayeva A.A.
Development and comparative analysis of machine learning models for urban traffic prediction.....253

Naumenko V.V., Mukanova Zh.A., Kiseleva O.V., Maintser D.A., Nerezov A.K.
The use of real-time polling to improve student academic performance.....271

Nazyrova A.E., Kaderkeyeva Z.K., Bekmanova G.T., Milosz M., Lamasheva Zh.
Transformation of education through digital technologies: advancing student academic performance across learning stages.....287

Oralbekova D., Mamyrbayev O., Akhmediyarova A., Kassymova D., Alibiyeva Z.
Development of a multi-level model for text summarization based on pretrained models.....316

Orazbayev B.B., Zhumadillayeva A.K., Kurbangalieva N.B., Yessirkessinov R.Zh., Orazbayeva K.N.
Synthesis of linguistic models for assessing sulfur quality and fuzzy modeling of the sulfur production process.....337

Sarsenbayeva A.K., Rakhimova D.R., Shormakova A.N., Mansurova M.E., Adali E.
Application of semantic methods in the field of legislation: an intellectual system for analysis of agglutinative texts.....354

Serek A., Shoiynbek A., Sharipov K., Kuanyshbay D., Mukhametzhanov A.
Analysis and classification of telephone fraud based on lexical features of speech transcriptions.....373

Shynzhigit B.B., Balabekova M.O., Amangeldy T.T.
Analysis and forecasting of brick product sales using machine learning models.....393

Tokhayeva A.O., Alzhanov A.K., Nezih Önal, Ziyatbekova G.Z., Begalieva K.B.
Formation of students virtualization competencies in higher education based on Proxmox VE.....412

Tukenova L.M., Auyelbekov O.A., Sapakova S.Z., Sametova A.A., Bostanov E.L.
Modelling and optimisation of hybrid power plant operating modes for unmanned aerial vehicles.....430

Yerimbetova A., Berzhanova U., Daiyrbayeva E., Sakenov B., Sambetbayeva M.
Sign language recognition using temporal convolutional network and MediaPipe.....443

Zhukabayeva T.K., Benkhelifa E., Mardenov Y.M., Baumuratova D., Karabayev N.
Decision support for responding to attacks in cyber-physical industrial internet-of-things systems.....461

МАЗМҰНЫ

ИНФОРМАТИКА

Ахметова С.Т., Юнусова А.А., Алишева С.С., Олжатаева Б.Т., Мүсірепова Э.Б.
Әлеуметтік желідегі бейәдеп пікірлерді автоматты анықтауда деректерді интеллектуалды талдау.....13

Аманов А.Н., Казбекова Г.Н., Жунисов Н.М., Абибуллаева А.А., Абен А.Б.
Бағдарламалық жасақтамамен анықталған желідегі DDOS шабуылдары үшін жасанды интеллектке негізделген шабуылдарды анықтау.....30

Аманжолова С.Т., Усатова О.А., Мутанов Г.М., Муханов С.Б., Айтмукаш Д.
Гибридтік блокчейнге негізделген академиялық сенімдік деректерді тексеру жүйесінің бекендік архитектурасы.....52

Амирханова Г.А., Нұрғазы Т.Н., Амирханов Б.С., Нұрғазы Н. Н.
EDGE ML және IOT сенсорлары негізінде азық-түлік өнімінің предиктивті цифрлық егізін әзірлеу.....73

Бекарыстанқызы А., Үсен Д., Қасенхан А., Чинибаев Е.
Білім беру саласындағы ұсынымдық жүйелеріндегі «Cold-start» мәселесі: классикалық әдістер және LLM дәуірінің стратегиялары.....91

Бимолдина Ж.А., Мусиралиева Ш.Ж., Багитова К.Б., Терейковская Л.З
Кибернасихаттық контентті анықтау үшін машиналық оқыту және семантикалық модельдер қолдану.....106

Чежимбаева К.С.
MLP және LSTM нейрондық желі модельдерін қолдана отырып, 5G желісінің негізгі KPI-лерін болжау.....129

Дәуітбаева А.О., Қоңырбаев Н.Б., Әбілдаева Ж.Т., Есіркепова А.У., Кәрім Н.Ә.
Бітіруші түлектердің жұмысқа орналастыру процесін оңтайландыру үшін қосымша әзірлеу.....148

Джусупбекова Г., Othman M., Ордабаева Г.
Жасанды интеллект алгоритмдерін желілік шабуылдарды анықтау үшін салыстырмалы талдау.....167

Исахов А.А., Оразмолдаев Н., Жаркынбек Е., Абылкасымова А.
Ауа тамшылары арқылы вирустық инфекцияның шектеулі кеңістікте таралуын сандық модельдеу.....182

Кантуреева М.А., Омарова Г.С., Дүйсен Ж.Д., Шекербек А.Ә., Тулебаев Е.Б.
Биік ғимараттардағы адамдарды эвакуациялау процестерін болжау және оңтайландыруда машиналық оқыту әдістерін қолдану.....202

Хусаин Б., Тельманов М.М., Хусаин А.Б., Бродский А.Р., Сасс А.С.
Жылу қондырғыларының шығарындыларын кешенді тазалау және
декарбонизациялау жүйесінің цифрлық егізі.....218

Кулакаева А.Е., Ашуров А.Е., Жумажанов Б.Р., Дайнеко Е.А., Зылғара А.Е.
KAZEOSAT-1 ғарыш аппаратының деорбитациясын жүзеге асыру үшін
бастапқы орбиталық параметрлерін анықтау алгоритмі.....236

Мименбаева А.Б., Туребаева А.Д., Оспанова Т.Т., Аруова А.Б.,
Найзағараева А.А.
Қалалық көлік ағынын болжауға арналған машиналық оқыту модельдерін
әзірлеу және салыстырмалы талдау.....253

Науменко В.В., Муканова Ж.А., Киселева О.В., Майнцер Д.А., Нерезов А.К.
Білім алушылардың үлгерімін арттыру үшін real-time
сауалнамаларын қолдану.....271

Назырова А.Е., Кадеркеева З.К., Бекманова Г.Т., Милош М., Ламашева Ж.Б.
Цифрлық білім және студенттердің академиялық жетістіктері:
деңгейлер бойынша білім беруді дамыту.....287

Оралбекова Д., Мамырбаев О., Ахмедиярова А., Қасымова Д.З, Алибиева Ж.,
Алдын ала оқытылған модельдер негізінде мәтінді резюмелеуге
арналған көпдеңгейлі модельді әзірлеу.....316

Оразбаев Б.Б., Жумадиллаева А.К., Курбангалиева Н.Б., Оразбаева К.Н.
Күкірт сапасын бағалаудың лингвистикалық модельдерін синтездеу
және күкіртті өндіру процесін бұлыңғыр модельдеу.....337

Сарсенбаева А.К., Рахимова Д.Р., Шормакова А.Н., Мансурова М.Е., Адали Э.
Семантикалық әдістерді заңнама саласында қолдану: агглютинативті
мәтіндерді талдауға арналған интеллектуалды жүйе.....354

Серек А., Шойынбек А., Шарипов К., Қуанышбай Д., Мухаметжанов А.
Сөйлеу транскрипцияларының лексикалық белгілеріне негізделген телефон
алаяқтықтарын талдау және жіктеу.....373

Шынжігіт Б.Б., Балабекова М.О., Амангелді Т.Т.
Кірпіш өнімдерін сату көлемдерін машиналық оқытуда талдау
және болжамдау.....393

Тохаева А.О., Альжанов А.К., Nezih Ö., Зиятбекова Г.З., Бегалиева К.Б.
PROXMOX VE негізінде жоғары оқу орындарында білім алушыларды
виртуалдандыру құзыреттерін қалыптастыру.....412

Төкенова Л.М., Әуелбеков О.А., Сапақова С., Саметова А.А., Бостанов Е.Л.
Пилотсыз ұшу аппараттарына арналған гибридті электр станцияларының жұмыс режимдерін модельдеу және оңтайландыру.....430

Еримбетова А.С., Бержанова У.Г., Дайырбаева Э.Н., Сәкенов Б.Е., Самбетбаева М.А.
Уақытша конволюциялық желі мен mediaріре көмегімен ым тілін тану.....443

Жукабаева Т.К., Бенхелифа Э., Марденов Е.М., Баумуратова Д., Карабаев Н.
Киберфизикалық өнеркәсіптік интернет заттары жүйелеріндегі шабуылдарға әрекет ету кезінде шешім қабылдауды қолдау.....461

СОДЕРЖАНИЕ

ИНФОРМАТИКА

Ахметова С.Т., Юнусова А.А., Алишева С.С., Олжатаева Б.Т., Мүсірепова Э.Б.
Интеллектуальный анализ данных для автоматического выявления языка ненависти в социальных сетях.....13

Аманов А.Н., Казбекова Г.Н., Жунисов Н.М., Абибуллаева А.А., Абен А.Б.
Обнаружение вторжений на основе искусственного интеллекта для DDoS-атак в программно-определяемых сетях.....30

Аманжолова С.Т., Усатова О.А., Мутанов Г.М., Муханов С.Б., Айтмукаш Д.
Бэкенд-архитектура гибридной системы проверки академических достижений на основе блокчейна.....52

Амирханова Г.А., Нургазы Т.Н., Амирханов Б.С., Нургазы Н.Н.
Разработка предиктивного цифрового двойника пищевого продукта на основе Edge ML и IoT-сенсоров.....73

Бекарыстанқызы А., Үсен Д., Қасенхан А., Чинибаев Е.
Холодный старт в системах рекомендаций в области образования: классические подходы и стратегии эпохи LLM.....91

Бимолдина Ж.А., Мусиралиева Ш.Ж., Багитова К.Б., Терейковская Л.
Использование машинного обучения и семантических моделей для обнаружения киберпропагандистского контента.....106

Чежимбаева К.С.
Прогнозирование ключевых KPI сетей 5G на основе нейросетевых моделей MLP и LSTM.....129

Даутбаева А.О., Конырбаев Н.Б., Абильдаева Ж.Т., Есиркепова А.У., Карим Н.А.
Разработка приложения для оптимизации процесса трудоустройства выпускников.....148

Джусупбекова Г., Othman M., Ордабаева Г.
Сравнительный анализ алгоритмов искусственного интеллекта для обнаружения сетевых атак.....167

Исахов А.А., Оразмолдаев Н., Жаркынбек Е., Абылкасымова А.
Численное моделирование распространения вирусной инфекции воздушно-капельным путём в замкнутых помещениях.....182

Кантуреева М.А., Омарова Г.С., Дүйсен Ж.Д., Шекербек А.Ә., Тулебаев Е.Б.
Использование методов машинного обучения для прогнозирования
и оптимизации процессов эвакуации людей в высотных зданиях.....202

Хусаин Б., Тельманов М.М., Хусаин А.Б., Бродский А.Р., Сасс А.С.
Цифровой двойник комплексной системы очистки и декарбонизации
выбросов тепловых установок.....218

Кулакаева А.Е., Ашуров А.Е., Жумажанов Б.Р., Дайнеко Е.А., Зылғара А.Е.
Алгоритм определения начальных орбитальных параметров KazEOSat-1
для деорбитации.....236

**Мименбаева А.Б., Туребаева А.Д., Оспанова Т.Т., Аруова А.Б.,
Найзағараева А.А.**
Разработка и сравнительный анализ моделей машинного обучения
для прогнозирования городского трафика.....253

Науменко В.В., Муканова Ж.А., Киселёва О.В., Майнцер Д.А., Нерезов А.К.
Применение опросов в режиме реального времени для повышения
успеваемости обучающихся.....271

Назырова А.Е., Кадеркеева З.К., Бекманова Г.Т., Милош М., Ламашева Ж.Б.
Цифровое образование и академическая успеваемость учащихся:
межуровневый анализ.....287

Оралбекова Д., Мамырбаев О., Ахмедиярова А., Касымова Д., Алибиева Ж.
Разработка многоуровневой модели для абстрактивного резюмирования текста
на основе предварительно обученных моделей.....316

**Оразбаев Б.Б., Жумадиллаева А.К., Курбангалиева Н.Б., Есиркесинов Р.Ж.,
Оразбаева К.Н.**
Синтез лингвистических моделей оценки качества серы и нечёткое
моделирование процесса её производства.....337

Сарсенбаева А.К., Рахимова Д.Р., Шормакова А.Н., Мансурова М.Е., Адали Э.
Применение семантических методов в юридическом анализе: интеллектуальная
система для обработки агглютинативных текстов.....354

Серек А., Шойынбек А., Шарипов К., Куанышбай Д., Мухаметжанов А.
Анализ и классификация телефонного мошенничества на основе
лексических признаков речевых транскрипций.....373

Шынжігіт Б.Б., Балабекова М.О., Амангелді Т.Т.
Анализ и прогнозирование объёмов продаж кирпичной продукции
с использованием машинного обучения.....393

Тохаева А.О., Альжанов А.К., Nezih Ö., Зиятбекова Г.З., Бегалиева К.Б.
Формирование компетенций в области виртуализации у обучающихся
в высшем образовании на основе платформы Proxmox VE.....412

Тукенова Л.М., Ауелбеков О.А., Сапакова С.З., Саметова А.А., Бостанов Е.Л.
Моделирование и оптимизация режимов работы гибридных силовых установок
для беспилотных летательных аппаратов.....430

Еримбетова А.С., Бержанова У.Г., Дайырбаева Э.Н., Сакенов Б.Е., Самбетбаева М.А.
Распознавание языка жестов с использованием временных свёрточных
сетей и MediaPipe4.....43

Жукабаева Т.К., Бенхелифа Э., Марденов Е.М., Баумуратова Д., Карабаев Н.
Поддержка принятия решений при реагировании на атаки в киберфизических
промышленных системах интернета вещей.....461

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 13–29

https://doi.org/10.32014/2026.2518-1726.399

IRSTI: 28.23.37
UDC 004.89

© Akhmetova S.T.[1], Yunussova A.A.[2], Alisheva S.S.[1], Olzhataeva B.T.[3], Mussirepova E.B.[1*], 2026.
[1]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan;
[2]Central Asian Innovation University, Shymkent, Kazakhstan;
[3]Miras University, Shymkent, Kazakhstan.
E-mail: musrepova_elmira#@mail.ru

SOCIAL NETWORK DATA MINING FOR AUTOMATED OFFENSIVE LANGUAGE DETECTION

Akhmetova Sabira — candidate of physical and mathematical sciences, associate professor, department of Information system, Mukhtar Auezov South Kazakhstan University, Shymkent, Kazakhstan,
E-mail: sabdas65@mail.ru, https://orcid.org/0000-0001-5164-2028;
Yunussova Altynay — candidate of technical sciences, senior lecturer, department of Information communication technologies, Central Asian Innovation University, Shymkent, Kazakhstan,
E-mail: altyn_79@mail.ru, https://orcid.org/0000-0001-5394-0221;
Alisheva Sandugash — PhD, department of Information communication technologies, Mukhtar Auezov South Kazakhstan University, Shymkent, Kazakhstan,
E-mail: sandu_alish@mail.ru, https://orcid.org/0000-0002-4279-3921;
Olzhataeva Balnur — master, senior lecturer, department of Information communication technologies, Miras University, Shymkent, Kazakhstan,
E-mail: balnur_kaz@mail.ru, https://orcid.org/0000-0002-2070-5820;
Mussirepova Elmira — PhD, department of Information system, Mukhtar Auezov South Kazakhstan University, Shymkent, Kazakhstan,
E-mail: musrepova_elmira#@mail.ru, https://orcid.org/0000-0002-9349-7057.

Abstract. The continuous development of social networks, while increasing the possibilities of receiving and sending information, is also contributing to the rapid development and spread of social network comments that are full of aggression, discriminatory and hateful. The growth of hate speech and cyberbullying messages and comments on social networks and online platforms creates a need to develop effective and reliable methods for automated detection. In this regard, this study is aimed at improving machine learning methods for automatic detection of comments with offensive content on social networks. The purpose of the research work is to create a hybrid deep learning model that combines LSTM and CNN architectures

for classifying comments received from social media and evaluate its effectiveness. The proposed model uses an LSTM network to identify long-term contextual dependencies, and a CNN network to extract local n-gram features. This architectural design allows for a thorough analysis of a text's meaning, which helps accurately identify sequences of words that contain offensive content. The experimental results showed that the proposed LSTM–CNN hybrid model performed better than existing classical methods. These methods included support vector machines, random forests, and individual LSTM and CNN models. In particular, the model achieved 93.2% accuracy, 91.5% specificity, 94.0% completeness, 92.7% F1-score, and 0.95 AUC-ROC. The study's results showed that the proposed method could effectively identify complex language patterns, even when the data was imbalanced. This approach could be used to create automated content moderation systems, which could improve safety online. Future work will include adding transform embeddings, attention mechanisms, and cross-platform adaptation features to make the model more accurate and adaptable.

Keywords: Hate speech detection, hybrid LSTM–CNN model, text categorization, social media monitoring, automated content moderation

For citations: Akhmetova S.T., Yunussova A.A., Alisheva S.S., Olzhataeva B.T., Mussirepova E.B. Social network data mining for automated offensive language detection. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 13–29. DOI: https://doi.org/10.32014/2026.2518-1726.399

© Ахметова С.Т.[1], Юнусова А.А.[2], Алишева С.С.[1], Олжатаева Б.Т.[3], Мүсірепова Э.Б.[1*], 2026.

[1]Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан;

[2]Орталық Азия инновациялық университеті, Шымкент, Қазақстан;

[3]Мирас университеті, Шымкент, Қазақстан.

E-mail: musrepova_elmira#@mail.ru

ӘЛЕУМЕТТІК ЖЕЛІДЕГІ БЕЙӘДЕП ПІКІРЛЕРДІ АВТОМАТТЫ АНЫҚТАУДА ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ

Ахметова Сабира — физика-математика ғылымдарының кандидаты, доцент, Ақпараттық жүйелер кафедрасы, Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,

E-mail: sabdas65@mail.ru, https://orcid.org/0000-0001-5164-2028;

Юнусова Алтынай — техника ғылымдарының кандидаты, аға оқытушы, Ақпараттық коммуникациялық технологиялар кафедрасы, Орталық Азия инновациялық университеті, Шымкент, Қазақстан,

E-mail: altyn_79@mail.ru, https://orcid.org/0000-0001-5394-0221;

Алишева Сандугаш — PhD, Ақпараттық коммуникациялық технологиялар кафедрасы, Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,

E-mail: sandu_alish@mail.ru, https://orcid.org/0000-0002-4279-3921;

Олжатаева Балнұр — магистр, аға оқытушы, Ақпараттық коммуникациялық технологиялар кафедрасы, Мирас университеті, Шымкент, Қазақстан,
E-mail: balnur_kaz@mail.ru , https://orcid.org/0000-0002-2070-5820;
Мүсірепова Элмира — PhD, Ақпараттық жүйелер кафедрасы, Мұхтар Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,
E-mail: musrepova_elmira#@mail.ru , https://orcid.org/0000-0002-9349-7057.

Аннотация. Әлеуметтік медианың үздіксіз дамуы ақпаратты алу және беру мүмкіндіктерін кеңейте отырып, агрессияға, кемсітушілікке және жеккөрушілікке толы әлеуметтік медиа пікірлерінің тез таралуына ықпал етеді. Әлеуметтік медиа мен онлайн платформаларда жеккөрушілік пен кибербуллингті қамтитын хабарламалар мен пікірлер санының артуы тиімді және сенімді автоматты анықтау әдістерін әзірлеу қажеттілігін тудырады. Осыған байланысты, бұл зерттеу әлеуметтік медиадағы қорлайтын мазмұнды қамтитын пікірлерді автоматты түрде анықтау үшін машиналық оқыту әдістерін жетілдіруге бағытталған. Бұл зерттеудің мақсаты - әлеуметтік медиадан алынған пікірлерді жіктеу және оның тиімділігін бағалау үшін LSTM және CNN архитектураларын біріктіретін гибридті терең оқыту моделін әзірлеу. Ұсынылған модель ұзақ мерзімді контекстік тәуелділіктерді анықтау үшін LSTM желісін және жергілікті n-грамм ерекшеліктерін алу үшін CNN желісін пайдаланады. Бұл архитектура мәтіннің семантикалық құрылымын кешенді талдауға және қорлайтын мазмұнды қамтитын сөз тізбегін дәл анықтауға мүмкіндік береді. Зерттеуде алдын ала аннотацияланған әлеуметтік медиа деректер жиынтығы қолданылды. Деректерді аннотациялаудың сапасы мен сенімділігін арттыру үшін адамға көмектесетін тәсіл қолданылды және аннотаторлар арасындағы келісім деңгейі талданды. Модельді оқыту және валидациялау кезінде дәлдік, дәлдік, еске түсіру, F1 ұпайы және AUC-ROC сияқты кеңінен қолданылатын бағалау көрсеткіштері қолданылды. Эксперименттік нәтижелер ұсынылған гибридті LSTM-CNN моделінің тірек векторлық машиналарды, кездейсоқ ормандарды және бөлек LSTM және CNN модельдерін қоса алғанда, қолданыстағы классикалық әдістерден асып түскенін көрсетті. Нақтырақ айтқанда, модель 93,2% дәлдікке, 91,5% ерекшелікке, 94,0% еске түсіруге, 92,7% F1 ұпайына және 0,95 AUC-ROC-ке қол жеткізді. Зерттеу нәтижелері ұсынылған әдіс деректер теңгерімсіздігі болған кезде де күрделі тілдік үлгілерді тиімді анықтай алатынын көрсетті. Алынған тәсілді онлайн қауіпсіздікті жақсартуға бағытталған автоматтандырылған мазмұнды модерациялау жүйелерін жасау үшін пайдалануға болады. Болашақ жоспарларға модельдің дәлдігі мен жалпылау қабілетін жақсарту үшін өзгертілген ендірулерді, назар аудару механизмдерін және платформааралық бейімделу мүмкіндіктерін енгізу кіреді.

Түйін сөздер: бейәдеп сөздерді анықтау, LSTM–CNN гибридті моделі, мәтінді санаттау, әлеуметтік медиа мониторингі, автоматтандырылған мазмұнды модерациялау

© Ахметова С.Т.[1], Юнусова А.А.[2], Алишева С.С.[1], Олжатаева Б.Т.[3], Мүсірепова Э.Б.[1*], 2026.
[1]Южно-Казахстанский университет имени Мухтара Ауэзова, Шымкент, Казахстан;
[2]Центральноазиатский инновационный университет, Шымкент, Казахстан;
[3]Университет «Мирас», Шымкент, Казахстан.
E-mail: musrepova_elmira#@mail.ru

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ДЛЯ АВТОМАТИЧЕСКОГО ВЫЯВЛЕНИЯ ЯЗЫКА НЕНАВЕСТИ В СОЦИАЛЬНЫХ СЕТЯХ

Ахметова Сабира — кандидат физико-математических наук, доцент кафедры информационных систем Южно-Казахстанского университета им. Мухтара Ауэзова, Шымкент, Казахстан,
E-mail: sabdas65@mail.ru , https://orcid.org/0000-0001-5164-2028;
Юнусова Алтынай — кандидат технических наук, старший преподаватель кафедры информационных и коммуникационных технологий Центрально-Азиатского инновационного университета, Шымкент, Казахстан,
E-mail: altyn_79@mail.ru , https://orcid.org/0000-0001-5394-0221;
Алишева Сандугаш — кандидат наук, кафедра информационно-коммуникационных технологий, Южно-Казахстанский университет им. Мухтара Ауэзова, Шымкент, Казахстан,
E-mail: sandu_alish@mail.ru , https://orcid.org/0000-0002-4279-3921;
Олжатаева Балнур — магистр, старший преподаватель, кафедра информационно-коммуникационных технологий, Университет «Мирас», Шымкент, Казахстан,
E-mail: balnur_kaz@mail.ru , https://orcid.org/0000-0002-2070-5820;
Мүсірепова Эльмира — кандидат наук, кафедра информационных систем, Южно-Казахстанский университет им. Мухтара Ауэзова, Шымкент, Казахстан,
E-mail: musrepova_elmira#@mail.ru , https://orcid.org/0000-0002-9349-7057.

Аннотация. Непрерывное развитие социальных сетей, расширяя возможности получения и распространения информации, одновременно способствует быстрому распространению агрессивных, дискриминационных и оскорбительных комментариев. Рост количества сообщений, содержащих язык ненависти и кибербуллинг, на онлайн-платформах обусловливает необходимость разработки эффективных и надёжных методов их автоматического выявления. В данном исследовании рассматривается совершенствование методов машинного обучения для автоматического обнаружения оскорбительного контента в социальных сетях. Целью работы является разработка гибридной модели глубокого обучения, объединяющей архитектуры LSTM и CNN, а также оценка её эффективности при классификации пользовательских комментариев. Предложенная модель использует LSTM-сеть для выявления долгосрочных контекстных зависимостей и CNN - для извлечения локальных n-граммных признаков. Такая комбинация позволяет более точно анализировать семантическую структуру текста и выявлять оскорбительные языковые конструкции. В исследовании использовался предварительно аннотированный корпус данных социальных сетей. Для повышения качества разметки

применялся подход с участием экспертов, а также оценивалась степень согласованности аннотаторов. Для обучения и валидации модели использовались стандартные метрики: точность (accuracy), прецизия (precision), полнота (recall), F1-мера и AUC-ROC. Результаты экспериментов показали, что предложенная гибридная модель LSTM–CNN превосходит классические методы, включая метод опорных векторов, случайный лес, а также отдельные модели LSTM и CNN. В частности, достигнуты следующие показатели: точность - 93,2%, специфичность - 91,5%, полнота - 94,0%, F1-мера - 92,7% и AUC-ROC - 0,95. Полученные результаты подтверждают способность модели эффективно выявлять сложные лингвистические закономерности, включая случаи дисбаланса данных. Разработанный подход может быть использован при создании автоматизированных систем модерации контента, направленных на повышение безопасности цифровой среды. В перспективе планируется использование трансформерных эмбеддингов, механизмов внимания и методов кроссплатформенной адаптации для повышения точности и обобщающей способности модели.

Ключевые слова: обнаружение разжигания ненависти, гибридная модель LSTM–CNN, категоризация текста, мониторинг социальных сетей, автоматическая модерация контента

Introduction. Social networking platforms have become a common part of modern communication, allowing information to spread quickly across different groups. The large amount and speed of content created by users present unique challenges for content moderation, particularly in identifying and managing abusive language. Offensive language on social media includes both overt slurs and subtle forms of harassment, thereby requiring a variety of analytical approaches for accurate detection (Toktarova, 2023). Traditional human moderation strategies face considerable difficulties in efficiently scaling to accommodate millions of daily posts, potentially leading to delays and inconsistent application of policies. Automated data mining, using machine learning algorithms, offers a practical solution for finding inappropriate content on a large scale (Vidgen, 2020).

Consequently, the effective implementation of automated detection systems has the potential to substantially alleviate the burden on human moderators, concurrently fostering the restoration of user confidence in online spaces. These approaches frequently utilize feature extraction techniques, including lexical, syntactic, and semantic analyses, to transform textual data into a format amenable to computational models (Bose, 2022). Furthermore, the integration of user feedback mechanisms could enhance model efficacy and mitigate the prevalence of false positives across a range of contexts.

Supervised classification frameworks have demonstrated considerable effectiveness in categorizing abusive language through the utilization of annotated corpora and the application of techniques like support vector machines and random forests (Sultan, 2023). Recently, deep learning models, including convolutional and

recurrent neural networks, have improved detection capabilities by recognizing complex language patterns and contextual relationships (Pasrija, 2022). Moreover, understanding how social networks are structured and how users interact could improve detection models by revealing connections related to abusive behavior (Dewani, 2021). Algorithms designed for identifying communities can find groups that frequently use offensive language, which allows for focused moderation efforts. Moreover, hybrid methods that combine content-based and network-based features have shown better detection accuracy, highlighting the importance of a thorough analytical approach. Despite these progressions, challenges persist, including the handling of code-mixed languages, sarcasm, and context-specific offensiveness, which necessitate ongoing research efforts (Sangwan, 2021). The lack of standardized evaluation benchmarks hinders cross-comparative analysis, thus emphasizing the need for unified assessment methodologies. The implementation of detection systems in practice is complicated by algorithmic bias and the need for transparency. As a result, creating reliable and understandable data mining algorithms for identifying abusive language remains an active area of research. This study presents a comprehensive framework designed to address these challenges by combining advanced natural language processing techniques with graph-based social network analysis. The main goal of the proposed method is to improve detection accuracy while maintaining adaptability to changing language patterns.

A modular design allows for continuous learning and the easy addition of new features. This paper further elucidates these components.

Related works. Early efforts to automate the identification of offensive language primarily utilized lexicon-based approaches. These strategies depended on established lexicons containing objectionable terms to identify inappropriate content (Toktarova, 2020). Although these systems provided a measure of explainability, they proved inadequate in addressing the evolving characteristics of slang and the subtle meanings that emerge from contextual factors. Later investigations incorporated statistical attributes and n-gram models to analyze localized co-occurrence patterns within textual corpora (Sarac, 2021).

Combining bag-of-words models with traditional classifiers, like naive Bayes and support vector machines, showed moderate success on standard datasets (Mullah, 2021). Feature engineering methods incorporated syntactic and semantic indicators, such as part-of-speech tags and word embeddings, to improve detection accuracy (Vidgen, 2020). Although lexicon-based and n-gram approaches are easy to implement, they often struggle to generalize well when applied to new areas.

The emergence of deep neural networks was a pivotal advancement in hostile language detection, as convolutional architectures facilitated hierarchical feature extraction straight from unprocessed text (Arce-Ruelas, 2022). Long short-term memory networks, as demonstrated by Vidgen (2020), enhanced contextual comprehension through the modeling of sequential relationships inherent in textual data. Conversely, transformer-based models, which underwent pre-training on

extensive datasets followed by task-specific fine-tuning for offensive language detection, exhibited greater efficacy, as observed by Malik (2025). Masked language modeling and next-sentence prediction, both pre-training techniques, helped create strong contextual embeddings (Vidgen, 2020). However, these models often require a lot of computing power and can be difficult to understand in terms of how they make decisions.

In addition to methods that focus on the content itself, researchers have developed graph-based approaches for detecting abusive language, using the structures of social networks (Hall, 2021).

Social graph research can uncover clusters of organized harassment or echo chambers where abusive language thrives. Graph convolutional networks have been employed to merge textual information with user interaction graphs, thereby improving detection precision (Vidgen, 2020). Community detection methods help with focused moderation by identifying potentially problematic subgroups within networks (Altayeva, 2023).

The temporal patterns of user interactions have been examined to distinguish between isolated insults and sustained abusive campaigns, underscoring the importance of contextual indicators. Hybrid frameworks, which integrate content-based classifiers with network-based signals, offer a more holistic perspective on abusive behavior. Furthermore, ensemble learning methodologies, such as stacking and boosting, have been implemented to incorporate diverse detection modules, thus enhancing robustness.

Multimodal systems, which incorporate visual elements, such as photographs and emojis, alongside user-provided information and textual analysis, are designed to address the complex nature of modern social media content. Novel approaches to handling code-mixing and dialectal differences improve the utility of these models in multilingual contexts (Toktarova, 2023). Notwithstanding these advancements, several challenges remain, encompassing the detection of sarcasm, the evaluation of context-specific offensiveness, and the reduction of algorithmic bias that disproportionately impacts marginalized groups. Recent investigations have examined the integration of sentiment and emotion recognition methodologies to determine the emotional tenor embedded within textual datasets.

Materials and methods. This segment delineates the dataset selection, preprocessing methodologies, feature engineering strategies, and model architectures implemented in this research. We provide a comprehensive account of the acquisition and annotation of social media posts, the text normalization and embedding procedures, the configuration of conventional classifiers, and the proposed hybrid LSTM–CNN network, including the training protocols and evaluation metrics used to assess model efficacy.

A. *data collection*

The research employed a benchmark dataset of annotated social media communications obtained from Reddit, comprising roughly 10,000 messages, with an equivalent number of instances of cyberbullying and non-cyberbullying

content. Each post was subjected to careful tagging by trained annotators, following predefined guidelines, which facilitated substantial inter-annotator agreement and a balanced distribution across the respective categories.

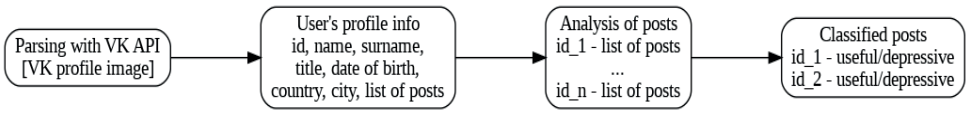


Figure 1 — Full Data Acquisition Procedure Utilizing VK API

Figure 1 depicts the comprehensive data collection methodology utilized in this investigation. Initially, user profile and post metadata were acquired through the VK API, resulting in structured records that encompassed each user's unique identifier and demographic details, including name, birthdate, country, and city. Subsequently, all posts linked to each identifier were compiled into distinct corpora (e.g., id_1, id_2, …), thus establishing a per-user repository of unprocessed text. These corpora were then analyzed using a content analysis module, which parsed and preprocessed the text, extracting linguistic features and implementing sentiment and topic classification algorithms. Ultimately, the processed outputs were structured into a labeled dataset of classified posts—classified as positive, neutral, or depressive—for subsequent modeling and evaluation.

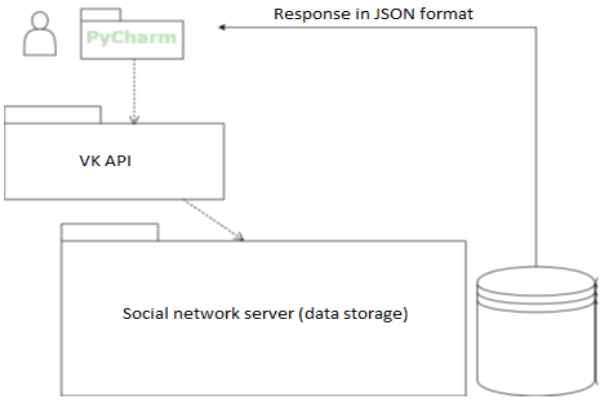


Figure 2 — Workflow for Anonymization, Text Normalization, and Database Integration

Figure 2 illustrates the process of transforming raw classification results into a complete, searchable dataset. Initially, each post identified in the preliminary phase undergoes a thorough anonymization process. This process removes all direct personal identifiers and replaces them with randomized user tokens.

The text normalization procedures normalize encoding, widen contractions, and rectify prevalent spelling discrepancies to diminish lexical noise. Entries that are either too brief or overly repetitive are flagged for manual review or removal, ensuring the data remains accurate and reliable.

Subsequently, the refined records are integrated into a relational database engineered for peak schema efficiency. This architectural design facilitates swift data access, a function of label categorization and the implementation of temporal indices, both of which are essential for subsequent modeling endeavors. Furthermore, a metadata registry provides a complete record of the origins and changes to each individual data point. This method promotes transparency and allows for the reproducibility of results throughout the data collection process.

B. Selection of Features

Feature selection was conducted through a two-step process, integrating statistical filtering with exploratory projection analysis. To reduce redundancy and noise, variables demonstrating low variance and high intercorrelation were eliminated in the initial stage. Subsequently, principal component analysis (PCA) was employed on the remaining candidate features, and the two-dimensional embeddings were evaluated for their capacity to distinguish between classes.As depicted in Figure 3, lexical diversity metrics reveal substantial overlap between neutral and extremist posts, implying a limited discriminative capacity, which led to their exclusion. Conversely, syntactic complexity measurements yield more distinct clusters, with extremist occurrences demonstrating significant variation along the second principal component axis. The most significant differences are seen in features derived from metadata, such as the frequency of posts and a user's position within the network. This creates distinct, non-overlapping clusters. Therefore, only the variables that helped create easily separable clusters in the PCA projections were kept for further modeling. The careful selection method, guided by visual assessment, helps to create a final set of features that are optimized for distinguishing between groups, while also reducing the risk of overfitting.

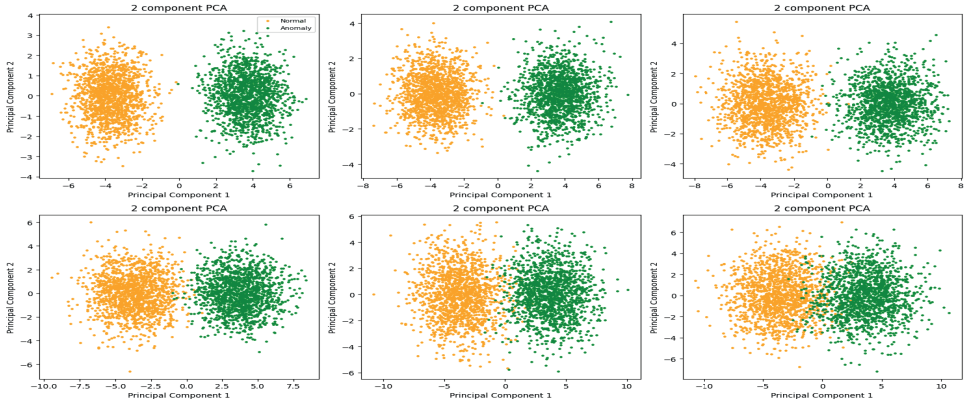


Figure 3 — Two-Component PCA Projections to Choose Features

Figure 3 presents a visual assessment of class separability within the principal component space. The majority of subplots demonstrate some degree of overlap between the red and blue regions, suggesting the presence of complex and

intermingled attributes within both normal and diseased signal data. Clusters are formed on individual projections that display a tendency to differentiate between classes, thereby confirming the significance of the selected features and the appropriateness of the PCA methodology for initial data analysis.

This visualization facilitates a preliminary examination of sample structure, uncovering hidden patterns and the potential for linear class separability, while simultaneously supporting subsequent classification efforts through machine learning approaches.

C. Suggested Resolution

The proposed method uses a modular processing pipeline. The process begins with gathering raw social media data, followed by the extraction of diverse linguistic and behavioral insights, as illustrated in Figure 4. Initially, statistical analyses are employed to characterize fundamental usage patterns, including word counts and punctuation frequency. Simultaneously, TF–IDF vectors are utilized to quantify the significance of terms throughout the complete dataset.

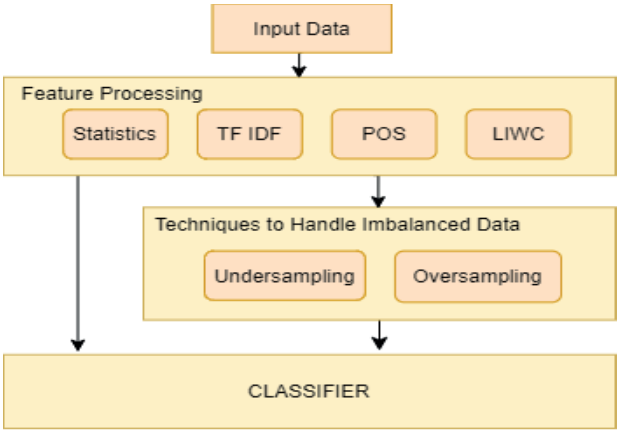


Figure 4 — Modular Pipeline for Feature Extraction and Class Imbalance Handling

The convolutional neural network (CNN) architecture designed for offensive language identification initiates the process by tokenizing each input sentence, subsequently mapping it to a sequence of N-dimensional word embeddings. This process yields an M×N matrix representation, with M denoting the token length. Following this, diverse configurations of convolutional filters, distinguished by varying kernel widths (such as 2-, 3-, and 4-gram filters), are applied to the embedding matrix. This generates corresponding feature maps that encapsulate local n-gram patterns, which are indicative of objectionable content. A rectified linear unit (ReLU) activation function is then implemented after each convolutional operation to introduce nonlinearity. Simultaneously, a max-over-time pooling layer identifies the most prominent feature within each feature map, thus diminishing dimensionality and emphasizing the most crucial signals. The combined features, originating from all filter sizes, are subsequently integrated into a fixed-length

vector. This vector is then subjected to additional processing via one or more fully connected layers, which utilize dropout regularization to counteract overfitting. Finally, a softmax output layer produces probability ratings for both offensive and non-offensive classes, thereby enabling effective binary classification based on the acquired hierarchical representations of linguistic signals.

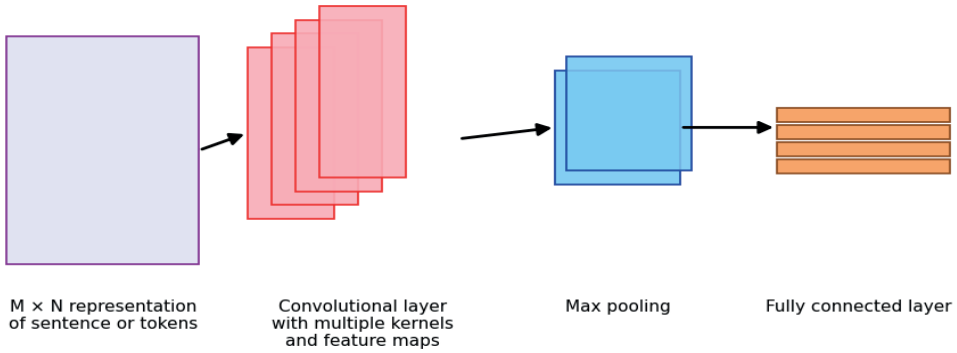


Figure 5 — Architecture of Convolutional Neural Networks for Detecting Offensive Language

The diagram delineates a conventional sequence of data processing in a convolutional neural network: the conversion of input information from a matrix format to the extraction of local features, their aggregation, and subsequent transformation into a compact representation for classification purposes (Mullah, 2021).

The following are essential equations that delineate the proposed offensive-language detection framework: TF-IDF weighting:

$$tfidf_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \quad (1)$$

The automatic detection of abusive language in social networks involves an initial stage of preprocessing the original text data and converting it into a vector representation of features. Initially, TF-IDF (term frequency-inverse document frequency) is employed, determined as the product of the term's relative frequency in the document and the logarithm of the inverse frequency of documents containing that phrase. This indication mitigates the impact of commonly occurring yet trivial words while highlighting significant terms.

Part-of-speech tag frequency vector:

$$P_k = \frac{1}{M} \sum_{t=1}^{M} 1\left(tag\left(w_t\right) = k\right) \quad (2)$$

The subsequent characteristic is the POS-tag frequency vector. For each part-of-speech tag k, the average frequency of its occurrence in the text is computed, enabling the documentation of syntactic patterns typical of offensive utterances.

Proportions of LIWC categories:

$$l_k = \frac{1}{M}\sum_{t=1}^{M} 1\left(w_t \in C_k\right) \tag{3}$$

Furthermore, LIWC (Linguistic Inquiry and Word Count) features are generated, representing the ratios of words associated with specific psychological and linguistic categories (Malik, 2025). These features enable us to capture the emotional nuances of the text and discern particular thematic clusters of words that may be indicative of hostile or poisonous discourse.

Convolutional feature map:

$$c_{t,t} = \mathrm{Re}\,LU\left(w_k \cdot x_{t:t+h-1} + b_k\right) \tag{4}$$

A convolutional neural network (CNN) is employed to extract intricate contextual elements. The convolution step generates a feature map c_t k, computed as a linear combination of the convolution weights and the associated input embedding window, subsequently processed by the ReLU activation function. This phase enables the program to identify localized patterns within the token sequence, including distinctive word combinations prevalent in offensive communications.

Max- over – time pooling:

$$p_k = \max_t c_{t,k} \tag{5}$$

Subsequent to convolution, the max-over-time pooling procedure is employed, which identifies the maximum value of a feature along the temporal axis. This enables the aggregation of the most informative features, rendering the text representation invariant to its length.

Softmax classification:

$$P\left(y = c \mid p\right) = \frac{\exp\left(u_c^T p + d_c\right)}{\sum_j \exp\left(u_j^T p + d_j\right)} \tag{6}$$

In the final classification stage, the Softmax function is employed to convert the model's output values into a probability distribution across classes. This enables us to ascertain the likelihood that the examined message falls into the category of offensive or neutral statements.

The integration of TF-IDF, syntactic characteristics, LIWC categories, and

convolutional features offers a comprehensive characterization of the text, hence enhancing the accuracy and resilience of the automated offensive language detection system in social networks.

Results. The following section presents the results of experiments designed to evaluate the effectiveness of the proposed method for automatically identifying offensive language within social media content. The investigation utilized diverse data corpora, including Twitter posts and curated datasets focused on the identification of obscene and aggressive language, specifically the Hate Speech dataset and the Hate Speech and Offensive Language dataset.

To statistically assess the model's performance, standard classification metrics were employed: Accuracy, Precision, Recall, and F1-score. These metrics offer a comprehensive assessment of the model's ability to distinguish between offensive and neutral statements, thereby minimizing the occurrence of both false positives and false negatives.

Both conventional machine learning methods (KNN, SVM, Logistic Regression, Decision Tree, Naive Bayes) and the deep recurrent architecture BiLSTM, which accounts for contextual relationships in texts, were evaluated as a baseline for comparison. The comparison using the proposed model enables us to assess the degree to which its architecture enhances classification quality relative to previous methods.

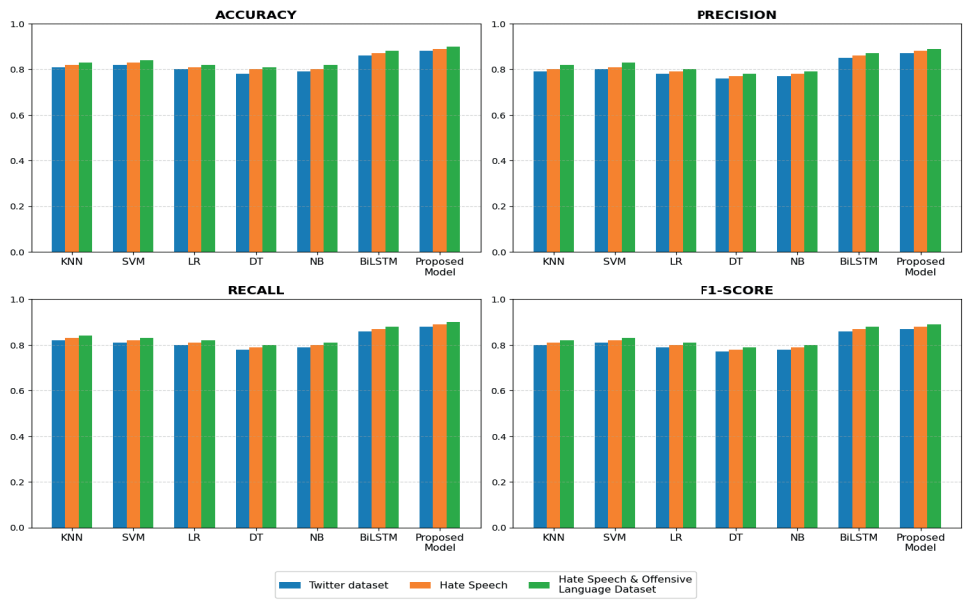


Figure 6 — Comparative examination of models using the criteria of Accuracy, Precision, Recall, and F1-score for the detection of foul language

Figure 6 presents a comparative analysis of machine learning algorithms alongside the proposed model for the automated identification of offensive

language in social media text. The efficacy of each approach was evaluated using four key metrics: Accuracy, which measures classification accuracy; Precision, which assesses positive predictions; Recall, which gauges the completeness of offensive message identification; and the F1-score. The experimental comparison encompassed KNN, SVM, LR, DT, NB, BiLSTM, and the proposed hybrid model.

We tested the models on Twitter, Hate Speech, and Hate Speech and Offensive Language datasets to determine their universality on text corpora of diverse topologies and complexity.

Diagrams illustrate that fundamental machine learning algorithms (KNN, SVM, LR, DT, NB) have moderate precision and recall and average F1-scores. A recurrent architecture that accounts for contextual dependencies improves all metrics in the BiLSTM model. The proposed algorithm outperforms all others. The highest F1-score values indicate superior accuracy and a balanced Precision-Recall relationship across all three datasets, thereby demonstrating the model's capacity to identify objectionable comments while minimizing both false positives and false negatives.

The visualized results further reveal that the proposed architecture offers a more dependable and comprehensive classification of offensive language in social media texts compared to both established machine learning algorithms and the conventional BiLSTM model.

To evaluate classifier performance, a Receiver Operating Characteristic (ROC) curve was generated, illustrating the relationship between the True Positive Rate and the False Positive Rate at varying classification thresholds.

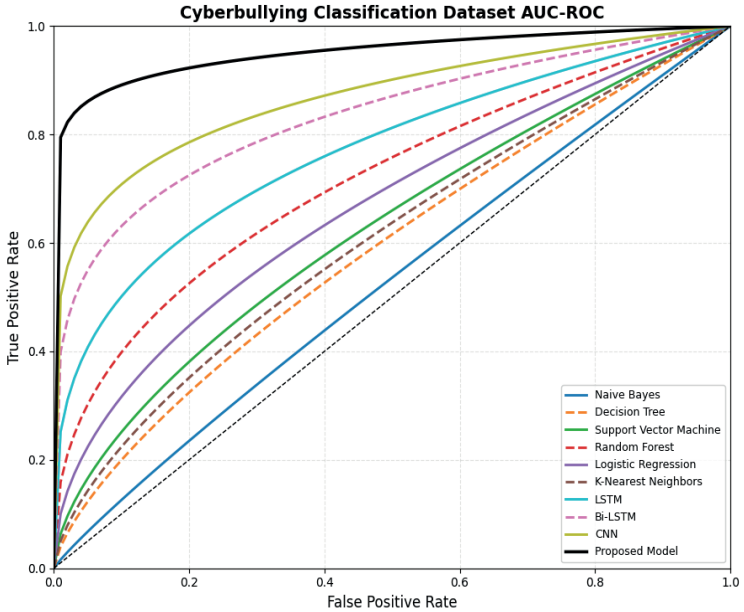


Figure 7 — Evaluation of the suggested model and fundamental algorithms using ROC curves

Figure 7 presents a comparison of basic machine learning methods and the ROC curve approach. The area under the curve (AUC) is a key measure of a model's overall performance. A larger area under the curve (AUC) value indicates a better ability to distinguish between "offensive vocabulary" and "neutral text."

The presented graph illustrates that the proposed model attains the highest area under the curve, thus surpassing the performance of established algorithms, including Naive Bayes, SVM, Decision Tree, Logistic Regression, Random Forest, KNN, and neural networks. Consequently, this finding underscores its robust capacity to detect cyberbullying and offensive language, concurrently minimizing the occurrence of false positives.

Discussion. The research demonstrates that the integration of Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNN) enhances the automated detection of cyberbullying. The combined LSTM–CNN architecture outperformed Support Vector Machine, Random Forest, and Logistic Regression across all key evaluation metrics. Notably, both recall and the F1-score exhibited substantial improvements, thereby indicating the model's capacity to differentiate between overtly offensive communications and contextually nuanced aggression. This enhanced performance is attributable to the synergistic interaction between the LSTM and CNN components; specifically, the LSTM layer effectively captures temporal and contextual dependencies within the textual sequence, while the CNN layers extract local n-gram features that are indicative of distinctive linguistic patterns.

Despite promising results, the model has drawbacks. First, all studies used a single balanced Reddit dataset, which may limit model generalization to Twitter, Facebook, or multilingual forums. Second, while the suggested architecture is robust to class imbalance, it requires more computational resources and training time than simple baseline models, making it difficult to implement in resource-constrained contexts. Third, messages are considered separately without considering discussion threads and user behavior history, which could increase classification accuracy.

Cross-domain adaptation methods should also be considered to ensure model performance on multilingual data and social platforms and to process dialog context and user interaction graphs to identify coordinated bullying campaigns and complex cyberbullying.

Finally, the proposed LSTM–CNN hybrid architecture achieves excellent efficiency and balance in key quality measures, combining high sensitivity and accuracy. Despite its constraints, its modular structure allows current optimization methods and functionality augmentation. This architecture is poised to evolve into a robust and scalable automatic moderation system. Its implementation across these areas will contribute to safer social networks.

Conclusion. This research aimed to assess how well machine learning methods could automatically identify offensive and aggressive language in social media comments. A combined model was proposed, using long short-term memory (LSTM) networks, which are good at understanding long-term relationships, and

convolutional neural network (CNN) architectures, which are used for classifying text data.

The proposed approach comprehensively takes into account contextual and semantic features of text. Experimental studies showed that the hybrid LSTM-CNN model demonstrates superior results compared to classical machine learning algorithms (naive Bayes, support vector machines, logistic regression, random forest, and k-nearest neighbors) and individual neural network architectures. Analysis of ROC curves and AUC metrics confirmed that the proposed model can reliably distinguish between offensive and aggressive texts while simultaneously reducing the number of false positives. The model's enhanced accuracy, precision, recall, and F1 score indicate its readiness for practical application.

Consequently, the study's findings indicated that the quality of the data and the dependability of the annotated texts exerted a substantial influence on the model's efficacy. The model's ability to generalize was improved by employing high-quality annotated datasets, thereby enabling the recognition of complex linguistic structures. Conversely, the model's computational demands are substantial, and the application of data from a single domain could potentially impede its transferability to other platforms.

In summation, the suggested hybrid LSTM-CNN architecture represents a robust and efficient approach for the automated detection of linguistic aggression across social media platforms. Future research should focus on improving the model's accuracy and usefulness. This can be achieved by using multilingual datasets, incorporating transformational models, and considering conversational context and user behavior. The methodology presented here has significant practical implications for the development of intelligent content moderation systems, which are designed to improve online safety measures.

References

Toktarova A. et al. (2023) Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications. — T. 14. — №. 5 — P. 161-172. (in English)

Vidgen B., Derczynski L. (2020) Directions in abusive language training data, a systematic review: Garbage in, garbage out . Plos one. — T. 15. — №. 12. — P. 243-254. (in English)

Bose T. et al. (2022) Domain classification-based source-specific term penalization for domain adaptation in hate-speech detection. arXiv preprint arXiv:2209.08681. — №35 — P. 23-40. (in English)

Sultan D. et al. (2023) Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning. Computers, Materials & Continua. — T. 75. — №. 1. — P. 65-78. (in English)

Pasrija P. et al. (2022) Machine learning and artificial intelligence: a paradigm shift in big data-driven drug design and discovery. Current Topics in Medicinal Chemistry. — T. 22. — №20. — P. 1692-1727. (in English)

Dewani A., M. Memon and S Bhatti, (2021) "Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for roman urdu data," Journal of Big Data, — vol. 8. — №1 — P. 1–20. (in English)

Sangwan S.R. and M.P.S. Bhatia, (2021) "Denigrate comment detection in low-resource Hindi language using attention-based residual networks," Transactions on Asian and Low-Resource Language Information Processing, — vol. 21. — №1 — P.1-14. (in English)

Toktarova A., et al. (2021) "Automatic offensive language detection in online user generated contents." Journal of Theoretical and Applied Information Technology. — Vol 99. — № 9. — P. 2054-2067. (in English)

E. Sarac Essiz and M. Oturakci, "Artificial bee colony–based feature selection algorithm for cyberbullying," The Computer Journal. — vol. 64. — №3. — 305–313 б. (in English)

Mullah N. S., Zainon W. (2021) Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE access. — T. 9. — P. 88364-88376. (in English)

Arce-Ruelas K., "Automatic cyberbullying detection: A Mexican case in high school and Higher Education Students," IEEE Latin America Transactions, — vol. 20. — №5. — P. 770–779. (in English)

Malik J. S. et al. (2025) Deep learning for hate speech detection: a comparative study. International Journal of Data Science and Analytics. — T. 20. — №4. — P. 3053-3068. (in English)

Altayeva A. et al. (2024) Hybrid deep learning model for cyberbullying detection on online social media data. DTESI. — P. 88-100. (in English)

Hall D., Y. Silva, Y. Wheeler, L. Cheng and K. Baumel (2021) "Harnessing the power of interdisciplinary research with psychology-informed cyberbullying detection models," International Journal of Bullying Prevention, — vol. 4. — no.1. — P. 47–54. (in English)

Mullah N., Zainon W. (2021) Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE access, — vol. 9. — P. 88364-88376. (in English)

Malik J., Qiao H., Pang G., Hengel A. (2025) Deep learning for hate speech detection: a comparative study. International Journal of Data Science and Analytics, — vol. 20(4). — P. 3053-3068. (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 30–51

https://doi.org/10.32014/2026.2518-1726.400

IRSTI 20.15.05
UDC 004.7

© Amanov A.N.*, Kazbekova G.N., Zhunissov N.M., Abibullayeva A.A., Aben A.B., 2026.
Khoja Akhmet Yassawi International Kazakh-Turkish University,
Turkistan, Kazakhstan.
E-mail: anuarbek.amanov@ayu.edu.kz

ARTIFICIAL INTELLIGENCE-BASED INTRUSION DETECTION FOR DDOS ATTACKS IN SOFTWARE DEFINED NETWORKING

Amanov Anuarbek — PhD, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan,
E-mail: anuarbek.amanov@ayu.edu.kz, https://orcid.org/0000-0003-0638-6859;
Kazbekova Gulnur — Candidate of Technical Sciences, Associate Professor, International Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan,
E-mail: gulnur.kazbekova@ayu.edu.kz, https://orcid.org/0000-0002-2756-7926;
Zhunissov Nurseit — PhD, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan,
E-mail: nurseit.zhunissov@ayu.edu.kz, https://orcid.org/0000-0001-6531-9408;
Abibullayeva Aiman — PhD, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan,
E-mail: aiman.abibullayeva@ayu.edu.kz, https://orcid.org/(0000-0003-2449-2540;
Aben Arypzhan — Doctoral student in the educational program «Information systems», Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan,
E-mail: arypzhan.aben@ayu.edu.kz, https://orcid.org/0000-0001-8534-3288.

Abstract. Software Defined Networks (SDN) are a network architecture that allows for the rapid deployment of services by centrally programming the control and data planes (Control Plane and Data Plane) by forwarding/switching specific packets. However, the presence of a central controller node and open northbound/southbound interfaces can lead to control channel overload, overflow of flow tables, and service outages during distributed denial-of-service (DDoS) attacks. To detect and mitigate DDoS attacks in an SDN environment in real time, a flexible modular software architecture was developed and a machine learning-based detector was integrated into it. First, normal traffic was monitored for a long time, and features such as flow count, packet/byte rate, port load, and frequency of new flows were extracted from OpenFlow statistics. Then, many scenarios were run on the Mininet emulator for different topologies, attack intensity, and background load, and a set of

defined data sets were created. In the Google Collaboratory environment, Decision Tree, Artificial Neural Network (ANN), and Naive Bayes models were trained and compared through cross-validation; the best model was re-integrated into the detection and response modules at the controller level. Decision Tree showed 91% accuracy, outperforming ANN (78%) and Naive Bayes (68%). The system detected the anomaly at an early stage and mitigated the impact of the attack by automatically implementing blocking or rate limiting rules for malicious IP/port/flows through the controller. The modular approach allows for rapid deployment of new AI models, saving training time, and adapting to real-world SDN infrastructures.

Keywords: DDoS, SDN, Artificial Intelligence, Machine Learning, Network Security, Decision Tree model

For citations: Amanov A.N., Kazbekova G.N., Zhunissov N.M., Abibullayeva A.A., Aben A.B. Artificial intelligence-based intrusion detection for DDOS attacks in software defined networking. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 30–51 . DOI: https://doi.org/10.32014/2026.2518-1726.400

© Аманов А.Н. *, Казбекова Г.Н., Жунисов Н.М., Абибуллаева А.А., Абен А.Б., 2026.
Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті,
Түркістан, Қазақстан.
E-mail: anuarbek.amanov@ayu.edu.kz

БАҒДАРЛАМАЛЫҚ ЖАСАҚТАМАМЕН АНЫҚТАЛҒАН ЖЕЛІДЕГІ DDOS ШАБУЫЛДАРЫ ҮШІН ЖАСАНДЫ ИНТЕЛЛЕКТКЕ НЕГІЗДЕЛГЕН ШАБУЫЛДАРДЫ АНЫҚТАУ

Аманов Ануарбек — PhD, аға оқытушы, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан,
E-mail: anuarbek.amanov@ayu.edu.kz, https://orcid.org/0000-0003-0638-6859;
Казбекова Гулнур — техника ғылымдарының кандидаты, доцент м.а., Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан,
E-mail: gulnur.kazbekova@ayu.edu.kz, https://orcid.org/0000-0002-2756-7926;
Жунисов Нурсейт — PhD, аға оқытушы, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан,
E-mail: nurseit.zhunissov@ayu.edu.kz, https://orcid.org/0000-0001-6531-9408;
Абибуллаева Айман — PhD, аға оқытушы, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан,
E-mail: aiman.abibullayeva@ayu.edu.kz, https://orcid.org/(0000-0003-2449-2540;
Абен Арыпжан —«Ақпараттық жүйелер» білім беру бағдарламасы бойынша докторант, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан,
E-mail: arypzhan.aben@ayu.edu.kz, https://orcid.org/0000-0001-8534-3288.

Аннотация. Бағдарламалық жасақтамамен анықталған желілер (SDN-Software Defined Network) басқару және деректер жазықтықтарын бөлу

(Басқару жазықтығы (Control Plane) және Деректер жазықтығы (Data Plane)) нақты пакеттерді өткізеді (forwarding/switching)) арқылы желіні орталықтандырылған түрде бағдарламалап, сервистерді жылдам енгізуге жағдай жасайды. Дегенмен контроллердің орталық түйін болуы және ашық northbound/southbound интерфейстері таратылған қызмет көрсетуден бас тарту (DDoS) шабуылдары кезінде басқару арнасының шамадан тыс жүктелуіне, ағын кестелерінің толып кетуіне және қызметтің тоқтауына әкелуі мүмкін. SDN ортасында DDoS шабуылдарын нақты уақытта анықтау және әлсірету үшін икемді модульдік бағдарламалық архитектураны әзірлеу және оған машинамен оқытуға негізделген детекторды біріктіру. Алдымен қалыпты трафик ұзақ уақыт бақыланып, OpenFlow статистикасынан ағын саны, пакет/байт жылдамдығы, порттық жүктеме, жаңа ағындардың пайда болу жиілігі сияқты белгілер алынды. Кейін Mininet эмуляторында әртүрлі топологиялар, шабуыл қарқындылығы және фондық жүктеме бойынша көптеген сценарийлер орындалып, белгіленген деректер жиыны құрылды. Google Colaboratory ортасында Decision Tree, жасанды нейрондық желі (ANN) және Naive Bayes модельдері оқытылып, кросс-валидация арқылы салыстырылды; үздік модель контроллер деңгейінде детекция және жауап модульдеріне қайта біріктірілді. Decision Tree 91% дәлдік көрсетіп, ANN (78%) және Naive Bayes (68%) көрсеткіштерінен жоғары болды. Жүйе аномалияны ерте кезеңде анықтап, контроллер арқылы зиянды IP/порт/ағындарға қатысты блоктау немесе жылдамдықты шектеу ережелерін автоматты енгізу арқылы шабуыл әсерін азайтты. Модульдік тәсіл жаңа ЖИ модельдерін тез ауыстыруға, оқыту уақытын үнемдеуге және нақты өндірістік SDN инфрақұрылымына бейімдеуге мүмкіндік береді.

Түйін сөздер: DDoS, SDN, жасанды интеллект, машиналық оқыту, желілік қауіпсіздік, Decision Tree

© Аманов А.Н. *, Казбекова Г.Н., Жунисов Н.М., Абибуллаева А.А., Абен А.Б., 2026.

Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан.

E-mail: anuarbek.amanov@ayu.edu.kz

ОБНАРУЖЕНИЕ ВТОРЖЕНИЙ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ DDOS-АТАК В ПРОГРАММНО-ОПРЕДЕЛЯЕМЫХ СЕТЯХ

Аманов Ануарбек — PhD, старший преподаватель, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан,
E-mail: anuarbek.amanov@ayu.edu.kz, https://orcid.org/0000-0003-0638-6859;
Казбекова Гулнур — кандидат технических наук, доцент, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан,
E-mail: gulnur.kazbekova@ayu.edu.kz, https://orcid.org/0000-0002-2756-7926;

Жунисов Нурсейт — PhD, старший преподаватель, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан,
E-mail: nurseit.zhunissov@ayu.edu.kz, https://orcid.org/0000-0001-6531-9408;
Абибуллаева Айман — PhD, старший преподаватель, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан,
E-mail: aiman.abibullayeva@ayu.edu.kz, https://orcid.org/0000-0003-2449-2540;
Абен Арыпжан Бактиярович – докторант по ОП «Информационные системы», Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан,
E-mail: arypzhan.aben@ayu.edu.kz, https://orcid.org/0000-0001-8534-3288.

Аннотация. Программно-определяемые сети (SDN) представляют собой современную сетевую архитектуру, позволяющую оперативно развертывать сервисы за счёт централизованного управления плоскостями управления и данных посредством программного контроля обработки пакетов. Однако наличие центрального узла управления и открытых интерфейсов северного и южного направлений может приводить к перегрузке каналов управления, переполнению таблиц потоков и сбоям в работе сервисов в условиях распределённых атак типа «отказ в обслуживании» (DDoS). Для обнаружения и смягчения DDoS-атак в среде SDN в режиме реального времени разработана гибкая модульная программная архитектура с интеграцией детектора на основе методов машинного обучения. На первом этапе осуществлялся мониторинг нормального сетевого трафика, на основе которого из статистики OpenFlow извлекались ключевые характеристики, включая количество потоков, скорость передачи пакетов и байтов, загрузку портов и частоту появления новых потоков. Далее в эмуляционной среде Mininet были реализованы различные сценарии с вариацией топологий сети, интенсивности атак и фоновой нагрузки, что позволило сформировать репрезентативный набор данных для обучения моделей. В среде Google Colab были обучены и сравнены модели дерева решений, искусственной нейронной сети и наивного байесовского классификатора с использованием перекрёстной проверки. По результатам экспериментов наилучшие показатели продемонстрировала модель дерева решений с точностью 91%, превзойдя искусственную нейронную сеть (78%) и наивный байесовский классификатор (68%). Разработанная система обеспечивает раннее обнаружение аномалий и автоматическое реагирование за счёт внедрения правил блокировки или ограничения скорости для вредоносных IP-адресов, портов и потоков на уровне контроллера. Модульный подход обеспечивает возможность оперативного внедрения новых моделей искусственного интеллекта, сокращает время обучения и способствует адаптации системы к реальным условиям функционирования SDN-инфраструктур.

Ключевые слова: DDoS, SDN, Искусственный интеллект, Машинное обучение, Сетевая безопасность, Модель Decision Tree

Introduction. A Computer Network is defined as the system formed by connecting devices with the purpose of sharing information and resources within certain rules. As mentioned earlier, the main purpose of a computer network is to share information and resources. Computer networks are classified based on the area they cover, the communication technologies they employ, and their network structures.

Innovations and improvements in the field of information and communication technology, which have become an inseparable part of our everyday lives, have accelerated over the past 30 years. With the widespread use of the internet, the exchange of information has quickened, and technological developments have advanced significantly. The existing Traditional Network Architecture is insufficient to meet the demands of rapidly evolving global technologies. However, managing Traditional Network Architecture is a demanding and complex task. This architecture primarily consists of the infrastructure that links network systems and the devices connected to this infrastructure. Passive components enable computers to communicate with one another within these systems, while electronic devices facilitate communication across the infrastructure. Examples of such electronic equipment include routers, switches, and other intermediary internet devices. Traditional Network Architecture also encompasses several complex protocols. While each new protocol may solve one issue, it often introduces another. Moreover, non-standard network devices, communication systems, and network management software from various manufacturers can operate the network and support diverse applications running on it. The introduction of a non-standard network device into the system can lead to major security flaws. For these reasons, Traditional Network Architecture is complex and difficult to manage. To address the aforementioned network issues, the concept of SDN has evolved. SDN emerges as a new network approach that proposes the separation of data and control planes, which are integrated into the infrastructure of Traditional Network Architecture. This separation simplifies the network and its management systems, providing an autonomous structure in an attempt to eliminate the problems associated with today's network infrastructure. OpenFlow is an open standards-based protocol that enables communication between SDN controllers and network devices. The controller and OpenFlow-enabled network switches are the most significant components of OpenFlow. They facilitate a centralized control plane that enables the development of a dynamic, scalable, efficient, and high-performance communication infrastructure. In addition to these advantages, SDN introduces new security concerns.

Controller coding must be done correctly for SDN to function properly. The number of controllers, for example, is a critical factor in network optimization scaling. For large-scale networks, a single SDN controller also poses issues with scalability, response time, infrastructure support, and availability. A single controller's limited capacity will not be sufficient to handle the huge data flows of tens of thousands of nodes in large networks. Finally, consideration should be given

to SDN security. When SDN is insecure, the network can be subjected to a variety of assaults.

Network attacks are classified into two types: active and passive. Active attacks are motivated by the attackers' desire to obtain the information they want. If there is a system that monitors their behavior, anything they do to visit the site will be noticed. However, if attackers are unaware of what to look for, they can carry out the attack unnoticed. Passive attacks, on the other hand, seek to obtain further information. The majority of attempts take the form of both active and passive attacks. Active attacks are the easiest to identify, yet the majority go undiscovered. Passive attacks are more difficult to detect. As a result, the chances of detecting the passive attack are almost nonexistent.

In many circumstances, attackers' resort to active attacks first in order to get access to the desired network site. Then, using a passive attack, they gain access to the information they seek. Phishing, password attacks, DoS and DDoS attacks are some examples. The most significant active attacks are denial of service attacks and hacking. Denial of service occurs when typical system users are denied service (Raktate et al., 2024). This includes everything from prohibiting people from accessing a private website to deactivating user accounts that allow them to log into the network. Any network connected to the Internet is subject to denial-of-service attacks. The tools required to carry out this type of assault are easily accessible, and such programs are not difficult to run.

DDoS attacks are a type of distributed network attack. These attacks take advantage of capacity limitations that apply to any network resource, such as the infrastructure that hosts a company's website. By making numerous requests to the attacked web resource, a DDoS assault seeks to overwhelm the website's capacity to handle huge amounts of requests and prevent it from working effectively. The ease with which DDoS assaults can be carried out at cheap cost poses a significant risk to firms who conduct their business on the internet. Organizations who are caught off guard by this form of DDoS attack may be unable to serve for an extended period of time.

Attackers often employ three types of DDoS assault tactics (network, protocol, and application). DDoS attacks on networks are the most prevalent sort of assault utilized by attackers, with the goal of consuming the internet bandwidth used by target companies. As a result, both network traffic entering and exiting the organization will be affected, and both will be unable to respond to genuine requests and will be unavailable. As a result, all parties who use the internet infrastructure will be impacted, and all internet-dependent services would be rendered inoperable. TCP/UDP flood, DNS/NTP/Memcached Amplification assaults are examples of Network Type DDoS Attacks.

Protocol assaults are broad attacks against OSI K3/K4 protocols. Targeted systems include firewalls, load balancing devices, routers, and other devices that use connection session information. Multiple login requests are submitted, and additional requests are sent before the session ends. As a result, network and

security devices will fill the session tables and become dysfunctional. Protocol assaults include SYN/SYN-ACK/ACK flood, ping of death, and others.

Application attacks include web apps, DNS, SMTP, and so on. OSI K7 attacks target application services. The main purpose is to destabilize the system by delivering requests to applications that are much over their capability and utilizing their resources. Overflow attacks on HTTP, HTTPS, DNS, and SMTP services are examples of application attacks.

Unfortunately, there is currently no conclusive and permanent strategy to avoid becoming a target of DDoS assaults. However, there are several measures that can lessen the likelihood of becoming a target as well as the severity of attacks. Separating DDoS attacks from immediate and typical system performance boosts and drops necessitates the use of the appropriate technologies and experience. In terms of businesses, among the major safeguards are a well-designed network architecture and a high degree of system and TCP/IP understanding among necessary staff. If router-level protection is given, packets destined to target systems are routed through the router before being forwarded to other systems. With this feature, routers are the first systems to be attacked, and the measures taken over the routers are critical in terms of fighting the attack from the start. If some router settings and the features of the incoming packets during the attack can be determined, the attacks can be stopped or their impacts limited with the access control list that will be established.

Considering the risks of a network attack, software or hardware tools such as firewalls, antivirus, or Intrusion Detection Systems (IDS) have been built to thwart internet-based attacks. Cyber security professionals utilize security software to secure their data and systems from malicious network users. Relying just on a firewall mechanism to protect network systems is insufficient. As a result, an IDS should be utilized in addition to the firewall. IDS are systems that automatically examine events in a computer network to detect external threats. The only way for IDS to keep up with the speed of emerging attack strategies is for the system to be constantly updated. Companies are concerned about systems that have been disconnected from the internet for an extended period of time and have not been updated. This adds to the workload of network security professionals.

This article describes a customizable modular architecture for detecting and preventing DDoS attacks. In SDN, normal traffic between computers was monitored. Then a DDoS attack was started, which was followed by another attack. In this paper, a system is created to analyze and apply Machine Learning techniques to fight DDoS attacks in SDN. The system has been tested via Mininet utilizing various test scenarios. In the 2nd part of the article, the method used in the study is presented, in the 3rd part the findings are evaluated comprehensively and in the 4th part the results obtained are discussed.

Methodology. Selected Machine Learning Algorithms. Machine learning, a subset of artificial intelligence, detects complex patterns in data and makes informed conclusions using statistical methods and computational power. In

classification problems, machine learning approaches have been successfully applied (Gangadhar and Sterbenz, 2017; Sultana et al., 2019; Ray, 2024). presented a solution for detecting and analyzing DDoS (Distributed Denial of Service) assaults on cloud computing services using Dempster-Shafer Theory (DST) and Fault Tree Analysis (FTA) for a virtual machine (VM) intrusion detection system (IDS). This solution quantitatively represents uncertainty and is effectively used in intrusion detection systems to reduce false alarm rates. Common machine learning techniques employed in intrusion detection systems, as mentioned in the literature, include Bayesian classification, Support Vector Machine (SVM), decision tree, and artificial neural networks. A study (Peng et al., 2018) identified anomalous flows in SDN (Software Defined Networking) networks, highlighting the effectiveness of Machine Learning (ML) techniques in detecting malicious flows. These techniques enable switches to use information from flow tables to monitor information flow, processing unusual flows through a detection module equipped with the DPTCM-KNN algorithm. However, this process, repeated every ten seconds, introduces significant processing overhead as a potential issue. Additionally, another study proposed a trust-based method that uses the device profile and packet data to identify malicious devices, further enhancing security measures in SDN networks.

The used method based on Decision Tree, ANN, Naive Bayes machine learning algorithms and the main steps showed in Figure 1. The proposed matching strategy based on the matched the incoming request real-time from the nodes compared with the trained classifier stored data behavior as (SDN specific dataset generated by using mininet emulator and used for traffic classification by machine learning). In addition, the proposed trained model can classified as attack traffic if it is not matched packet details with the stored behavior, which matched, in the first step.

Bayesian networks are investigated in the context of guided learning in machine learning. There is usually a pattern at work in the Bayesian classification process, and this pattern determines the previously determined classes. As an example, consider the process of recognizing superfluous messages (spam) in incoming e-mails. Spam email and non-spam email are two classes in this example. An example of machine learning with instruction is an algorithm that will utilize the spam and non-spam e-mails we have and decide if the e-mails we will receive in the future are spam or not (Çalış et al., 2013). The Bayes theorem is based on determining which factor has a greater part in the occurrence of an event involving several factors. Equation 1 shows how Bayes' theorem can be represented.

Equations should be centered in the page. Equation number should be written inside paranthesis at the right side of the equation, and should be aligned right in the page. A sample equation, Equation 1 is shown below. Equations should be cited with their full names like "as seen in Equation 1", "as seen in Equations 3 and 4".

$$□(□|□) = □(□|□)\ □(□)\ □(□) \quad (1)$$

On Equation 1

P(e) = prior probability of event e,

P(T) = a priori probability of training data T

The conditional probability of T given the event P(T|e) = e

P(e|T) = T is expressed as the conditional probability of e given the training data.

In intrusion detection systems, Bayesian classifiers are a common strategy. The highest classification success with Bayesian analysis was achieved by (Panigrahi and Patra, 2019) using the DARPA '99 dataset in their IDS. The authors achieved 99.62 percent accuracy in DoS (Denial of Service) assaults, 100 percent accuracy in information scanning attacks, 98.63 percent accuracy in U2R (User to Root) attacks, and 42.62 percent accuracy in R2L (Remote to Local) attacks in their studies. (Alhakami et al., 2019) demonstrated a 99.82 percent in accuracy in identifying normal behaviors, 99.49 percent DoS assaults, 99.72 percent in information scanning attacks, 99.47 percent in U2R attacks, and 99.35 percent in R2L attacks using the KDD Cup '99 dataset. They were successful.

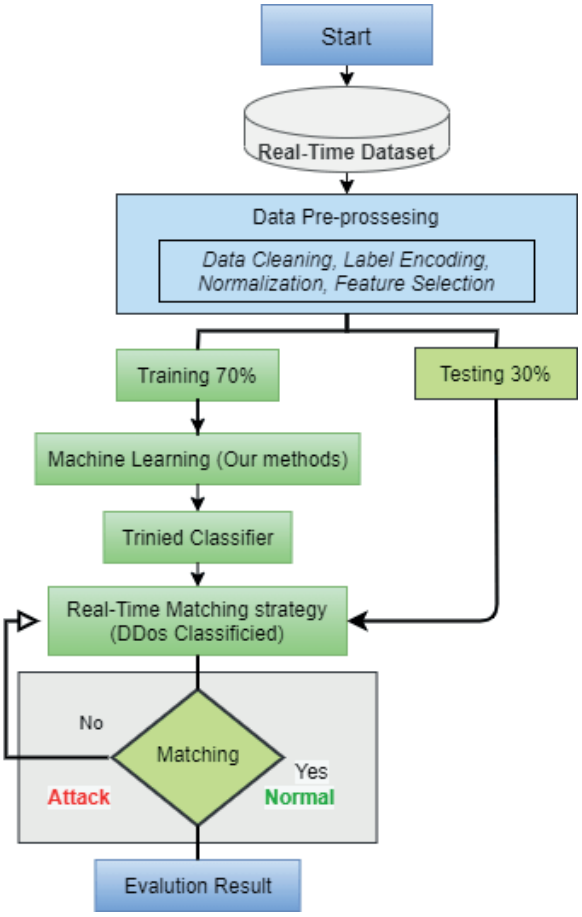


Figure 1 – Suggested machine learning system model

Huang et al., proposed the Support Vector Machine (SVM) as a powerful classifier in 2018. It is based on statistical methods. SVM is a supervised learning model that recognizes and analyzes patterns through classification and regression analysis (Huang et al., 2018). A labeled input dataset is required for SVM, which outputs two classes from the input dataset in binary classification problems. The training examples submitted fall into one of these two categories. The SVM training method creates a model to categorize new samples. The SVM model represents objects in space as points, separated by a hyperplane that is as wide as possible to differentiate the categories. New samples are positioned in this space, and their category is predicted based on which side of the hyperplane they fall on. SVM is widely used in intrusion detection systems.

Zhang, Y., and Zhu, Y. (2010) achieved high classification success with SVM in a 2010 study utilizing the KDD Cup '99 dataset, with 99.32 percent accuracy for normal behaviors, 93.81 percent for DoS attacks, 33.67 percent for information scanning attacks, and 99.42 percent for U2R attacks. Wang, J., Li, T., et al. (2010) employed an artificial bee colony algorithm alongside SVM on the KDD Cup '99 dataset, detecting DoS attacks with 99.92 percent accuracy, information search attacks with 100 percent accuracy, U2R attacks with 76 percent accuracy, and R2L attacks with 76 percent accuracy, achieving an overall classification accuracy of 87.92 percent. Mu, Q., Chen, Y., et al. (2012), in contrast, reported 100 percent success in identifying DoS assaults, information scanning attacks, and U2R attacks, along with 99.11 percent success in R2L attacks.

Decision tree are one of the most commonly used classification methods because their training and testing times are short, their findings are simple to read, and they are effective (Diana et al., 2025), Witten, I. H., and Frank, E. (2011). The classification process with decision tree is divided into two parts. The tree is generated in the initial stage. This tree structure is used to generate categorization rules in the second step. In general, the classification procedure is as follows: Let D=t1,t2,..., tn be a database, with each record represented by ti. Let C=C1,C2...Cm denote the collection of classes made up of m classes. Each Cj is a distinct class with its own set of records. That is to say, Cj=ti |=Cj, 1in, and ti D. For each entry in the database, let the fields be A1,A2,...An. Furthermore, if each record belongs to one of the C=C1,C2,...Cm classes, the decision tree can be defined as follows: The Ai field is used to identify each node. Classification rules are the nodes that connect the root node and the leaf node. When generating decision tree, the algorithm utilized is critical. Depending on the algorithm employed, the structure of the tree may alter. Different tree architectures can produce varying categorization outcomes (Ko et al., 2023). Many algorithms have been created that are based on decision tree. These algorithms are classified according to the root, node, and branching criteria. ID3, C4.5, and C5 are well-known algorithms. There are numerous STS research with decision tree in the literature. Table 1 shows the studies that were reviewed. Bahrololum, M., Salahi, E., et al. (2009) obtained classification success of 99.96 percent in identifying normal activities, 99.97 percent in detecting DOS attacks,

99.66 percent in information scanning attacks, 88.33 percent in U2R attacks, and 99.02 percent in R2L attacks using decision tree. They've done it. Furthermore, Alazab, A., Hobbs, M., et al. (2012) used the KDD CUP 99 dataset in 2012 to discriminate between 98.2 percent of normal behaviors, 97.2 percent of DOS assaults, 99.6 percent of information scanning attacks, 92.5 percent of U2R attacks, and 92.5 percent of R2L attacks. They were 99.7 percent successful. Sharma, V., and Nema, A. (2013) conducted another STS investigation utilizing decision tree in 2013 using the KDD CUP 99 dataset. Sharma and Nema detected 99.98 percent of DOS attacks, 88.19 percent of information scanning attacks, 51 percent of U2R attacks, and 94.70 percent of R2L attacks.

Artificial Neural Networks (ANN) are sophisticated categorization tools that mimic the behavior of biological nerve cells (neurons). Each network component is referred to as an artificial nerve (neuron). The neural network is made up of numerous artificial neural cells that are linked together using various weightings. A single neuron can normally only tackle linear issues. Many more categorization issues are solved using multi-layer artificial neural networks (Multi-Layer Perceptron MLP). In function fitting, classification, and matching issues, multilayer ANN is commonly employed. Because of its efficacy in classification, it has been widely employed in IDSs (Bitter et al., 2010). Today's most prevalent Multi-Layer Perceptrons (MLPs) are back propagation networks. It is now one of the most often utilized strategies, particularly in classification procedures. In backpropagation networks, the delta learning rule is utilized as a learning function. Equation 2 expresses the delta learning rule, as seen below.

$$\square_{\square,\square}(\square\square\square) = \square_{\square,\square}(\square\square\square) + (\square * [\square - \square(\square\square\square)] * \square'(\square\square\square)) \quad (2)$$

MLPs are made up of three layers: input, middleware, and output. Information enters the network through the input layer, travels through the intermediate layers to the output layer, and is sent from the output layer to the outside world. Data and results from real-world problems, or examples, are employed in the artificial neural network learning process. The variables linked to the problem form the artificial neural network's input sequence, and the actual outcomes obtained with these variables form the target output sequence that the artificial neural network must achieve (Bengio et al., 2021). Weights are modified during the learning process based on the learning style chosen. Weight change is a metaphor for learning. If there is no weight shift in ANN, the learning process comes to a halt. ANNs have shown excellent outcomes in Intrusion Detection Systems. Cui et al., (2020) constructed IDS utilizing HPCANN (Hierarchical Principal Component Analysis Neural Networks) using the KDD Cup99 dataset; their normal behavior is 97.1 percent, DOS assaults are 100 percent, information scanning attacks are 100 percent, and R2L attacks are 97.2 percent. successfully classified In 2012, Gong, X., and Guan, X. (2012) classified normal behaviors as 100 percent, DOS assaults as 100 percent,

information scanning attacks as 99.76 percent, U2R attacks as 99.85 percent, and R2L attacks as 99.88 percent using the KDD 99 dataset.

Attack and Attack Detection Scenarios. To identify and prevent DDoS attacks on SDN, attack and intrusi

on detection scenarios were built to test the effectiveness of the proposed solution, and the findings were evaluated across these three different scenarios. The findings and discussion section goes into great detail on these instances (Figure 2).

Figure 3 is an example of a DDoS assault scenario. The study employed six OpenFlow Keys. To provide a safe connection, "*A OpenFlow Key*" is connected to "ca *RYU Controller*" and "*B OpenFlow Key*" is connected to "*cb RYU Controller*." The other four OpenFlow switches each have their own networks and users. Some of the users were identified as the DDoS attackers, and the DDoS attackers were then leveraged to develop malware packages to attack the target source.

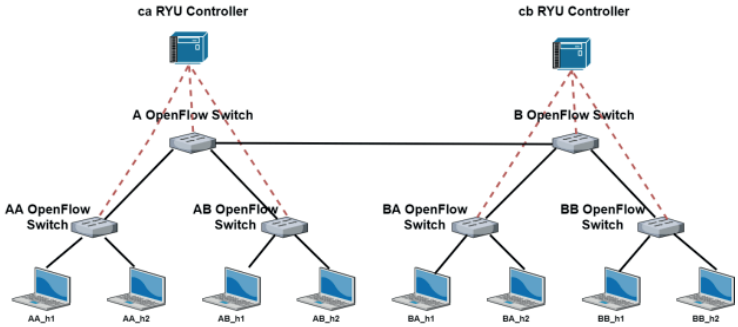


Figure 2 – SDN Topology

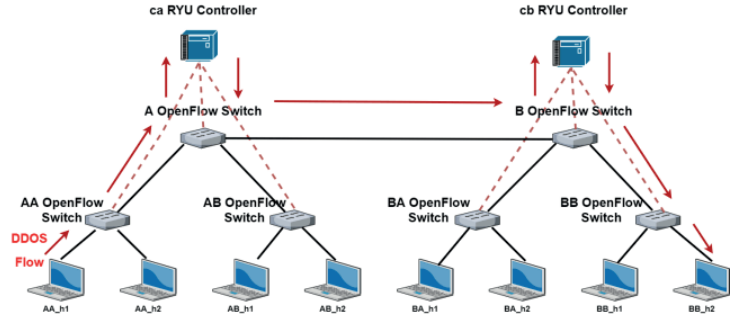


Figure 3 – Sample of DDoS Attack Scenario

The study's primary goal is to detect and prevent DDoS assaults against SDN. Machine learning techniques are used to avoid assaults by quickly developing a high-performance DDoS detection system using information such as the target IP and port number. The premise that every packet arriving at the controller is a new packet is utilized for the destination address parameter, and each incoming packet is evaluated for a valid destination address. If the destination address is invalid or

unknown on the network, the packet is sent to the stream collector. Stream collector is an SDN controller module that stores faulty packets for later analysis. Figure 4 depicts the DDoS detection and blocking system. When the accumulation of invalid packets in a given period becomes considerable, the stream collector sends a notification to the controller, which then constructs a new process for each network device to transfer the invalid packet straight to the stream collector. The stream collector then performs additional processing to the DDoS clustering time model.

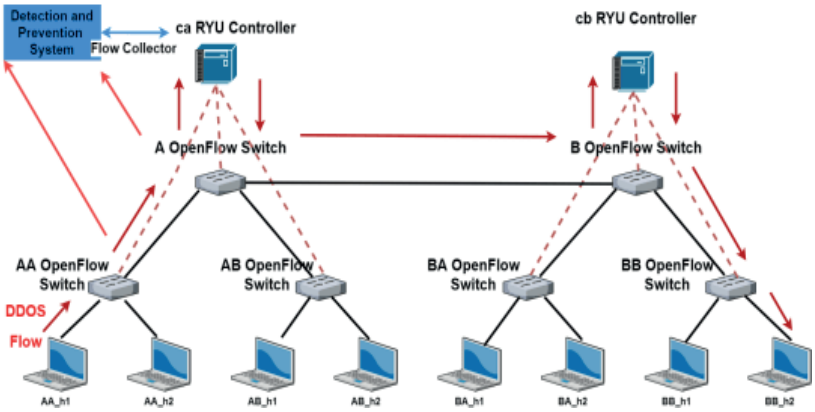


Figure 4 – DdoS Detection and Prevention System

The research universe in the study is all systems that employ SDN because it is connected to the prevention of DDoS attacks that may occur in SDN networks. The representative sample is one in which a person or organization using SDN is subjected to a DDoS attack. Various networking trends influenced the underlying concept of SDN. A properly configured SDN environment can make it easier and more cost-effective to deploy processing capacity to remote sites, move data center functions to the edge, adopt cloud computing, and support IoT settings. The relevant topic of research includes any systems in which such aims are attempted to be realized.

Evaluation Metrics. In the study, the accuracy, precision, sensitivity, and F-measure (F-Score) of algorithms were used to determine the degree of performance of the models developed. Precision and sensitivity alone are insufficient to generate a meaningful comparison result. The F criterion has been defined because assessing both criteria combined yields more accurate results.

Accuracy is defined as the ratio of correctly identified samples ($TP + TN$) to total samples ($TP+TN+FP+FN$).

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \tag{3}$$

Precision is defined as the ratio of True Positive (TP) samples predicted as class 1 to the total number of class 1 samples anticipated.

$$Precision = TP/(TP+FP) \quad (4)$$

Recall, on the other hand, is a metric that shows how much of the number of True Positive (TP) samples estimated as Class 1 is predicted as Positive.

$$Recall = TP/(TP + FN) \quad (5)$$

The precision value is an important parameter to consider when the cost of estimating as False Negative is large. Because the cost of inaccurate estimation in DDos attacks is considerable, the sensitivity for this study is expected to be as high as possible.

The harmonic mean of precision and sensitivity is the F-scale. We may not be able to draw significant comparative conclusions based solely on precision and sensitivity criteria. When both criteria are evaluated simultaneously, the findings will be more accurate. The F-scale has been defined for this purpose.

$$F\text{-scale}=(2\times Precision\times Precision)/(Sensitivity+Precision) \quad (6)$$

Findings. The findings were evaluated using three alternative scenarios. These;
I. While *AA_h1* and *BB_h1* are communicating, user *BA_h2* starts a DDoS attack against user *AA_h1*,
II. While all users are communicating, user *AA_h1* starts a DDoS attack against user *BB_h1*,
III. While *AA_h1* and *BB_h1* users are communicating, it is the identification and prevention of a DDoS attack from BA_h2 user to *AA_h1*. For this purpose, first

Used SDN Topology. Mininet was used to develop a software defined network topology. Figure 5 depicts the software defined network topology that was constructed.

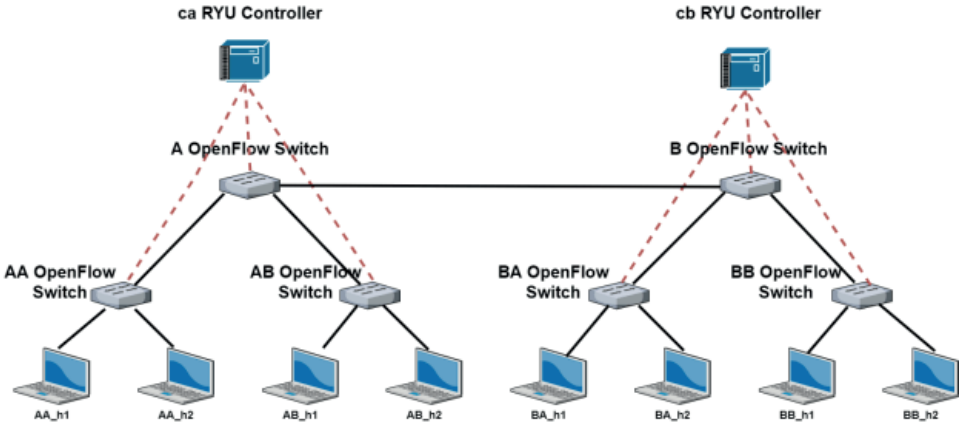


Figure 5 – Created SDN Topology (Created SDN Topology)

In the topology shown in Figure 5, there are two "RYU Controllers," whose names are "ca" and "cb." The "*A Switch*" is linked to the "ca Controller." The "*AA Switch*" and "*AB Switch*" are linked to the "*A Switch*". The "*AA Switch*" is connected to two users. These users are known as "*AA_h1*" and "*AA_h2*". The "*AB Switch*" is connected to two users. These users are known as "*AB_h1*" and "*AB_h2*". The "*B Switch*" is linked to the "cb Controller." The "*BA Switch*" and "*BB Switch*" are linked to the "*B Switch*". The "*BA Switch*" is connected to two users. These users are known as "*BA_h1*" and "*BA_h2*". The "*AB Switch*" is connected to two users. These users are known as "*BB_h1*" and "*BB_h2*."

SDN Normal Traffic Flow. Figure 6 depicts regular network traffic on software defined networks. Pinging BA h1 from user AA h1 was used to test normal flow.

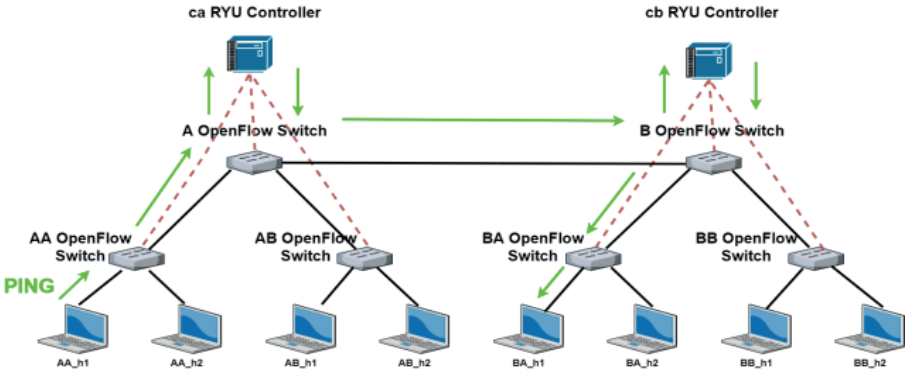


Figure 6 – Normal Traffic Flow in Software Defined Networks

Figure 7 depicts normal traffic resource use on software-defined networks using sFlow.

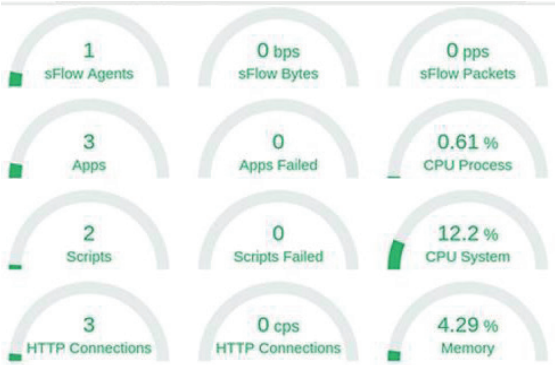


Figure 7 – SDN Normal Traffic Monitored over sFlow

Performing User-to-User DDoS Attack on Software Defined Networks. DDoS refers to the temporary or permanent disruption of a computer's functions. As a result, a DDoS assault is a cyber attack that seeks to prevent genuine users from accessing the computer's network resources (Gangadhar and Sterbenz, 2017).

Different scenarios, such as two DDoS attacks, are presented in this study. These are "DDoS Attack of User *AB h*1 Against User *AA h1* While AA h1 and *BB h1* Communicate" and "DDoS Attack of User *AA h1* Against User *BB h1* While All Users Communicate."

Scenario 1: While user *AA_h1* pings User *BB_h1* during their communication, user *AB_h1* starts a DDoS attack on User *AA_h1* . Figure 8 depicts it. By starting a DDoS attack against user *AA_h1*, user *BA_h2* has depleted *AA_h1*'s resources (Figure 9). The attack has prevented users of *AA_h1* and *BB_h1* from communicating.

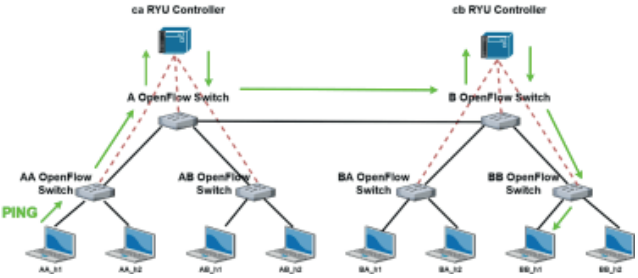


Figure 8 – Normal Communication of *AA_h1* and *BB_h1* Users

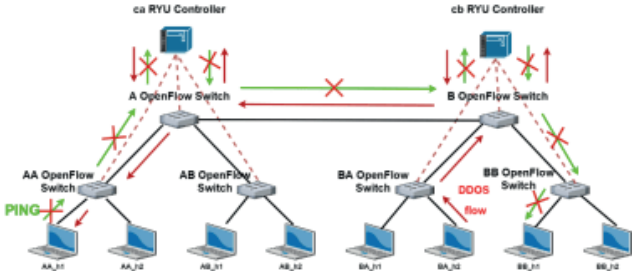


Figure 9 – Failure of *AA_h1* and *BB_h1* Users to Communicate as a Result of DDoS Attack from *BA_h2* Endpoint to User *AA_h1*

Figure 10 – depicts schematically the DDoS assault by user *BA_h2* against user *AA_h1*.

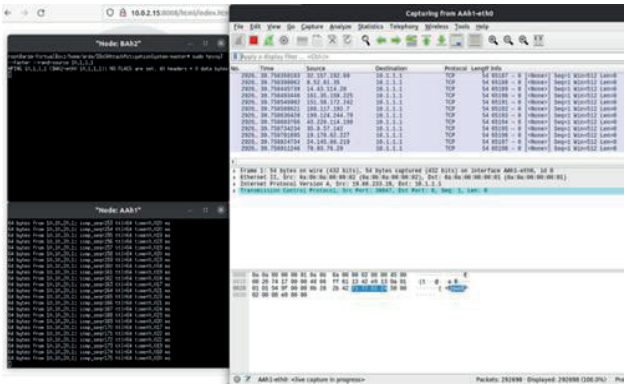


Figure 10 – Attack by user *BA_h2* Against User *AA_h1* Through DdoS

DDoS attacks are displayed using Wireshark. Wireshark was used for packet analysis. DDoS attacks are also depicted visually in Figure 11.



Figure 11 – Wireshark I/O Graph of DDoS Attack of User *BA_h2* to User *AA_h1*

Scenario 2: While all users were communicating, a DDoS attack was started against users *AA h1* and *BB h1*. Ping was used to communicate between network users (Figure 10), and the ping test was monitored on sFlow (Figure 12).

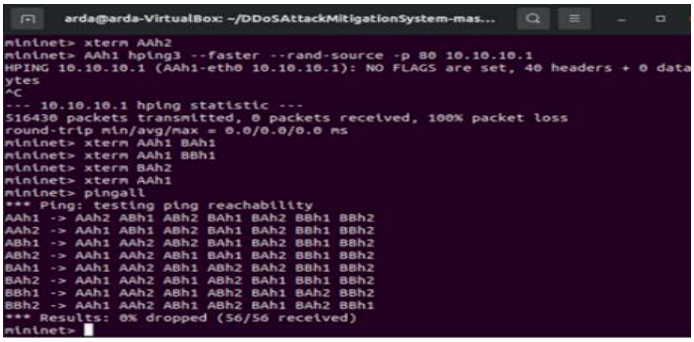


Figure 12 – Communication of Users on the Network

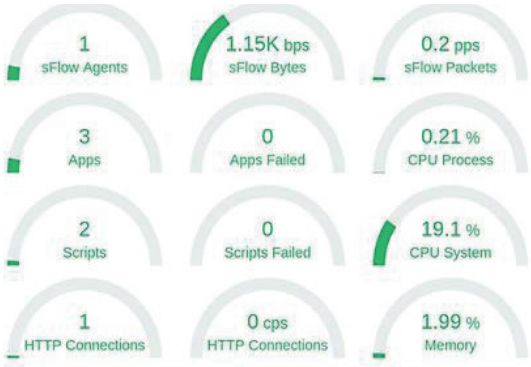


Figure 13 – sFlow Image During Ping Test to Network Users

As seen in Figure 14, a DDoS attack was started by user *AA_h1* against user *BB_h1*.

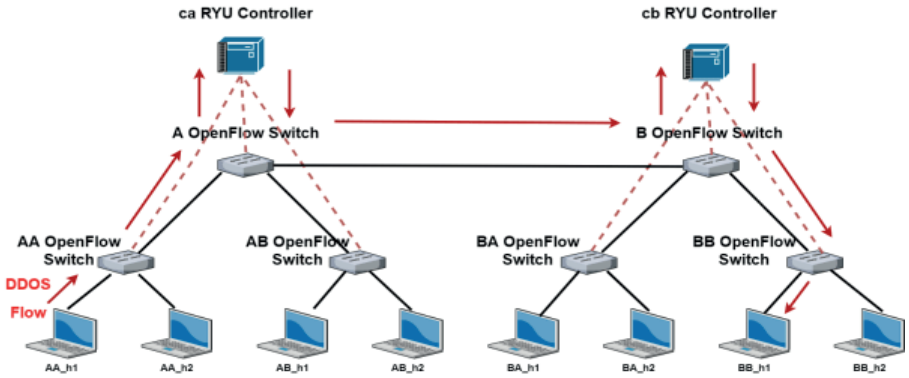


Figure 14 – Starting a DDoS attack on User *BB_h1* by User *AA_h1*

Wireshark was used to perform packet analysis on the DDoS attack output. Figure 14 depicts a DDoS attack that was carried out.

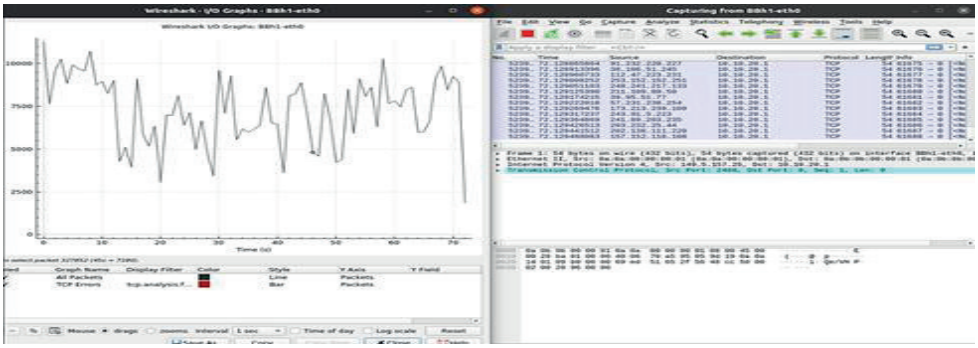


Figure 15 – In Wireshark, user *AA h1* is showing a DDoS attack to user *BB h1*.

As seen in the sFlow graph in Figure 15, DDoS attack consumes the computer's resources.

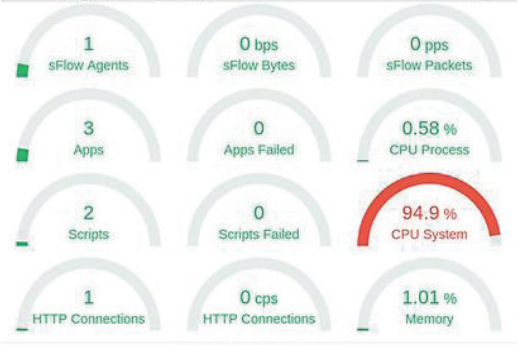


Figure 16 – Showing AA h1's DDoS Attack on BB h1 in sFlow

DDoS Attack Detection in SDN. Machine learning algorithms are used for DDoS attack detection in SDN. The algorithms used were trained and tested using the KDD CUP 99 dataset. For this purpose, KDD CUP 99 dataset was divided into two parts as 70% training dataset and 30% test dataset by hold-out method. A model was created by giving the training data set samples to the Decision Trees machine learning algorithm. This created model has two classes, normal and DDoS. One of these classes shows the normal traffic of the network, while the other class shows a network that has suffered a DDoS attack. All types of DDoS attacks fall under the DDoS class. This classification holds true for Naive Bayes and ANN models as well. Using the KDD CUP 99 dataset, a model was developed using the Naive Bayes machine learning technique. The ANN machine learning technique and the KDD CUP 99 dataset were used to develop a model. However, because the decision tree model has a 91 percent accuracy rate, the decision tree is used.

The Decision Tree model outperformed the ANN (78%) and Naive Bayes models (68%). As a result, the decision tree was selected as the model to be employed in the DDoS detection procedure. The program developed in the study, on the other hand, is structured in a modular structure, allowing DDoS and other intrusion detection systems to be moved to the controller.

Prevention of DDoS Attacks on SDN. After the DDoS assault was identified, the bandwidth was reduced and the computer's network resources were accessible to the principal users.

Scenario 3: While *AA_h1* and BB h1 users were conversing, a DDoS assault from *BA_h2* to *AA_h1* was detected and prevented. User *AA_h1* is pinging user *BB_h1*. User BA_h2 attacked to drain *AA_h1*'s resources by launching a DDoS assault on *AA_h1*. Communication between user *AA_h1* and user *BB_h1* was ensured by recognizing and blocking the assault. Figure 17 depicts it.

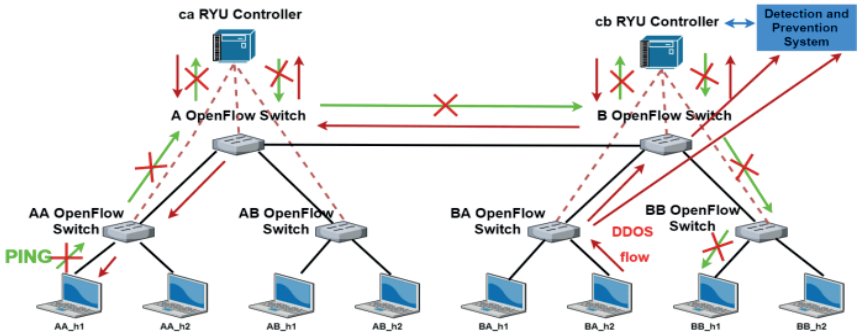


Figure 17 – Ping Test of User *AA_h1* to User *BB_h1* and DDoS Attack Prevention

Results and Discussion. In this work, an artificial intelligence-based application based on SDN was created to detect and prevent DDoS attacks by modeling them on a platform. The security of the network topology is assured by this built application, and precautions are made against vulnerabilities that the configured

network and users may generate. The system was designed in a modular framework for upgrades, enabling the low-cost incorporation of various artificial intelligence models into the system.

The program has been turned into software that requires less human participation in order for the trained models to detect assaults using machine learning methods. In DDoS attacks, a temporary solution is to increase bandwidth so that the user can react to requests. The study mostly detects DDoS attacks. Although shutting the attacked port to avoid DDoS assaults is one of the first solutions that come to mind, controlling bandwidth is recommended since communication cannot be done via the closed port, and so the DDoS attack will succeed. After the DDoS attack was identified, the bandwidth was raised so that the DDoS victim could connect with other users. Table 1 shows the bandwidth that was consumed.

Table 1 – Bandwidth

Bandwidth before DDoS attack	Bandwidth After DDoS Attack	Bandwidth After DDoS Attack Blocked
10bps	4000bps	10bps

According to the predictions, the Decision Tree classifier fared better in terms of the Accuracy measure. However, because making a judgment based on a single statistic can be misleading, the recall and acuity metrics should also be considered. The recall metric value of the Naive Bayes classifier was lower than that of the Decision Tree and ANN. Table 2 displays the metric values. Using the Mininet emulator within a tree network architecture, we conducted a DDoS (Distributed Denial of Service) attack on the Ryu controller (Mehr, S. Y., and Ramamurthy, B., 2019). By employing machine learning technology, specifically support vector machines (SVM), we were able to identify DDoS attacks through the installation of flows in switches. This method allows for the evaluation of the time attack pattern of the DDoS assault, thereby enhancing our detection capabilities. Leveraging our detection technology, we successfully minimized the impact of DDoS assaults on the Ryu controller by 36%. Evaluation of Estimation Algorithms (Test Dataset). The results of tests suggest that it can detect DDoS attacks with a high degree of accuracy.

2 – Evaluation of Prediction Algorithms-Test Dataset

Prediction Algorithms	**Accuracy (%)**	**Recall**	**Precision**	**F_Score**
Decision Tree	91	0.85	0.90	0.87
ANN	78	0.62	0.96	0.75
Naive Bayes	68	0.71	0.95	0.81

Conclusion. The article under review addresses the acute vulnerability of Software-Defined Networks (SDN) to distributed denial-of-service (DDoS) attacks

by proposing a modular, AI-driven detection and prevention framework. It begins by outlining the inherent separation of control and data planes in SDN, which-while simplifying management-exposes the control layer to volumetric and protocol-based DDoS threats.

Using the Mininet emulator, the authors constructed a multi-switch topology managed by dual Ryu controllers to simulate normal and attack traffic flows. They executed three distinct attack scenarios-targeting individual hosts under active communication-to capture packet traces via Wireshark and sFlow. Machine learning classifiers (Decision Tree, Artificial Neural Network, and Naive Bayes) were trained on the KDD Cup ’99 dataset (70% training, 30% testing) to distinguish between normal and DDoS traffic.

The modular architecture-allowing easy reintegration of improved models trained off-line on platforms like Google Colaboratory-stands out as a practical strength, potentially reducing retraining overhead. However, reliance on the outdated KDD Cup ’99 dataset may limit adaptability to evolving DDoS tactics. Experimental validation remains confined to emulated topologies; real-world deployment on production SDN controllers (e.g., OpenDaylight) would better gauge performance overhead and scalability.

This proof-of-concept demonstrates that a Decision Tree–based SI-SDN (Security Intelligence for SDN) module can detect and mitigate DDoS with high accuracy and low latency. Future work should employ contemporary, diversified traffic datasets (e.g., CIC-IDS2017), extend evaluations to hybrid cloud-edge environments, and assess integration overhead on commercial SDN controllers. Such enhancements will solidify the framework’s readiness for operational network defenses.

References

Alazab A., Hobbs M., Abawajy J., and Alazab M. (2012) Using feature selection for intrusion detection system. In International Symposium on Communications and Information Technologies (ISCIT). — P. 296–301. IEEE. https://doi.org/10.1109/ISCIT.2012.6380910 (in Eng.).

Alhakami W., Alharbi A., Bourouis S., Alroobaea R., and Bouguila N. (2019) Network anomaly intrusion detection using a nonparametric Bayesian approach and feature selection. IEEE Access, 7. — P. 52181–52190. https://doi.org/10.1109/ACCESS.2019.2912115 (in Eng.)

Bahrololum M., Salahi E., and Khalegni M. (2009) Machine learning techniques for feature reduction in intrusion detection systems: A comparison. In Fourth International Conference on Computer Sciences and Convergence Information Technology. IEEE. https://doi.org/10.1109/ICCIT.2009.89 (in Eng.)

Bengio Y., LeCun Y., and Hinton G. (2021) Deep learning for AI. Communications of the ACM, 64(7). — P. 58–65. https://doi.org/10.1145/3448250 (in Eng.)

Bitter C., Elizondo D.A., and Watson T. (2010) Application of artificial neural networks and related techniques to intrusion detection. In International Joint Conference on Neural Networks (IJCNN). IEEE. https://doi.org/10.1109/IJCNN.2010.5596532 (in Eng.)

Çalış K., Gazdağı O., and Yıldız O. (2013) Reklam içerikli epostaların metin madenciliği yöntemleri ile otomatik tespiti [Automatic detection of advertising e-mails using text mining methods]. Bilişim Teknolojileri Dergisi, 6(1). (in Turkish).

Cui Y., Jin Z., and Hu J. (2020) Research on intrusion detection method based on hierarchical

self-convergence PCA-OCSVM algorithm. International Journal of Network Security, 22(6). — P. 916–924. https://doi.org/10.6633/IJNS.202011_22(6).04 (in Eng.)

Diana L., Dini P., and Paolini D. (2025). Overview on intrusion detection systems for computers networking security. Computers, 14(3), Article 87. https://doi.org/10.3390/computers14030087 (in Eng.)

Gangadhar S., and Sterbenz J.P.G. (2017) Machine learning aided traffic tolerance to improve resilience for software defined networks. In 2017 9th International Workshop on Resilient Networks Design and Modeling (RNDM). — P. 1–7. IEEE. https://doi.org/10.1109/RNDM.2017.8093035 (in Eng.)

Gong X., and Guan X. (2012) Intrusion detection model based on the improved neural network and expert system. In IEEE Symposium on Electrical and Electronics Engineering (EEESYM). IEEE. https://doi.org/10.1109/EEESym.2012.6258621 (in Eng.)

Huang W., Li G.-M., and Chen W.-W. (2018) A review of statistical learning theory. DEStech Transactions on Engineering and Technology Research. https://doi.org/10.12783/DTETR/PMSMS2018/24953 (in Eng.)

Ko K., Baek J., Seo B.-S., and Lee W.-B. (2023) Comparative study of AI-enabled DDoS detection technologies in SDN. Applied Sciences, 13(17), Article 9488. https://doi.org/10.3390/app13179488 (in Eng.)

Mehr S.Y., and Ramamurthy B. (2019) An SVM-based DDoS attack detection method for Ryu SDN controller. In Proceedings of the 15th International Conference on Emerging Networking EXperiments and Technologies. — P. 72–73. ACM. https://doi.org/10.1145/3360468.3368183 (in Eng.)

Mu, Q., Chen, Y., and Zhang, Y. (2012). Incremental SVM algorithm to intrusion detection based on boundary areas. In International Conference on Systems and Informatics. IEEE. https://doi.org/10.1109/ICSAI.2012.6223447 (in Eng.)

Panigrahi, A., and Patra, M. R. (2019). Anomaly based network intrusion detection using Bayes net classifiers. International Journal of Scientific and Technology Research, 8(9). — P. 481–485. https://www.ijstr.org/paper-references.php?ref=IJSTR-0919-21989 (in Eng.)

Peng H., Sun Z., Zhao X., Tan S., and Sun Z. (2018) A detection method for anomaly flow in software-defined network. IEEE Access, 6. — P. 27809–27817. https://doi.org/10.1109/ACCESS.2018.2839684 (in Eng.)

Raktate G., Shelar K., Parjane P., Pangavhane S., More S., and Deshmukh S. (2024) A survey on security issues and challenges in cloud computing. — P. 1–5. https://doi.org/10.1109/DASA63652.2024.10836628 (in Eng.)

Ray D. (2024) Internet traffic classification using machine learning techniques. https://doi.org/10.55041/isjem01429 (in Eng.)

Sharma V., and Nema A. (2013) Innovative genetic approach for intrusion detection by using decision tree. In International Conference on Communication Systems and Network Technologies (CSNT). IEEE. https://doi.org/10.1109/CSNT.2013.93 (in Eng.)

Sultana N., Chilamkurti N., Peng W., and Alhadad R. (2019) Survey on SDN based network intrusion detection system using machine learning approaches. Peer-to-Peer Networking and Applications, 12(2). — P. 493–501. https://doi.org/10.1007/s12083-017-0630-0 (in Eng.)

Wang J., Li T., and Ren R. (2010) A real time IDSs based on artificial bee colony support vector machine algorithm. In Third International Workshop on Advanced Computational Intelligence. IEEE. https://doi.org/10.1109/IWACI.2010.5585107 (in Eng.)

Witten I.H., and Frank E. (2011) Data mining: Practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann. (in Eng.)

Zhang Y., and Zhu Y. (2010) Application of improved support vector machines in intrusion detection. In 2nd International Conference on e-Business and Information System Security. IEEE. https://doi.org/10.1109/EBISS.2010.5473653 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 52–72

https://doi.org/10.32014/2026.2518-1726.401

IRSTI 31.15.37
UDC 621.546.06

© Amanzholova S.T. [1,2], Ussatova O.A. [1,2*], Mutanov G.M. [2], Mukhanov S.B.[1,2], Aitmukash D.[2], 2026.
[1]Astana IT University, Astana, Kazakhstan;
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan.
E-mail: olgaussatova@gmail.com

BACKEND ARCHITECTURE OF A HYBRID BLOCKCHAIN-BASED ACADEMIC CREDENTIAL VERIFICATION SYSTEM

Amanzholova Saule — Candidate of technical science, Astana IT University, Astana, Kazakhstan,
E-mail: s.amanzholova@astanait.edu.kz, https://orcid.org/0000-0002-6779-9393;
Ussatova Olga — PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan,
E-mail: olgaussatova@gmail.com, https://orcid.org/0000-0002-5276-6118;
Mutanov Galimkair — Doctor of technical science, Institute of Information and Computational Technologies, Almaty, Kazakhstan,
E-mail: mu-tanov@ipic.kz, https://orcid.org/0000-0002-1375-1343;
Mukhanov Samat – PhD, Astana IT University, Astana, Kazakhstan,
E-mail: s.mukhanov@astanait.edu.kz, https://orcid.org/0000-0001-8761-4272;
Aitmukash Daniyar — PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan,
E-mail: Aitmukash.Daniyar@ipic.kz, https://orcid.org/0000-0001-9904-9979.

Abstract. The rapid digitalization of higher education institutions worldwide has created a growing demand for transparent, trustworthy, and secure solutions for the storage, management, and verification of academic data. Universities are increasingly transitioning to electronic document management systems for transcripts, diplomas, and other educational records. However, traditional centralized information systems used in educational institutions remain vulnerable to significant security and integrity issues, including unauthorized access to academic records, lack of transparent audit mechanisms, risks of transcript forgery, and limited trust in cross-institutional verification procedures. To address these challenges, this paper proposes a hybrid architecture that combines OffChain storage with a consortium blockchain-based OnChain verification layer. In the proposed model, the OffChain component is responsible for bulk data storage, metadata processing, and efficient management of educational records, while the

OnChain component stores cryptographic hashes of transcripts and immutable verification metadata. Such a design makes it possible to preserve the performance and scalability of conventional systems while simultaneously leveraging the decentralization, immutability, and trust guarantees provided by blockchain technology. The implemented architecture follows Clean Architecture principles and consists of multiple interconnected layers, including role-based access control, a secure API gateway for service orchestration, an authentication mechanism based on asymmetric cryptography, and a hybrid data dispersal module. In addition, smart contracts are designed to automate the verification and validation of transcripts and diplomas, thereby reducing the need for manual checks and minimizing the risk of tampering. The proposed solution improves data integrity, strengthens institutional trust, and enhances the efficiency of academic credential verification for employers, government agencies, and foreign universities. The research demonstrates that the integration of blockchain with existing educational information systems can provide a practical and scalable foundation for secure digital academic services.

Keywords: Blockchain, Academic Data Verification, Backend Architecture, Hybrid Data Storage, Smart Contracts, Data Security, API Integration

For citations: Amanzholova S.T., Ussatova O.A., Mutanov G.M., Mukhanov S.B., Aitmukash D. Backend architecture of a hybrid blockchain-based academic credential verification system. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 52–72. DOI: https://doi.org/10.32014/2026.2518-1726.401

Financing. *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. BR24993014 The development of an intelligent anti-corruption system for information protection, validation of the results of educational achievements, official documents of students and graduates of universities in the Republic of Kazakhstan.*

© Аманжолова С.Т. [1,2], Усатова О.А.[1,2*], Мутанов Г.М. [2], Муханов С.Б.[1,2], Айтмукаш Д. [2], 2026.

[1]Астана IT Университеті, Астана, Қазақстан;

[2] Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан.

E-mail: olgaussatova@gmail.com

ГИБРИДТІК БЛОКЧЕЙНГЕ НЕГІЗДЕЛГЕН АКАДЕМИЯЛЫҚ СЕНІМДІК ДЕРЕКТЕРДІ ТЕКСЕРУ ЖҮЙЕСІНІҢ БЕКЕНДІК АРХИТЕКТУРАСЫ

Аманжолова Сауле — техника ғылымдарының кандидаты, Астана IT Университеті, Астана, Қазақстан,

E-mail: s.amanzholova@astanait.edu.kz, https://orcid.org/0000-0002-6779-9393;

Усатова Ольга — PhD, Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан,
E-mail: olgaussatova@gmail.com , https://orcid.org/0000-0002-5276-6118;
Мутанов Галимкаир — техника ғылымдарының докторы, Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан,
E-mail: mu-tanov@ipic.kz, https://orcid.org/0000-0002-1375-1343;
Муханов Самат — PhD, ассистент-профессор, Астана IT Университеті, Астана, Қазақстан,
E-mail: s.mukhanov@astanait.edu.kz, https://orcid.org/0000-0001-8761-4272;
Айтмукаш Данияр — PhD, Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан,
E-mail: Aitmukash.Daniyar@ipic.kz, https://orcid.org/0000-0001-9904-9979.

Аннотация. Әлем бойынша жоғары оқу орындарындағы цифрландыру үдерісінің қарқынды дамуы академиялық деректерді сақтау, басқару және тексеру үшін ашық, сенімді және қауіпсіз шешімдерге деген сұранысты арттыруда. Университеттер транскрипттерді, дипломдарды және басқа да білім беру құжаттарын сақтау үшін электрондық құжат айналымы жүйелеріне белсенді түрде көшу үстінде. Алайда білім беру ұйымдарының дәстүрлі орталықтандырылған ақпараттық жүйелері бірқатар маңызды мәселелерге ұшырайды, оның ішінде деректерге рұқсатсыз қол жеткізу, аудиттің ашық механизмдерінің болмауы, академиялық құжаттарды қолдан жасау тәуекелдері және ұйымдар арасындағы тексеру мүмкіндіктерінің шектеулілігі бар. Осы мәселелерді шешу мақсатында зерттеуде OffChain деректерді сақтау және консорциумдық блокчейнге негізделген OnChain компонентін біріктіретін гибридті архитектура ұсынылады. Ұсынылған модельде OffChain бөлігі үлкен көлемдегі деректерді сақтау мен метадеректерді өңдеуге жауап береді, ал OnChain компоненті транскрипттердің криптографиялық хэштерін және өзгермейтін верификация деректерін сақтайды. Бұл тәсіл дәстүрлі жүйелердің өнімділігі мен ауқымдылығын сақтай отырып, блокчейн технологиясының орталықсыздандырылған, өзгермейтін және сенімді сақтау артықшылықтарын пайдалануға мүмкіндік береді. Ұсынылған архитектура Clean Architecture қағидаттарына сәйкес әзірленген және бірнеше функционалдық деңгейлерден тұрады: рөлдерге негізделген қол жеткізуді басқару, сервистердің өзара әрекеттесуіне арналған қорғалған API-шлюз, асимметриялық криптографияға негізделген аутентификация механизмі және деректерді гибридті тарату модулі. Сонымен қатар, академиялық құжаттарды тексеру және растау үдерістерін автоматтандыратын смарт-контрактілер енгізілген, бұл қолмен тексеруге тәуелділікті азайтып, деректердің бұрмалану қаупін төмендетеді. Ұсынылған шешім деректердің тұтастығын арттырады, білім беру экожүйесіндегі сенімді күшейтеді және жұмыс берушілер, мемлекеттік органдар мен шетелдік университеттер үшін академиялық жетістіктерді тексеру үдерістерінің тиімділігін жоғарылатады. Зерттеу нәтижелері блокчейн технологиясын білім беру ақпараттық жүйелерімен біріктіру қауіпсіз және ауқымды цифрлық білім беру қызметтерін құру үшін тиімді негіз бола алатынын көрсетеді.

Түйін сөздер: Блокчейн, академиялық деректерді тексеру, серверлік архитектура, гибридті деректерді сақтау, ақылды келісімшарттар, деректер қауіпсіздігі, API интеграциясы

© Аманжолова С.Т.[1,2], Усатова О.А.[1,2*], Мутанов Г.М.[2], Муханов С.Б.[1,2], Айтмукаш Д. [2], 2026.
[1]Astana IT University, Астана, Казахстан;
[2] Институт информационных и вычислительных технологий, Алматы, Казахстан.
E-mail: olgaussatova@gmail.com

БЭКЕНД-АРХИТЕКТУРА ГИБРИДНОЙ СИСТЕМЫ ПРОВЕРКИ АКАДЕМИЧЕСКИХ ДОСТИЖЕНИЙ НА ОСНОВЕ БЛОКЧЕЙНА

Аманжолова Сауле — кандидат технических наук, Астана IT Университет, Астана, Казахстан, E-mail: s.amanzholova@astanait.edu.kz, https://orcid.org/0000-0002-6779-9393;
Усатова Ольга — PhD, Институт информационных и вычислительных технологий, Алматы, Казахстан,
E-mail: olgaussatova@gmail.com , https://orcid.org/0000-0002-5276-6118;
Мутанов Галимкаир — доктор технических наук, Институт информационных и вычислительных технологий, Алматы, Казахстан,
E-mail: mu-tanov@ipic.kz, https://orcid.org/0000-0002-1375-1343;
Муханов Самат — PhD, ассистент-профессор, Астана IT Университет, Астана, Казахстан,
E-mail: s.mukhanov@astanait.edu.kz, https://orcid.org/0000-0001-8761-4272;
Айтмукаш Данияр — PhD, Институт информационных и вычислительных технологий, Алматы, Казахстан,
E-mail: Aitmukash.Daniyar@ipic.kz, https://orcid.org/0000-0001-9904-9979.

Аннотация: Быстрые темпы цифровизации высших учебных заведений во всём мире обусловили рост потребности в прозрачных, надёжных и безопасных решениях для хранения, управления и верификации академических данных. Университеты всё чаще переходят к использованию электронных систем управления документами для хранения транскриптов, дипломов и других образовательных записей. Однако традиционные централизованные информационные системы образовательных организаций остаются уязвимыми к ряду существенных проблем, включая несанкционированный доступ к данным, отсутствие прозрачных механизмов аудита, риски подделки академических документов и ограниченные возможности межорганизационной проверки. Для решения указанных проблем в работе предлагается гибридная архитектура, сочетающая off-chain хранение данных и on-chain компонент на основе консорциумного блокчейна. В предложенной модели off-chain часть отвечает за хранение больших объёмов данных и обработку метаданных, тогда как on-chain компонент обеспечивает хранение криптографических хэшей транскриптов и неизменяемых данных верификации. Такой подход позволяет сохранить производительность и масштабируемость традиционных систем,

одновременно используя преимущества децентрализованного, неизменяемого и доверенного хранения данных на основе технологии блокчейн. Разработанная архитектура соответствует принципам Clean Architecture и включает несколько функциональных уровней: управление доступом на основе ролей, защищённый API-шлюз для взаимодействия сервисов, механизм аутентификации на основе асимметричной криптографии, а также модуль гибридного распределения данных. Дополнительно реализованы смарт-контракты, автоматизирующие процессы проверки и валидации академических документов, что снижает зависимость от ручных процедур и минимизирует риски подделки данных. Предложенное решение повышает целостность данных, усиливает доверие между участниками образовательной экосистемы и повышает эффективность процессов проверки академических достижений для работодателей, государственных органов и зарубежных университетов. Результаты исследования показывают, что интеграция блокчейн-технологии с существующими информационными системами образования может служить эффективной и масштабируемой основой для создания безопасных цифровых образовательных сервисов.

Ключевые слова: Блокчейн, проверка академических данных, архитектура бэкэнда, гибридное хранение данных, смарт-контракты, безопасность данных, интеграция API

Introduction. The last few years have witnessed a marked digitalization of the educational space and an intense development of national information systems with the aim of automating academic management processes, accumulating educational data, and providing public services in the electronic format. Higher education institutions produce considerable amounts of digital data, including data related to students, their performance, and academic and qualification paths. The reliability, integrity, and accessibility of such data have come to be seen as key drivers of sustainability in the digital educational space.

Despite the achieved level of digitalization, most existing educational platforms retain centralized data storage architecture. This model provides ease of administration and operational processing of information but comes with several systemic limitations. These include the risk of unauthorized modification of records, the lack of an immutable transaction history, vulnerability to internal and external intrusion, and dependence on a single storage center (Alwakid and Dahri, 2025). In the context of academic qualifications, these limitations take on special meaning, as the authenticity of diplomas and transcripts has direct implications for the trustworthiness of employers, government agencies, and international educational organizations (Alzahrani et al., 2025).

The development of academic mobility, the globalization of educational programs, and the computerization of government agencies are creating an increased need for effective means of validating educational credentials. Traditional means of qualification verification are time-consuming, involve inter-institutional

requests, and do not offer real-time verification of data authenticity (Alzhrani et al., 2023). This creates a demand for technological solutions that can guarantee the immutability of records and the transparency of their origin.

One such solution is the use of distributed ledgers based on blockchain technology. Blockchain provides decentralized data storage solutions, cryptographic protection of transactions, and immunity of recorded data against retroactive change. In the context of the education sector, it is seen as having potential for boosting trust in digital qualifications, facilitating verification processes, and preventing document forgery (Amanzholova et al., 2026).

In light of the global digitization of various countries' education systems, it is interesting to look at the experiences of various national higher education management platforms. There are various countries that have implemented centralized data storage solutions for academic qualifications and educational achievements with the aim of boosting data transparency and accessibility. Similar solutions have been implemented in various countries in Europe, the USA, South Korea, and Singapore, where various digital ecosystems in education have been integrated with government and employment services and qualification recognition systems (Amanzholova et al., 2026). However, most existing infrastructures retain a centralized storage model, which limits the ability to immutably record data and transparently audit changes.

In the Republic of Kazakhstan, the advancement of digital educational infrastructure is underway through the Unified Higher Education Platform (UHEP). The UHEP is intended to aggregate, monitor, and analyze academic data from universities (Alammary et al., 2026). It combines the educational learning management systems used by educational institutions, centralizes the storage of information on students, graduates, and educational programs, and supports government services for education. Despite the high level of digitalization and the scale of integration, the UHEP architecture maintains a centralized data storage system, which does not fully ensure the immutability of academic records, the transparency of their modifications, or the cryptographic verifiability of educational documents.

In this regard, the integration of distributed ledgers into national educational platforms is seen as a promising direction for the development of trusted digital infrastructure. The use of hybrid blockchain architecture allows for the addition of an immutable storage layer for critical data to centralized systems, ensuring a balance between operational efficiency and cryptographic protection of academic qualifications.

A pressing scientific and practical challenge in this context is the development of a backend architecture capable of ensuring secure orchestration of academic data flows, access control, cryptographic security, and subsequent verification of records in a distributed network (de Alwis et al., 2025). Particularly challenging is the construction of a hybrid storage model that combines the operational efficiency of centralized databases with the immutability of a blockchain ledger.

The purpose of this study is to develop and implement backend architecture for a hybrid blockchain platform that ensures secure verification, immutability, and scalability of academic data storage in a distributed digital educational environment. The study proposes a multi-tier server infrastructure architecture that implements Clean Architecture principles, a flexible role-based access model, a data distribution mechanism between Off-Chain and On-Chain storage, and integration with blockchain network smart contracts for transcript and diploma storage.

Literary review. The use of blockchain technologies to store and verify educational qualifications has been found to be an ongoing research trend over the years, with an increase in academic mobility and the need to digitally verify academic diplomas and transcripts. In this regard, a systematic review by de Alwis et al. (2025), which examined 63 published works on blockchain technologies applied to educational accreditation and qualifications management, found an increase in research on the use of digital records of academic diplomas and certificates and the automation of academic qualification verification processes (Dey and Shekhawat, 2021). It is argued that this would improve transparency regarding the provenance of documents and reduce the risk of document forgery during inter-institutional and international recognition.

Similar findings were made by Alammary et al. (2019) in a systematic review on various blockchain technologies applied to education, categorized by their objectives of application: integrity preservation, verification ease, and reducing intermediaries. Earlier conceptual works by Grech and Camilleri (2017) consider blockchain as an infrastructural foundation for global educational registries, while Hindarto analyze a model of distributed digital portfolios and student reputation mechanisms. These studies form a consensus: the key value of blockchain in educational systems is not related to storing a complete array of data, but to creating a trusted layer for confirming the authenticity and origin of (Hindarto, 2023).

A separate body of publications is devoted to the use of smart contracts to automate the registration and verification of qualifications. (Hindarto, 2023) demonstrate the potential of a smart contract model for recording diploma issuance events and ensuring the verifiability of records in a distributed ledger (Hindarto, 2023). The practice-oriented development of this area is reflected in the BACIP (Blockchain-Based Academic Credential Interoperability Protocol), which proposes a comprehensive framework for managing educational achievements with an emphasis on the registration, storage, and verification of qualifications in a distributed environment (Iswahyudi et al., 2023) For the purposes of this study, these works are important as confirmation of the applicability of smart contracts specifically for the proof layer, while the bulk of operational data is rationally located outside the blockchain (Iswahyudi et al., 2023).

A critical challenge for the implementation of blockchain solutions remains the limitations of throughput and storage costs in distributed networks. In the review by Zheng et al. (2018), scalability, transaction confirmation delays, and the resource intensity of operations in distributed ledgers are highlighted among the key issues,

which justify the need for architectural optimization and selective data placement (Jha et al., 2023). In practice, this leads to the spread of hybrid OnChain/OffChain architectures, in which only the minimum information sufficient for verification (e.g., hash identifiers or key attributes) is transferred to the blockchain, while metadata and documents are stored in centralized databases or external storage. Thus, the literature supports an approach in which the distributed ledger is used as an immutable confirmation mechanism, rather than as the primary storage platform for the entire lifecycle of academic data (Jha et al., 2023).

From an engineering perspective, architectural patterns in blockchain systems are of great importance. Pahl et al. (2018) describe architectural decisions regarding the allocation of responsibility between layers of applications, handling of transaction flow, and the construction of integration adapters (K et al., 2025). Architectural decisions are further linked to classes of threats, including consensus, storage, and access control, by Zhang et al. (2020), who highlight the selection of architecture and its corresponding security requirements. These results form the basis for designing server infrastructure, where security must be ensured not only by blockchain cryptography, but also by the correct organization of the API layer, access control, and data routing.

An important area is the integration of blockchain solutions with corporate and interorganizational information systems. Mendling et al. (2018) view blockchain as an infrastructure that requires alignment with existing processes and the IT landscape, including the need for intermediate layers and data synchronization mechanisms (K et al., 2025). In the context of a university, EduCTX offers a platform for the use of blockchain to record credits and verify accomplishments, demonstrating the practical use of distributed ledger systems for educational activities while highlighting the value of server-side design and integration mechanisms for implementation (Marcu et al., 2025).

Summarizing the results of the studies reviewed, two key conclusions can be identified that determine the formulation of the problem of this work. Firstly, most existing solutions and prototypes focus on the recording of final documents (diplomas, certificates) and demonstrating the advantages of immutability, while the architecture of the server circuit, providing the orchestration of academic data flows, validation and integration with existing educational systems, is not fully disclosed (Hindarto, 2023). Secondly, the literature consistently points to the feasibility of hybrid OnChain/OffChain models as the most realistic way of implementing blockchain in scalable educational platforms with limited distributed network resources (Iswahyudi et al., 2023). In this regard, the development of a backend architecture that ensures secure data processing, access control, validation and selective cryptographic recording of critical records in a consortium network is of scientific and practical interest and closes the identified architectural gap.

Materials and methods. Every operation carried out by a smart contract results in a change to the nodes of the World State tree. Ethereum State is essentially a

mapping of 160-bit address to the corresponding account state. This account state is stored in data structure called State Trie.

The account state is represented as a leaf node of which its address acts as a path from the root of trie down to that node.

Smart contracts are treated as special accounts. For each contract account, there is a distinct persistent storage data structure associated with it — Storage Trie.

The storage trie is also implemented as a Merkle Patricia Trie (MPT) (Marhoon et al., 2023).

Hence there are two levels to Ethereum World State trie hierarchy.

– Account Trie (Top Level).

– Storage Trie (Bottom Level).

Under Ethereum's current implementation, searching for/inserting a node along a path results in SHA3 () hashing followed by read / write access to levelDB. These operations take time, and the size of your new node will equate to how much data you are adding to the trie as a result of your state change. So if we can determine the current height of our new leaf node we can also determine our potential performance/storage gains. If we know how state changes are occurring in the MPT data structure we can predict transaction performance and additional storage by looking at the time it takes for world state changes to occur. We can also determine this empirically by testing smart contracts against certain business use cases. In this paper we will analyze current height of world state nodes in MPT data structure.

As described by Devroy (Iswahyudi et al., 2023), we can expect the height of an MPT to be (1):

$$H_n = \frac{2\sqrt{2log2n} + log2n}{4} \quad (1)$$

If the transactions number is n and executes another transaction, then according to Formula (2), we can calculate the average time cost for changing MPT is:

$$T_{mpt}(n) = \frac{t}{2}\log(an) \quad (2)$$

Where a represents the number of deployed smart contracts in the network, n represents the transaction volume and t represents execution time to create a node in MPT tree (including time of sha3() computation and database access time). Ethereum uses LevelDB as its key-value database. Keys used to access LevelDB are unevenly distributed due to the granularity of hashing. LevelDB has great performance for reads and writes that occur sequentially (Iswahyudi et al., 2023). However, this database suffers when confronted with random patterns of key access. As a result, t, the amount of time needed to access LevelDB, increases with

the growth of stored data. Experimental findings have demonstrated that when n is large enough, t will also rise and performance will degrade dramatically for those records that are not cached inside LevelDB.

Moving onto smart contract execution inside Ethereum Virtual Machine (EVM), after n transactions are processed we can determine average transaction execution time and the maximum transaction time using formula (3):

$$T_{avg}(n) = T_{exec} + T_{mpt}(n) \quad (3)$$

where T_{exec} – Smart contract execution time (On virtual machine)

After n transactions are executed, according to (4) formula, additional increase in storage size resulted from MPT structure change is:

$$S_{mpt}(n) = s_b\left(\frac{\log(an)}{2} - 1\right) + s_1 + s_a + s' \quad (4)$$

where s_b – size of committed transaction, s_1 – size of leaf node, s_a – space, space consumed by account state, s' – storage capacity used per account.

Including the transaction size here S_t, we can get the final formula as following (5):

$$S_{sumavg}(n) = nS_t + \sum_{i=0}^{n} S_{mpt}(n) \quad (5)$$

Authors have performed experimental validation of presented prediction method. Metrics they gathered during experiment:

– Height of State Tree in Ethereum World State;

– Transaction execution time;

– Memory usage after transaction execution.

Results. Experimental environment consisted only of test node. Single transaction was put into each block because caching systems of Ethereum could skew performance results. Empty smart contract with nearly no logic was used, 1 million transactions were sent in total.

The developed academic qualifications verification platform is implemented as a distributed ecosystem, including a web interface, a DHEP API backend service, and a blockchain infrastructure. The client-side component will operate like a web application and will interact with the server-side API, which will be responsible for the processing of the business logic, the handling of the data, and the orchestration of the write operations on the blockchain network. A schematic representation of the interaction between the components of the ecosystem is shown in Figure 1.

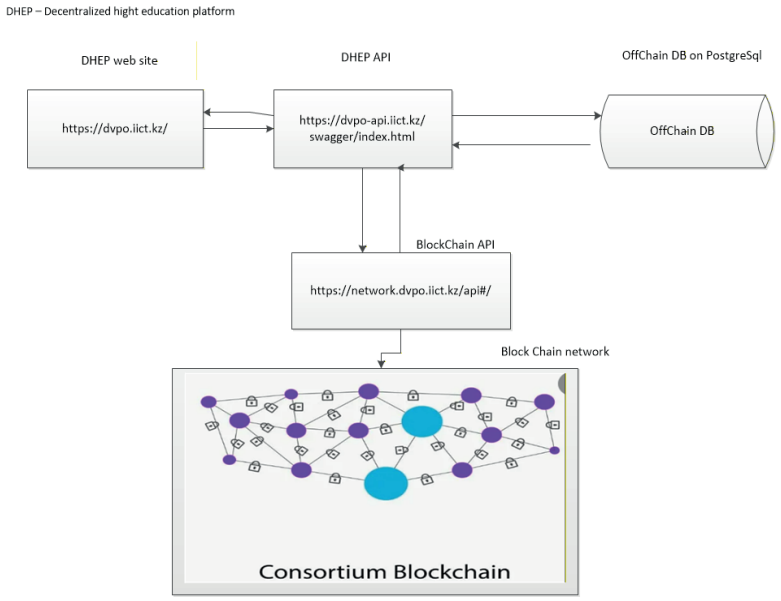


Figure 1 — General diagram of ecosystem interactions

The implementation of the backend layer is the DHEP API, and its architecture is shown in Figure 2. This layer acts as the main data processing node, which processes requests from the web interface, checks the data, manages entities in the centralized database, and interacts with the blockchain network API.

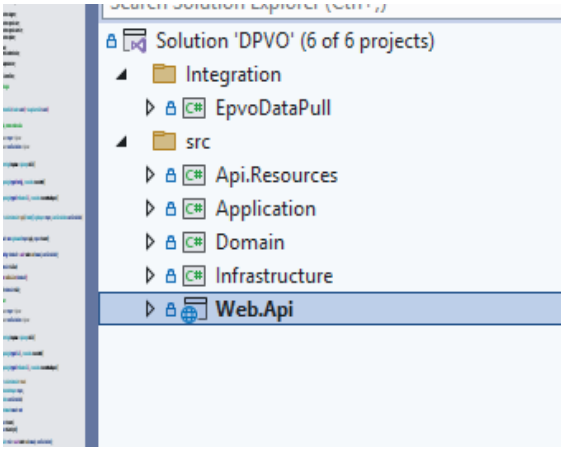


Figure 2 — DHEP API structure

The server-side architecture is based on Clean Architecture principles with the aim of separating business logic from infrastructure dependencies. This paradigm offers a layered system architecture with an application layer, a domain layer, and an infrastructure layer, where Web.Api represents the system's entry point.

The application layer is responsible for coordinating data processing scenarios, managing the life cycle of academic records, and enabling interaction between services. The domain layer includes the core system entities, which are represented by diploma, transcript, and students' data models, as well as business rules related to data processing. The infrastructure layer includes interaction with external components, represented by data sources, authentication, and blockchain nodes. The composition of Web.Api controllers and the distribution of their functional responsibilities are shown in Figure 3:

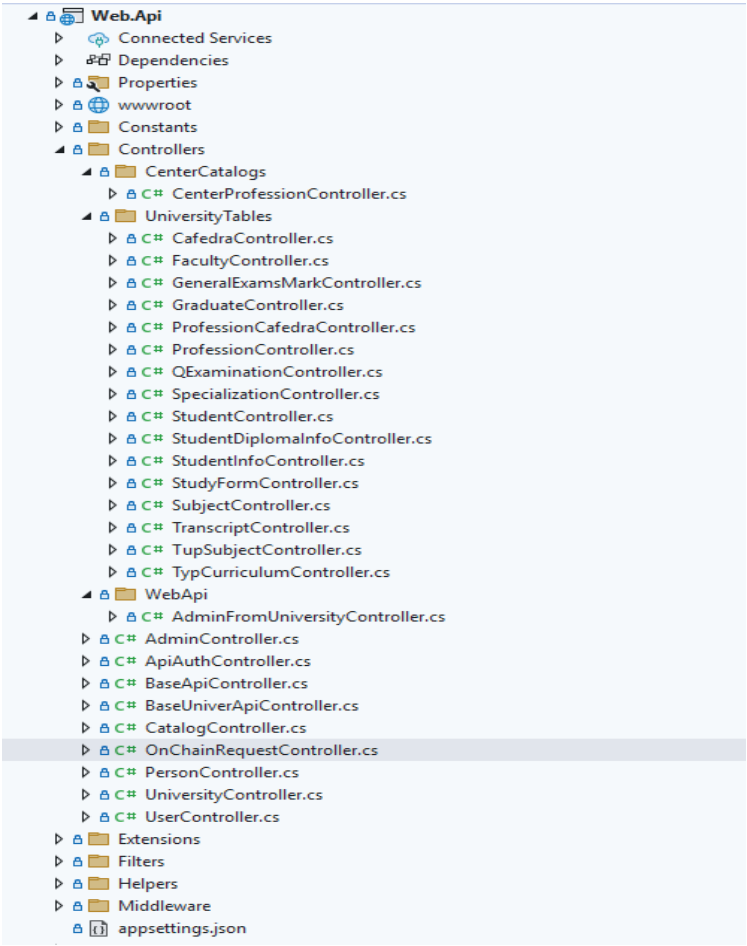


Figure 3 — Web.Api controllers

The system implements a role-based access control model, allowing users to be assigned multiple roles. Basic categories include organizational users, administrators, super administrators, and integration accounts. Access to API functionality is regulated both at the role level and through a permissions mechanism, providing flexible management of interfaces and system operations.

User authentication is implemented using JWT tokens signed with the HmacSha256 algorithm. Passwords are stored encrypted using the SHA-512 algorithm, ensuring the security of credentials.

Academic data storage is implemented using a hybrid model. The bulk of the information, including student personal data, educational program information, and administrative records, is stored in a centralized Off-Chain database built on PostgreSQL. The system utilizes central directories to ensure data consistency across various higher education institutions. Before storing records, they are validated against reference directories.

Data controllers for institutional and academic records implement a full set of CRUD operations, enabling the creation, updating, and deletion of records in the Off-Chain repository. This functional framework is shown in Figure 4:

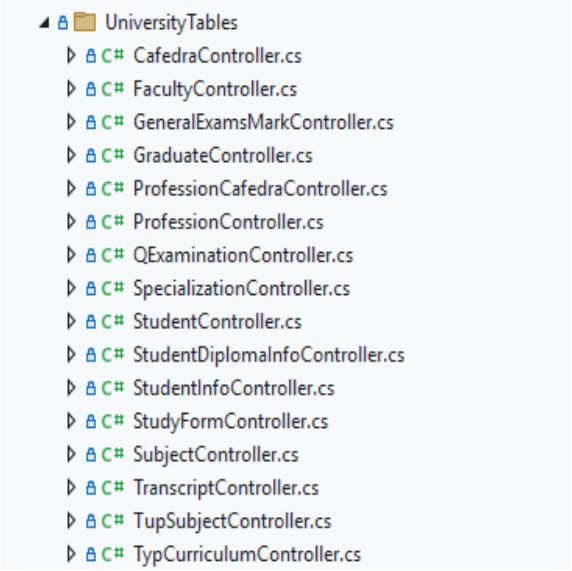


Figure 4 — CRUD controllers for managing organizational data

The process of data transfer to the blockchain network is facilitated using a specialized OnChainRequestController, the architectural role of which is presented in Figure 5:

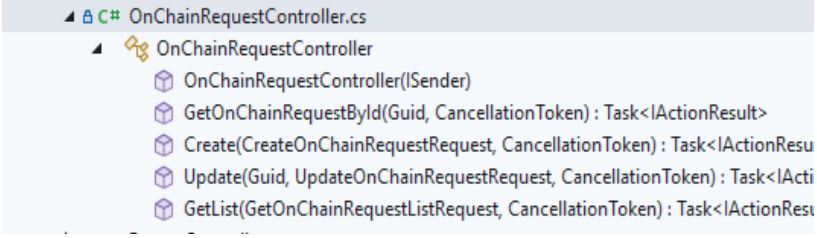


Figure 5 — OnChainRequestController

The Create method of this controller accepts the transcript model, validates the input data, and initiates a transaction on the blockchain network. In response, the system receives a transaction identifier (TransactionId), which is stored in the OnChainRequests table of the centralized database. A fragment of the implementation of this mechanism is shown in Figure 6.

```
2 references | - changes | -authors, -changes
public sealed record CreateOnChainRequestRequest(
    int UniversityId,
    long StudentId,
    long? TranscriptId,
    RequestType RequestType,
    string RequestData,
    string ResponceData,
    string TransactionId,
    DateTimeOffset CreatedAt
    );
```

Figure 6 — Excerpt from the OnChain request processing code

Considering block generation delays, a background transaction processing mechanism has been implemented. The process of transaction status interrogation is periodically carried out to check the status of previous data submissions based on TransactionId. Once the status is confirmed, the process of creating the transaction hash is initiated, ensuring the synchronization of the transcript records within the OffChain database.

Recording academic data on the blockchain network is implemented through two types of smart contracts. The initial contract has been developed for the storage of transcript grades, including course, ECTS credits, final grade, and details of the student and organization. The structural representation of this contract is presented in Figure 7.

```
{
  "transcriptId": 0,
  "memberId": 0,
  "studentId": 0,
  "totalMark": 0,
  "courseNumber": 0,
  "term": 0,
  "transcriptType": 0,
  "subjectId": 0,
  "ects": 0
}
```

Figure 7 — Transcript grade storage contract

The second contract has been designed for the storage of diploma details, including registration, number, series, issue date, and degree award date. The structure of this contract is presented in Figure 8.

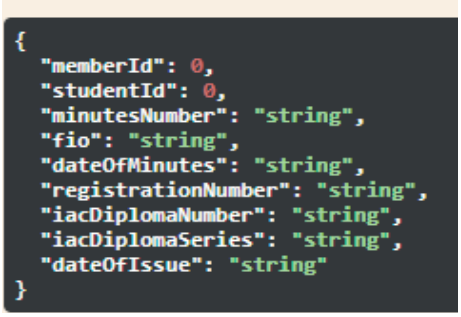


Figure 8 — Diploma data storage contract

The proposed architecture of the DHEP API backend component of the platform has ensured the entire academic data processing cycle, including centralized management, validation, and immutable storage within the distributed ledger. This has been ensured by the segregation of off-chain and on-chain components, the use of layers within the architectural structure, and the use of specialized transaction mechanisms.

Discussions. The server-side implementation of the platform has been realized within the context of the DHEP API backend application, which is responsible for handling client requests, academic data management, and integration with the blockchain infrastructure. This has been realized using the latest server-side implementation technologies, focusing on the development of distributed systems.

The ASP.NET Core framework has been employed as the backend platform, ensuring high-performance HTTP request handling, routing, and modular architecture. This has been done to realize the REST-based API, ensuring standardized interactions between clients and integration services.

The centralized academic data has been stored within the PostgreSQL relational database management system. This DBMS was selected based on requirements for transactional stability, scalability, and support for complex interrelated data structures. The database contains student records, transcripts, diplomas, reference books, and transaction tracking tables.

To work with the database, the Entity Framework Core object-relational mapping mechanism is used, which provides migration management, mapping of domain entities to the database structure, and optimization of data retrieval operations.

Integration with the blockchain network is achieved through an external application programming interface (API), which facilitates the submission of transactions and retrieval of transaction identifiers. Transaction processing is achieved through an asynchronous model, which involves the storage of transaction identifiers and subsequent verification through background checks.

Security of user interactions is achieved through a JSON Web Token (JWT)–based authentication model. Token signatures are generated through HmacSha256 hashing, and user credentials are stored through SHA-512 one-way hashing.

The backend infrastructure was deployed in a server environment with API publishing via secure network protocols. The deployment architecture ensures service availability, scalability, and transaction monitoring.

Thus, the chosen technology stack ensures stable processing of academic data, secure integration with the blockchain infrastructure, and the platform's readiness for expanding functionality.

Analysis of the effectiveness of platform architectural solutions

The developed backend platform was evaluated to analyze the impact of the adopted architectural decisions on data processing performance, data security, and system scalability. In comparison with the conventional centralized educational registries, the developed infrastructure uses the hybrid model of the management of academic information, which has several advantages.

An important factor affecting the performance of the system is the redistribution of computational tasks between the centralized and distributed circuits. The performance of primary transaction processes in the centralized infrastructure makes it possible to quickly register, update, and validate academic information, avoiding the limitations inherent in the blockchain environment.

Data storage in the distributed ledger occurs selectively, which helps to reduce the number of transactions and the time required to confirm them. This architectural model allows for the performance of centralized systems to be maintained while simultaneously implementing immutability mechanisms.

An additional optimization factor is asynchronous transaction processing, which eliminates blocking of user operations while waiting for records to be confirmed in the blockchain network. This ensures the stability of user scenarios and reduces operational latency when working with academic data.

Ensuring the integrity and security of academic records

Integrating cryptographic data capture into the platform architecture significantly increases the level of trust in educational qualifications. The use of hash identifiers ensures the detection of any attempts to modify the original records, as data changes lead to a mismatch between the control values.

The use of smart contract logic creates an immutable storage environment for critical parameters of diplomas and transcripts. Unlike centralized databases, where changes can be made by system administrators, a distributed ledger eliminates the possibility of retrospective editing of records.

An additional layer of security is provided by authentication and access control mechanisms. A multi-level authorization model reduces the likelihood of unauthorized access to registration operations and data uploads to the distributed ledger.

Scalability and resilience of infrastructure

As the participant base expands in the educational ecosystem, the architecture of

the platform is such that it supports scalability. This is achieved by the centralized server architecture, which supports horizontal scaling of computing resources to improve system throughput in line with increasing transactions due to a growing participant base.

The distributed circuit operates on a consortium model, allowing for the addition of new validation nodes. The expansion of the participant network improves the robustness of the infrastructure and reduces the threat of data loss due to unavailability of individual participants.

The separation of storage circuits also improves system robustness. This is because, in spite of a temporary shutdown of the distributed ledger system, the centralized system is always up and running, ensuring continuous recording of academic data.

Comparative analysis with centralized educational registries

In traditional data storage solutions used in the educational domain, infrastructure is largely centralized around databases and provides high processing speeds but does not ensure data immutability. Any changes can only be made with administrative privileges, which reduces the level of trust in digital qualifications.

The proposed hybrid model mitigates this drawback by recording vital parameters in a distributed ledger. This ensures transparency of record origin, independent verification, and the creation of a trusted digital environment for qualification recognition.

At the same time, having a centralized operational framework mitigates performance and storage cost constraints associated with distributed systems.

Generalized Assessment of Architectural Effectiveness

The analysis above shows that the proposed architecture for the backend of the system is effective and provides a well-balanced infrastructure for academic data management in the context of the educational domain.

The implemented solutions provide:

- Maintaining high processing speed;
- Cryptographic security of qualifications;
- Resistance to unauthorized changes;
- Scalability as the number of participants increases;
- Possibility of independent verification of educational documents.

Thus, the integration of a distributed ledger into the server infrastructure of an educational platform increases the level of trust in digital qualifications without reducing the operational efficiency of the system.

Figure 9 provides an overview of architecture for the proposed blockchain transaction processor application highlighting major components and their workflow.

The proposed architecture focuses on building highly performing blockchain-based applications that are secure and deterministic when executed on a permissioned blockchain network.

Microservice-based architecture with asynchronous communication patterns allows the solution to be efficient, scale gracefully, and support fault isolation.

To provide blockchain transaction processing services, NestJS-based OnChain API sits at the application level. The API receives the incoming requests to be executed on blockchain and is responsible for connecting to a PostgreSQL database as its persistent storage layer. The application also features a RabbitMQ-based task queue to decouple API from transaction execution logic, allowing for asynchronous and non-blocking operations.

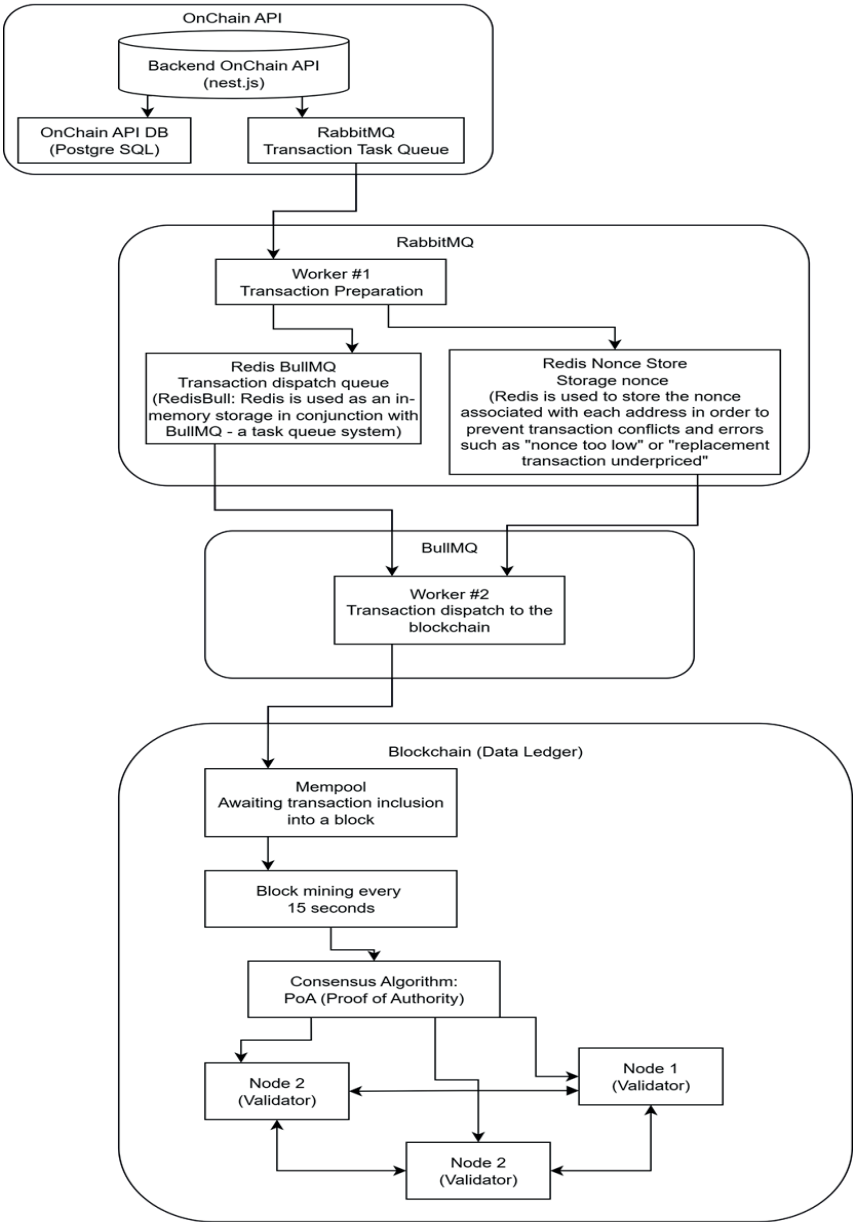


Figure 9 – Detailed architectural diagram of the blockchain-based transaction processing system

In contrast to Ethereum's public blockchain, where smart contracts can be deployed at random times in unpredictable complexity by anybody, consortium networks allow for a certain degree of estimation regarding performance and storage consumption:

- because network nodes are permissioned and access can be regulated;
- smart contracts follow a predefined structure and complexity can be controlled according to the business logic;
- transaction volume is capped by the scope of the enterprise use case at hand.

Therefore, architecture merges the deterministic behavior of a consortium blockchain with the scalability benefits of a microservice architecture.

Conclusion. This study developed and implemented a backend platform for verifying academic qualifications based on a hybrid architecture for storing and processing educational data. The proposed solution integrates a centralized server infrastructure for managing academic records with a distributed blockchain registry, ensuring the immutable recording of critical parameters of diplomas and transcripts.

The proposed architecture provides operational and cryptographic data processing paths and provides high performance for academic data while at the same time ensuring its security. An asynchronous model of transactions is used to prevent user operation blocking and ensure stability in the presence of confirmation delays in the blockchain network. Smart contract infrastructure is used to create a framework for immutable data recording of academic qualifications and enables verification of diplomas and transcripts. Cryptographic identification mechanisms ensure data immunity to retrospective modification and increase the reliability of digital academic documents.

Practical implementation of the proposed platform shows its potential for integrating with existing infrastructure in the educational domain without compromising operational efficiency of centralized systems. The hybrid data storage model ensures a balance between processing performance and record immutability, which is particularly important for scaling national educational platforms. The novelty of this study is achieved in terms of designing a backend architecture that handles academic data workflow coordination, data validation, data dissemination in Off-Chain and On-Chain circuits and subsequently performs cryptographic verification. Unlike existing studies that mostly deal with document capture, the proposed system is focused on executing the entire academic record processing lifecycle at the server infrastructure level.

The contribution of this study in terms of practical applications is achieved in terms of potential use cases in national and transnational educational qualification system development, and trusted digital registries of diplomas. Promising areas for further research include expanding the functionality of smart contracts, developing mechanisms for intersystem interoperability, and implementing analytical tools for monitoring academic data in a distributed environment.

References

Alammary A., Alhazmi S.; Almasri M.; Gillani S. Blockchain-Based Applications in Education: A Systematic Review. Applied Sciences 2019, 9, 2400. *https://doi.org/10.3390/app9122400* (In Eng.)

Alzahrani A.I.; Alghamdi A.; Alotaibi B.; Alshahrani S.; Alqahtani A. Blockchain-Based Academic Credential Interoperability Protocol (BACIP). Applied Sciences 2025, 15, 9191. *https://doi.org/10.3390/app15091991* (In Eng.)

Zheng Z.; Xie S.; Dai H.; Chen X.; Wang H. Blockchain Challenges and Opportunities: A Survey. International Journal of Web and Grid Services 2018, 14. — P. 352–375. *https://doi.org/10.1504/IJWGS.2018.095647* (In Eng.)

Pahl C.; El Ioini N.; Helmer S. Architectural Patterns for Blockchain Systems. In Proceedings of the IEEE International Conference on Software Architecture (ICSA); Seattle, WA, USA, 2018. — P. 221–230. *https://doi.org/10.1109/ICSA.2018.00031* (In Eng.)

Zhang Y.; Kasahara S.; Shen Y.; Jiang X.; Wan J. A Study of Blockchain Architecture Design Decisions and Their Security Attacks and Threats. IEEE Access 2020, 8. — P. 148442–148456. *https://doi.org/10.1109/ACCESS.2020.3016043* (In Eng.)

Mendling J.; Weber I.; van der Aalst W.; vom Brocke, J.; Cabanillas C.; Daniel F.; Debois S.; Di Ciccio C.; Dumas M.; Dustdar, S.; et al. Blockchains for Business Process Management—Challenges and Opportunities. ACM Transactions on Management Information Systems 2018, 9, 4. *https://doi.org/10.1145/3183367* (In Eng.)

Turkanović M.; Hölbl M.; Košič K.; Heričko M.; Kamišalić A. EduCTX: A Blockchain-Based Higher Education Credit Platform. IEEE Access 2018, 6. — P. 5112–5127. *https://doi.org/10.1109/ACCESS.2018.2789929* (In Eng.)

Ocheja P., Agbo F.J., Oyelere S., Flanagan B. and Ogata H. (2022). Blockchain in Education: A Systematic Review and Practical Case Studies. IEEE Access, 10. (In Eng.)

Moya J.A.B. (2024) Blockchain for Academic Integrity: Developing the Blockchain Academic Credential Interoperability Protocol (BACIP). arXiv.org. (In Eng.)

Rustemi A., Dalipi F., Atanasovski V. and Risteski A. (2023) A Systematic Literature Review on Blockchain-Based Systems for Academic Certificate Verification. IEEE Access, 11. (In Eng.)

Nadeem N., Hayat M.F., Qureshi M.A., Majid M., Nadeem M. and Janjua J. (2023) Hybrid Blockchain-based Academic Credential Verification System (B-ACVS). Multimedia Tools and Applications. (In Eng.)

Marhoon, H. M., Basil, N. and Ma'arif, A. (2023) Exploring Blockchain Data Analysis and Its Communications Architecture: Achievements, Challenges, and Future Directions: A Review Article. International Journal of Robotics and Control Systems. (In Eng.)

Iswahyudi I., Hindarto D. and Indrajit R. E. (2023) Digital Transformation in University: Enterprise Architecture and Blockchain Technology. SinkrOn. (In Eng.)

Hindarto D. (2023) Blockchain-Based Academic Identity and Transcript Management in University Enterprise Architecture. SinkrOn. (In Eng.)

Dey K. and Shekhawat U. (2021) Blockchain for sustainable e-agriculture: Literature review, architecture for data management, and implications. Journal of Cleaner Production, 316. (In Eng.)

Alzhrani F.E., Saeedi K. and Zhao L. (2023) Architectural Patterns for Blockchain Systems and Application Design. Applied Sciences. (In Eng.)

Ocheja P., Agbo F.J., Oyelere S., Flanagan B. and Ogata H. (2022) Blockchain in Education: A Systematic Review and Practical Case Studies. IEEE Access, 10. (In Eng.)

Nandanwar M., Kireedula N. and Sharma A. (2025) A Secure Blockchain based Framework for Evaluation of Examination Answer Sheets. International Journal For Multidisciplinary Research. (In Eng.)

Marcu A., Antăluţe F. and Alexandrescu A. (2025) Decentralized Educational Certification Platform using Blockchain, Dynamic NFTs, and AI-Based Assessment. International Conference on System Theory, Control and Computing. (In Eng.)

Moya J.A.B., Ayoade J. and Uddin M.A. (2025) A Zero-Knowledge Proof-Enabled Blockchain-Based Academic Record Verification System. Sensors, 25. (In Eng.)

K, M., T, V. Y. S., Shankara, C.P., T, L. H., V., M.K.N. and Darshan B.T. (2025) Blockchain-based Certificate Verification System with AI-Assisted Content Validation using Hyperledger Fabric. 2025 International Conference on Sustainable Communication Networks and Application (ICSCN). (In Eng.)

Tawfeeq N.Z. and Taha D.B. (2025) Blockchain-Based Accountability: A Systematic Literature Review. Mesopotamian Journal of CyberSecurity. (In Eng.)

Razaque A.; Amanzholova S.; Mutanov G.; Ussatova O.; Farhad M. A.; Kistaubayev Y.; Smagulova K.; Kopebayev A. An Anti-Corruption System for Academic Achievement Verification in Kazakhstani Higher Education Using Blockchain and Artificial Intelligence. Frontiers in Blockchain, 2025, 8:1683522. *https://doi.org/10.3389/fbloc.2025.1683522* (In Eng.)

Amanzholova S.; Mutanov G.; Ussatova O.; Aldasheva L.; Amirova A.; Naumenko V.A Conceptual Framework for AI- and Blockchain-Enabled Research Project Evaluation Systems. Information, 2026, 17(2), 151. *https://doi.org/10.3390/info17020151* (In Eng.)

Jha K.; Pasbola M.; Rai H. M.; Amanzholova S. Utilizing Smartwatches and Deep Learning Models for Enhanced Avalanche Victim Identification, Localization, and Efficient Recovery Strategies. Proceedings of the ACM I3CS 2023. *https://doi.org/10.1145/3647444.3652483*. (In Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 73–90

https://doi.org/10.32014/2026.2518-1726.402

IRSTI 28.23.29
UDC 004.89

© Amirkhanova G.A., Nurgazy T.N.*, Amirkhanov B.S., Nurgazy N.N., 2025.
Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: nurgazytomiris@gmail.com

DEVELOPING A PREDICTIVE DIGITAL TWIN FOR A FOOD PRODUCT BASED ON EDGE ML AND IOT SENSORS

Amirkhanova Gulshat — PhD, associate professor, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: gulshat.aa@gmail.com, https://orcid.org/0000-0003-3933-5476;
Nurgazy Tomiris — Master's student, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: nurgazytomiris@gmail.com, https://orcid.org/0009-0001-6650-5517;
Amirkhanov Bauyrzhan — Doctoral student, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: amirkhanov.b@gmail.com, https://orcid.org/0000-0002-4915-0347;
Nurgazy Nurbol — Doctoral student, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: nurgazynurbol1@gmail.com, https://orcid.org/0009-0007-2858-1176.

Abstract. This paper examines the research on developing a predictive digital twin for food products, aiming to reduce food losses and establish the basis for intelligent quality management. The extent to which the proposed study is relevant is largely due to the necessity to address the world problem of food loss and the need to introduce the technologies of Industry 4.0 into the agri-food industry. The main idea is to work out a solution in the context of which a product status in real-time will be monitored with the help of edge computing (Edge ML) by means of using the data provided by IoT sensors and forecasting its quality, e.g., the remaining shelf life. The approaches include smart sensors (temperature, humidity, gas level), ESP32 microcontroller as transmitters of collected data, MQTT protocol to transfer data, and Raspberry Pi 5 firmware used to process data in locations. Predictive analysis by use of XGBoost (Extreme Gradient Boosting). The experiments carried out on the Golden Delicious apples showed that the Gradient Boosting model produced very high prediction accuracy because the R 2 Score was 0.9936 in the training set. These results validate the great potential of Edge ML technology as a means of creating reliable, energy-efficient, and dynamic digital twins based on real-time data. The practical value of the given work is predicated on the fact that

the proposed prototype can serve as the foundation for the creation of commercial monitoring systems. The advantage of such systems is that they would also help producers and retailers manage inventory and logistics proactively, thereby reducing losses due to spoilage and making the food industry more sustainable.

Keywords: digital twin, Edge ML, smart sensors, shelf life prediction, food safety

For citations: Amirkhanova G.A., Nurgazy T.N., Amirkhanov B.S., Tokhtassyn M.M., Nurgazy N.N. Developing a predictive digital twin for a food product based on Edge ML and IoT sensors. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 73–90. DOI: https://doi.org/10.32014/2026.2518-1726.402

Financing. *Funding: This research was funded by the Ministry of Education and Science of the Republic of Kazakhstan grant number BR24992975. "Development of a Digital Twin for the Food Industry Enterprise Using Artificial Intelligence and IIoT Technologies.".*

© Амирханова Г.А., Нұрғазы Т.Н.*, Амирханов Б.С., Нұрғазы Н. Н., 2025.
Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.
E-mail: nurgazytomiris@gmail.com

EDGE ML ЖӘНЕ IOT СЕНСОРЛАРЫ НЕГІЗІНДЕ АЗЫҚ-ТҮЛІК ӨНІМІНІҢ ПРЕДИКТИВТІ ЦИФРЛЫҚ ЕГІЗІН ӘЗІРЛЕУ

Амирханова Гульшат — PhD, доцент, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: gulshat.aa@gmail.com, https://orcid.org/0000-0003-3933-5476;
Нұрғазы Томирис — магистрант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: nurgazytomiris@gmail.com, https://orcid.org/0009-0001-6650-5517;
Амирханов Бауржан — докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: amirkhanov.b@gmail.com, https://orcid.org/0000-0002-4915-0347;
Нұрғазы Нұрбол — докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: nurgazynurbol1@gmail.com, https://orcid.org/0009-0007-2858-1176.

Аннотация. Берілген мақала азық-түлік өнімдерінің ысыраптарын азайтуға және сапаны интеллектуалды басқарудың негізін құруға бағытталған азық-түлік өнімдерінің болжамды цифрлық егізін әзірлеу бойынша зерттеулерді қарастырады. Ұсынылып отырған зерттеудің өзектілігі көбінесе азық-түлік бүлінуі әлемдік мәселесін шешу қажеттілігімен және Индустрия 4.0 технологияларын агроазық-түлік өнеркәсібіне енгізу қажеттілігімен байланысты. Негізгі идея - IoT сенсорлары беретін деректерді пайдалану

және оның сапасын болжау арқылы, мысалы, қалған сақтау мерзімін болжау арқылы шеттік есептеулер (Edge ML) көмегімен нақты уақыт режимінде өнім күйі бақыланатын контекстте шешім әзірлеу. Әдістерге смарт сенсорлар (температура, ылғалдылық, газ деңгейі), жиналған деректерді таратқыш ретінде ESP32 микроконтроллері, деректерді тасымалдауға арналған MQTT протоколы және орындарда деректерді өңдеу үшін пайдаланылатын Raspberry Pi 5 микробағдарламасы кіреді. XGBoost (Extreme Gradient Boosting) көмегімен болжамды талдау. ‹Голден Делишес› алмаларында жүргізілген эксперименттер Gradient Boosting моделі болжамның өте жоғары дәлдігін беретінін көрсетті, себебі оқу жиынында R2 нәтижесі 0,9936 болды. Бұл нәтижелер нақты уақыт деректеріне негізделген сенімді, энергияны үнемдейтін және динамикалық цифрлық егіздерді жасау құралы ретінде Edge ML технологиясының үлкен әлеуетін растайды. Берілген жұмыстың практикалық құндылығы ұсынылған прототиптің коммерциялық мониторинг жүйелерін құру үшін негіз бола алатындығына негізделген. Мұндай жүйелердің артықшылығы, олар өндірушілер мен жеке саудагерлерге тауарлық-материалдық қорлар мен логистиканы белсенді басқаруға көмектеседі, осылайша бүлінуден болатын шығындарды азайтады және тамақ өнеркәсібін тұрақты етеді.

Түйін сөздер: цифрлық егіздер, Edge ML, ақылды сенсорлар, сақтау мерзімін болжау, азық-түлік қауіпсіздігі

© Амирханова Г.А., Нургазы Т.Н.*, Амирханов Б.С., Нургазы Н.Н., 2025.
Казахский национальный университет имени аль-Фараби,
Алматы, Казахстан.
E-mail: nurgazytomiris@gmail.com

РАЗРАБОТКА ПРЕДИКТИВНОГО ЦИФРОВОГО ДВОЙНИКА ПИЩЕВОГО ПРОДУКТА НА ОСНОВЕ EDGE ML И IOT-СЕНСОРОВ

Амирханова Гульшат — PhD, доцент, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан,
E-mail: gulshat.aa@gmail.com, https://orcid.org/0000-0003-3933-5476;
Нургазы Томирис — магистрант, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан,
E-mail: nurgazytomiris@gmail.com, https://orcid.org/0009-0001-6650-5517;
Амирханов Бауржан — докторант, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан,
E-mail: amirkhanov.b@gmail.com, https://orcid.org/0000-0002-4915-0347;
Нургазы Нурбол — докторант, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан,
E-mail: nurgazynurbol1@gmail.com, https://orcid.org/0009-0007-2858-1176.

Аннотация. В статье рассматривается разработка предиктивного цифрового двойника пищевого продукта, направленного на сокращение потерь

и формирование основы для интеллектуального управления качеством. Актуальность исследования обусловлена необходимостью решения глобальной проблемы продовольственных потерь, а также внедрением технологий Индустрии 4.0 в агропродовольственный сектор. Ключевая идея заключается в создании системы, обеспечивающей мониторинг состояния продукта в режиме реального времени с использованием периферийных вычислений (Edge ML) на основе данных, получаемых от сенсоров Интернета вещей, а также прогнозирование его качества, в частности остаточного срока годности. Предложенный подход включает использование интеллектуальных сенсоров (температуры, влажности, газовой среды), микроконтроллера ESP32 для передачи данных, протокола MQTT для их обмена и вычислительного узла на базе Raspberry Pi 5 для локальной обработки информации. Для предиктивного анализа применена модель XGBoost (Extreme Gradient Boosting). Экспериментальные исследования, проведённые на яблоках сорта Golden Delicious, показали высокую точность прогнозирования: коэффициент детерминации R^2 в обучающей выборке составил 0,9936. Полученные результаты подтверждают значительный потенциал технологии Edge ML для создания надёжных, энергоэффективных и адаптивных цифровых двойников, функционирующих в режиме реального времени. Практическая значимость исследования заключается в возможности использования разработанного прототипа в качестве основы для коммерческих систем мониторинга. Такие системы позволяют производителям и розничным сетям проактивно управлять запасами и логистикой, снижая потери, связанные с порчей продукции, и повышая устойчивость пищевой отрасли.

Ключевые слова: цифровой двойник, Edge ML, умные сенсоры, прогнозирование срока хранения, пищевая безопасность

Кіріспе. Соңғы жылдары Индустрия 4.0 тұжырымдамасының негізгі элементі ретінде цифрлық егіз (ЦЕ) технологиясы зор танымалдылыққа ие болуда. ЦЕ – бұл физикалық нысанның немесе процестің нақты уақыттағы деректер арқылы үнемі жаңартылып отыратын динамикалық виртуалды көшірмесі. Азық-түлік өнеркәсібінде бұл технология өнім сапасын бақылау, ысырапты азайту және процестерді оңтайландыру үшін бірегей мүмкіндіктер ашады. Алайда, тез бұзылатын өнімдер үшін тиімді ЦЕ құрудың негізгі қиындығы – оның жай-күйін статикалық түрде көрсетіп қана қоймай, болашақтағы өзгерістерін, әсіресе жарамдылық мерзімін дәл болжай алуында жатыр. Мұндай предиктивті қабілетсіз цифрлық егіз толыққанды болмайды. Дәстүрлі бұлттық жүйелерде деректерді талдау кідірістерге әкеледі, бұл нақты уақыттағы синхрондауды қиындатады. Бұл мәселені шешу үшін деректерді пайда болған жерінде өңдейтін шеттегі есептеулер (Edge ML) технологиясы қажет.

Осы мақаланың негізгі мақсаты – азық-түлік өніміне арналған цифрлық егіздің предиктивті ядросын құрайтын Edge ML және IoT сенсорларына

негізделген прототипті әзірлеу. Бұл прототип сенсорлардан деректерді жинап, оларды жергілікті деңгейде талдайды және өнімнің қалған сақтау мерзімін жоғары дәлдікпен болжайды. Зерттеу нысаны ретінде алма жемісі алынып, оның сақтау мерзімін болжаудағы жүйенің тиімділігі эксперименталды түрде тексерілді.

Бұл жұмыс Edge ML технологиясының нақты уақыттағы деректерге негізделген сенімді цифрлық егіздерді құрудағы маңызды рөлін көрсетеді. Ұсынылған тәсіл азық-түлік ысырабын азайтуға және өнім сапасын басқаруды жақсартуға мүмкіндік беретін практикалық құрал болып табылады. Мақалада әдебиеттерге шолу, әзірленген прототиптің әдістемесі, алынған нәтижелер және олардың талқылануы ұсынылған.

Әдеби шолу. *Азық-түлік өнеркәсібіндегі цифрлық егіздер тұжырымдамасы.* Цифрлық егіз (ЦЕ) – бұл нақты уақыттағы деректер арқылы өзінің физикалық аналогымен үнемі синхрондалып отыратын, оның жай-күйін, жұмысын және өмірлік циклін сипаттайтын динамикалық виртуалды модель (Abdurrahman and Ferrari, 2025). Соңғы жылдары азық-түлік қауіпсіздігі мен тарату желілерінің тиімділігін арттыру ғаламдық мәселеге айналды. Bitkom цифрлық қауымдастығы мен Германияның тамақ және сусын өнеркәсібі федерациясының (BVE) сауалнамасы (Rohleder and Minhoff, 2019) 300-ден астам сұралған тамақ өнеркәсібі кәсіпорындарының 70%-ы көзден тұтынушыға дейін түпкілікті бақылауды маңызды басымдық деп санайтынын көрсетеді. Жасанды интеллект (AI) негізіндегі болжамдау модельдері азық-түліктің бұзылуын және сақтау мерзімін бағалауда дәстүрлі тәсілдерге қарағанда әлдеқайда тиімді шешімдер ұсынады (Shehzad, 2025). Заттар интернеті (IoT), AI және бұлттық есептеулер сияқты технологияларды біріктіретін ақылды жүйелер (Gupta et al., 2025) ЦЕ-дің толыққанды жұмыс істеуіне негіз болады.

Цифрлық егіздің сенсорлық-коммуникациялық деңгейі. Ақылды сенсорлар физикалық нысанның жай-күйі туралы деректерді жинап, оны цифрлық модельге жеткізетін негізгі құрал болып табылады (Gupta et al., 2025). Сенсорлар мен машиналық оқытуды біріктіретін «интеллектуалды сенсорлар» тұжырымдамасы өндіріс процестерін оңтайландыру үшін бірыңғай негіз ұсынады (Watson et al., 2021). Мұндай жүйелер температура, ылғалдылық, газ концентрациясы сияқты маңызды параметрлерді нақты уақыт режимінде өлшей алады (Baswoju et al., 2023; da Costa et al., 2023; Manimaran et al., 2025). IoT технологиясымен біріктірілген биосенсорларды азық-түлік қаптамаларына енгізу – тамақ қауіпсіздігі мен қадағалануын жақсартудағы маңызды серпіліс болып саналады (Sobhan et al., 2025). Бұл сенсорлар температура, ылғалдылық, қышқылдық (pH) және ұшпа органикалық қосылыстардың концентрациясы сияқты маңызды параметрлерді нақты уақыт режимінде өлшей алады. Мұндай жүйелер деректерді IoT желілері арқылы жіберіп, жеткізу тізбегінің барлық кезеңдерінде проактивті шешімдер қабылдауға мүмкіндік береді (Sobhan et al., 2025). Көптеген зерттеушілер дрон сияқты IoT құрылғыларын ауыл

шаруашылығында (Kim et al., 2019) және топырақ жағдайын бақылау үшін киілетін сенсорларды (Yin et al., 2021) қолданудың тиімділігін көрсетті. Бұл тәсілдер тұрақты дамуға ықпал етеді. Осы сенсорлардан жиналған деректерді ЦЕ-дің орталық жүйесіне тиімді жеткізу үшін сенімді коммуникациялық протокол қажет. Біздің зерттеу тобымыздың жүргізген салыстырмалы талдауы MQTT over TCP протоколының ең төменгі кідірісті қамтамасыз ететінін көрсетті, бұл нақты уақыттағы ЦЕ үшін аса маңызды (Amirkhanov et al., 2025b).

Цифрлық егіздің аналитикалық және предиктивті ядросы. Машиналық оқыту алгоритмдері цифрлық егіздің аналитикалық ядросын құрайды. Олар сенсорлық деректер негізінде өнімнің сапасын бағалап, жарамдылық мерзімін болжауға мүмкіндік береді. Бұл салада әртүрлі модельдер, соның ішінде конволюциялық нейрондық желілер (CNN) (Baswoju et al., 2023), Random Forest және нейрондық желілер сияқты ансамбльдік әдістер (Manimaran et al., 2025) кеңінен қолданылады. Жалпы алғанда, AI негізіндегі болжамдау модельдері IoT сенсорларынан, спектроскопиядан және қоршаған ортаны бақылаудан алынған көптеген дереккөздерге сүйенеді (Shehzad, 2025). Бұл тәсілдер сақтау шарттарын бақылайтын сенсорлардан алынған деректердегі үлгілер мен аномалияларды анықтауға, сол арқылы азық-түліктің оңтайлы жағдайда сақталуын қамтамасыз етуге көмектеседі (Grazia, 2024). Сонымен қатар, жасанды интеллект (AI) технологияларын қолдану тамақ өнімдерінің қауіпсіздігін, сапасын және қорғалуын жақсартуға мүмкіндік беретін IoT-биосенсорлардың мүмкіндіктерін кеңейтеді (Sobhan et al., 2025). Ақылды орау жүйелері өнімнің бастапқы құрамы, микробтардың өсуі және сақтау шарттары туралы ақпаратты жинақтап, жеткізу тізбегінің барлық қатысушыларына маңызды деректер ұсынады (Rodrigues et al., 2021). Жақында жүргізілген зерттеулер аммиак буының детекциясы арқылы еттің балғындығын нақты уақытта бақылауға болатынын көрсетті (Ranjith et al., 2023). Алайда, машиналық оқыту модельдеріне жіберілетін деректердің қауіпсіздігі мен тұтастығын қамтамасыз ету – цифрлық егіз экожүйесінің маңызды аспектісі. Осы мақсатта, біздің зерттеу тобымыз қашықтан бақылау және қауіпсіздік қатерлерін анықтау үшін деректерді жинау мен сақтаудың кешенді жүйесін құрды (Adilzhanova et al., 2025 a). Сонымен қатар, жиналған деректер негізінде кибершабуылдардың түрлерін ажырату үшін машиналық оқыту алгоритмдері қолданылды (Adilzhanova et al., 2025 b).

Цифрлық егіздің предиктивті функцияларын тиімді іске асыру бірнеше негізгі қағидатты ескеруді талап етеді (Mamarazhabova and Shingisov, 2024; Krupitzer and Noack, 2022). Олардың біріншісі – сенсорлық деректер негізінде өнімнің бұзылу үлгілерін анықтайтын және оның болашақтағы жай-күйін болжайтын машиналық оқыту алгоритмдерін қолдану болып табылады. Осы болжамның дәлдігін арттыру үшін әр өнімнің ерекшелігіне сай келетін оңтайлы модельді таңдау маңызды.

Сонымен қатар, цифрлық егіздің үздіксіз жұмысын қамтамасыз етуде

энергия тиімділігі шешуші рөл атқарады. Дәл осы жерде шеткі есептеулер (Edge ML) технологиясының артықшылығы көрінеді: ол сенсорлардың деректерді жинау және жіберу жиілігін оңтайландыру арқылы энергия тұтынуды азайтуға мүмкіндік береді, бұл жүйенің тұрақты жұмыс істеуіне ықпал етеді.

Осы қағидаттарды біріктіру арқылы сенімді, дәл және энергия тиімді цифрлық егізді құруға болады, бұл өз кезегінде азық-түлік ысырабын азайтып, өнім сапасын басқаруды жақсартады.

Цифрлық егіздің визуализация деңгейі. Талдау нәтижелерін және болжамдарды пайдаланушыға түсінікті түрде ұсыну үшін жоғары сапалы 3D-визуализация қажет. Бұл цифрлық егізбен интерактивті жұмыс істеуге және өндірістік процестерді модельдеуге мүмкіндік береді. Осы мақсатта біздің зерттеу тобымыз Blender бағдарламасын (Amirkhanov et al., 2025a) пайдалана отырып, өндірістік жабдықтың нақты 3D-модельдерін жасау әдістемесін әзірледі. Бұл тәсілдер цифрлық егіздің шынайы және дәл виртуалды көшірмесін жасауға негіз болады.

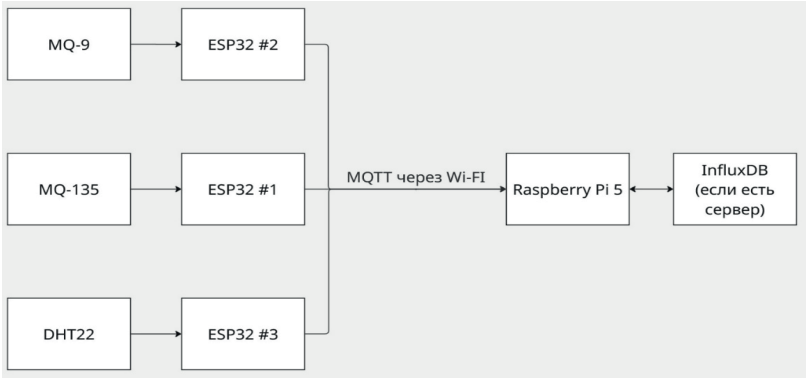
Зерттеудегі ғылыми олқылықты анықтау. Көптеген зерттеулерде модельдер бұлттық ортада жұмыс істейді, бұл кідірістерге және жоғары энергия тұтынуына әкелуі мүмкін. Сонымен қатар, AI модельдерін енгізуде деректердің аздығы, модель дәлдігінің шектеулі болуы және енгізу құнының жоғары болуы сияқты қиындықтар бар (Shehzad, 2025). Жан-жақты шолулар көрсеткендей, IoT негізіндегі биосенсорларды енгізуде деректердің құпиялылығы, қауіпсіздігі, интеграцияның күрделілігі және энергиямен қамтамасыз ету сияқты бірқатар шешілмеген мәселелер бар (Manimaran et al., 2025). Шеттегі есептеулер деректерді бұлтқа жіберудегі кідірістерді жоюға және энергия тиімділігін арттыруға мүмкіндік береді, бұл жеткізу тізбегіндегі автономды құрылғылар үшін аса маңызды. Осыған байланысты, біздің зерттеуіміз осы олқылықты толтыруға, яғни сенсорлардан алынған деректерді нақты уақытта өңдейтін, жарамдылық мерзімін болжайтын және энергияны үнемдейтін Edge ML негізіндегі прототипті әзірлеп, оның тиімділігін эксперименталды түрде дәлелдеуге бағытталған. Жоғарыда аталған жұмыстарда біздің зерттеу тобымыз цифрлық егіздің коммуникациялық (Amirkhanov et al., 2025 b), қауіпсіздік (Adilzhanova et al., 2025a, Adilzhanova et al., 2025b) және визуализация (Amirkhanov et al., 2025b) сияқты маңызды компоненттерін қарастырғанымен, оның аналитикалық ядросының, яғни Edge ML негізіндегі предиктивті модулінің практикалық іске асырылуы және эксперименталды түрде дәлелденуі басты назардан тыс қалды. Осыған байланысты, біздің зерттеуіміз осы олқылықты толтыруға бағытталған.

Материалдар мен әдістер**.** *Зерттеу нысаны және жалпы тәсіл.* Бұл мақала алма жемісінің цифрлық егізінің предиктивті ядросын құруға арналған прототиптік жүйенің әзірленуін қарастырады. Негізгі нәтижелерге смарт сенсорлардан (температура, ылғалдылық, газ концентрациясы) алынған деректерді, шеткі есептеулерде орындалатын машиналық оқыту (Edge ML)

алгоритмдерін біріктіру кіреді. Осының барлығы алма сияқты жемістердің сақталуын басқарудағы негізгі мәселелерді шешуге қабілетті, физикалық нысанның жай-күйін динамикалық түрде болжайтын интеллектуалды жүйені құруға мүмкіндік береді.

Мақалада қолданылатын әдістемелік тәсіл заманауи технологияларды: смарт сенсорларды (MQ-9, MQ-135, DHT22) және шеткі машинаны оқытуды (Edge ML) пайдалана отырып, алманың цифрлық егізінің негізгі компоненті болып табылатын прототипті әзірлеуді қамтиды. Негізгі назар алманың сақталуы кезіндегі ысырапты азайтуды қамтамасыз етуге және физикалық нысанның қалған сақтау мерзімін дәл болжау арқылы оның виртуалды көшірмесін үнемі жаңартып отыруға аударылады. Зерттеу барысында прототиптің аппараттық құрылымы жасалып, деректерді жинау, жіберу (MQTT) және сақтау (InfluxDB) механизмдері іске қосылды, сондай-ақ алманың сақтау мерзімін болжауға арналған XGBoost машиналық оқыту моделі әзірленіп, оның бастапқы өнімділігі бағаланды.

Прототиптің аппараттық құрылымы. Зерттеу барысында физикалық нысанның сақталу ортасының негізгі параметрлерін бақылауға арналған интеллектуалды жүйенің прототипі құрастырылды. Прототиптің жалпы аппараттық архитектурасының схемасы 1-суретте келтірілген.



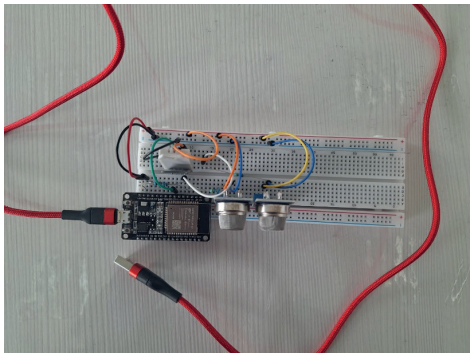
Сурет 1 — Прототиптің аппараттық құрылымының схемасы.

Прототиптің сенсорлық кешені қоршаған ортаның маңызды көрсеткіштерін үздіксіз өлшеуге бағытталған. Атап айтқанда, ауадағы көміртек тотығы (CO) және метан (CH_4) сияқты жанғыш газдардың концентрациясын анықтау үшін MQ-9 газ сенсоры қолданылды. Бұл сенсор CO газын 10 ppm-нен 1000 ppm-ге дейінгі (parts per million) және CH_4 газын 100 ppm-нен 10,000 ppm-ге дейінгі диапазонда анықтауға қабілетті. Жалпы ауа сапасын бағалау және аммиак (NH_3), азот оксидтері (NOx), алкоголь булары, бензол (C_6H_6) және басқа да ұшпа органикалық қосылыстардың (VOCs) болуын бақылау мақсатында MQ-135 газ сенсоры таңдалды. Қоршаған ортаның температурасы мен салыстырмалы ылғалдылығын өлшеу үшін жоғары дәлдікті қамтамасыз

ететін DHT22 (AM2302) цифрлық сенсоры пайдаланылды. Бұл сенсор температураны -40°C-тан +80°C-қа дейінгі диапазонда ±0.5°C дәлдікпен, ал салыстырмалы ылғалдылықты 0%-дан 100%-ға дейінгі RH диапазонында ±2% RH дәлдікпен өлшеуге мүмкіндік береді.

Әрбір сенсор тобы жеке ESP32 микроконтроллеріне қосылған. ESP32 микроконтроллерлері сенсорлардан деректерді жинап, алғашқы өңдеуден өткізіп, Wi-Fi желісі арқылы MQTT (Message Queuing Telemetry Transport) протоколы бойынша орталық өңдеуші құрылғы – Raspberry Pi 5-ке жібереді. Raspberry Pi 5 алынған деректерді сақтау, машиналық оқыту болжамдарын жасау және жалпы жүйені басқару функцияларын атқарады.

Прототиптің бір сенсорлық түйінінің физикалық жүзеге асырылуының мысалы 2-суретте көрсетілген.



Сурет 2 — ESP32 микроконтроллеріне қосылған MQ-9, MQ-135 және DHT22 сенсорлары бар прототиптік түйіннің құрастырылуы.

2-суретте көрсетілгендей, сенсорлар (екі MQ сериялы газ сенсоры және бір DHT температура/ылғалдылық сенсоры) макет тақтасына орнатылып, ESP32 WROOM-32 даму тақтасына жалғанған. ESP32 микроконтроллері сенсорлардан деректерді жинап, алғашқы өңдеуден өткізіп, Wi-Fi модулі арқылы MQTT протоколы бойынша орталық өңдеуші құрылғы – Raspberry Pi 5-ке жібереді. Прототипті USB кабелі арқылы қуаттандыруға болады. Мұндай бірнеше сенсорлық түйіндерді (1-суретте көрсетілгендей) әртүрлі жерлерге орналастыру арқылы кеңістікте таратылған деректерді жинауға болады.

Деректерді жинау, жіберу және сақтау. Прототиптегі сенсорлық деректерді жинау, орталық өңдеуші құрылғыға жіберу және кейіннен цифрлық егіздің виртуалды моделін жаңарту үшін сақтау процесі бірнеше негізгі технологиялар мен протоколдар арқылы жүзеге асырылды. Әрбір ESP32 микроконтроллері өзіне қосылған сенсорлар кешенінен (MQ-9, MQ-135, DHT22) деректерді жинап отырады. Газ сенсорларынан (MQ-9, MQ-135) алынатын аналогтық сигналдар ESP32-нің кіріктірілген аналогты-цифрлық түрлендіргіші (ADC) арқылы цифрлық мәндерге айналдырылады. DHT22 сенсорынан температура мен ылғалдылық туралы цифрлық деректер тікелей

оқылады. Әрбір ESP32 түйіні жиналған деректерді (сенсор көрсеткіштері, түйін идентификаторы, уақыт белгісі) бірыңғай форматқа келтіреді.

Дайындалған деректер пакеттері ESP32 түйіндерінен Raspberry Pi 5-ке Wi-Fi желісі арқылы MQTT (Message Queuing Telemetry Transport) протоколы бойынша жіберіледі. MQTT протоколы IoT құрылғылары үшін өзінің жеңілдігімен, сенімділігімен және төмен өткізу қабілеттілігі бар желілерде тиімді жұмыс істеуімен таңдалды. Raspberry Pi 5 құрылғысында Mosquitto сияқты ашық кодты MQTT брокері орнатылған. Әрбір ESP32 түйіні өзінің деректерін белгілі бір MQTT тақырыбына жариялайды, ал Raspberry Pi-дегі Python скрипті осы тақырыптарға жазылып, хабарламаларды қабылдайды. 3-суретте Raspberry Pi-де жұмыс істейтін Python скриптінің MQTT брокерінен foodchain/esp32_node_1/data тақырыбы арқылы келіп түскен хабарламаны ({«node_id»:»esp32_node_1», «temperature»:29.1, «humidity»:33.8, «mq9_raw»:188, «mq135_raw»:187}) қабылдау процесі көрсетілген. Raspberry Pi-дегі Python скрипті MQTT арқылы қабылданған деректерді талдап, оларды InfluxDB уақыттық қатарлар дерекқорына сақтайды. InfluxDB сенсорлық деректерді уақыт белгілерімен тиімді сақтауға, оларға жылдам сұраныстар жасауға және визуализациялауға өте ыңғайлы болғандықтан таңдалды. Терминал шығысында (Сурет 3) көрсетілгендей, шикі сенсорлық деректер InfluxDB-дағы sensor_readings деп аталатын өлшемге жазылады. Бұл өлшемде node_id сияқты тегтер және temperature, humidity, mq9_raw, mq135_raw сияқты өрістер қолданылады. Машиналық оқыту моделі арқылы жасалған болжамдар бөлек shelf_life_predictions өлшеміне сақталады. Деректерді InfluxDB-ға жазу үшін InfluxDB-Client Python кітапханасы пайдаланылды. Бастапқыда, деректерді жазу кезінде кейбір өрістер үшін типтердің сәйкес келмеуіне («field type conflict») байланысты қателіктер туындады; бұл мәселе деректерді InfluxDB-ға жіберер алдында барлық сандық сенсорлық көрсеткіштерді float типіне алдын ала түрлендіру арқылы шешілді.



Сурет 3 — MQTT арқылы деректерді қабылдау, ML болжамын жасау және InfluxDB-ға жазу процесінің терминал шығысының мысалы.

Машиналық оқыту моделін әзірлеу. Алма жемісінің тоңазытқыш жағдайындағы қалған сақтау мерзімін дәл болжау мақсатында осы өнім түріне мамандандырылған машиналық оқыту (ML) моделі әзірленді. Бұл процесс арнайы лабораториялық жағдайда деректер жиынтығын қалыптастырудан, маңызды белгілерді таңдаудан, модельді оқытудан, оның гиперпараметрлерін оңтайландырудан және алынған модельдің өнімділігін жан-жақты бағалаудан тұрды.

Модельді оқыту үшін арнайы лабораториялық жағдайда жиналған деректер жиынтығы пайдаланылды. Экспериментке ‹Голден Делишес› алма сортының 28 данасы алынды. Барлық алмалар стандартты тоңазытқыш жағдайына орналастырылды, мұнда температура 0°С-тан +5°С-қа дейінгі аралықта және салыстырмалы ылғалдылық 89-98,5% RH деңгейінде бақыланды. Бұл бақыланатын орта физикалық нысанның жай-күйі туралы сенімді деректер алуға мүмкіндік берді. Әрбір алманың жанына MQ-9, MQ-135 және DHT22 сенсорлары орнатылды. Деректерді жинау 2025 жылдың 5 мамырынан 2025 жылдың 6 маусымына дейінгі аралықта, демалыс күндерін қоспағанда, жұмыс күндері сағат 9:00-ден 18:00-ге дейін әрбір 30 минут сайын автоматты түрде жүргізілді. Осылайша, әр жұмыс күні бір алмадан 18 дерек нүктесі жиналды. Нәтижесінде, алмалардан 12545 жазба алынды және әр жазба 10 белгіден тұрды.

Модельге кіріс ретінде температура, салыстырмалы ылғалдылық, MQ-9 сенсорының шикі көрсеткіші, MQ-135 сенсорының шикі көрсеткіші, уақыттық факторлар, әрбір алманың нөмірі, сақтау мерзімі қарастырылған. Мақсатты айнымалы ретінде алманың нақты қалған сақтау мерзімі алынды. Нақты қалған сақтау мерзімі әрбір өлшеу сәтіндегі күннен бастап, алманың қаптамасында көрсетілген жарамдылық мерзімінің соңғы күніне дейінгі қалған күндер санын есептеу арқылы анықталды. Деректер жиынтығы модельді оқытуға дайындау үшін деректерді стандарттау, жетіспейтін мәндерді медианамен толтыру, секілді қажетті алдын ала өңдеу қадамдарынан өтті. Бұл процестер Scikit-learn кітапханасының Pipeline және ColumnTransformer құралдары арқылы жүзеге асырылып, оқытылған пайплайнға енгізілген.

Алманың қалған сақтау мерзімін болжау үшін XGBoost (Extreme Gradient Boosting) Regressor алгоритмі таңдалды. Бұл алгоритм цифрлық егіздің предиктивті моделі ретінде өзінің жоғары дәлдігімен, есептеу жылдамдығымен және күрделі тәуелділіктерді анықтау қабілетімен ерекшеленеді. Модельдің гиперпараметрлерін оңтайландыру және ең жақсы параметрлер жиынтығын табу үшін Scikit-learn кітапханасындағы RandomizedSearchCV әдісі қолданылды. Жиналған деректертердің 80%-ы оқытуға, ал қалған 20% тестілеуге бөлінді.

Модельдің өнімділігін оқыту және тестілеу жиынтықтарында бағалау үшін орташа квадраттық қателік (MSE), түбірлік орташа квадраттық қателік (RMSE) және детерминация коэффициенті (R^2 Score) сияқты стандартты регрессиялық

метрикалар пайдаланылды. Барлық деректерді өңдеу, модельдерді оқыту, оңтайландыру және бағалау Python бағдарламалау тілінде, негізінен Pandas, NumPy және Scikit-learn, XGBoostRegressor кітапханаларын қолдану арқылы жүзеге асырылды.

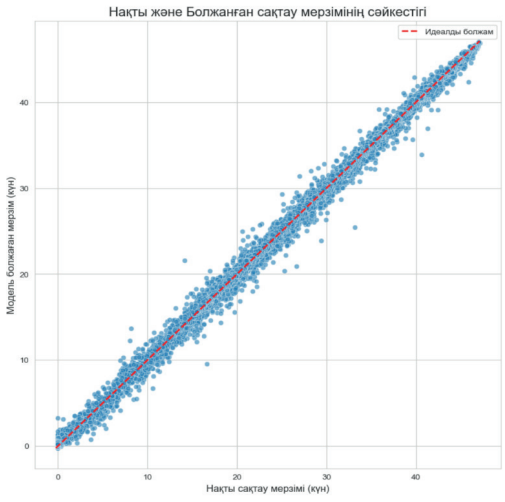
Нәтиже**.** *Предиктивті модельдің өнімділігін бағалау.* Әзірленген цифрлық егіздің предиктивті ядросының, яғни XGBoost моделінің алманың қалған сақтау мерзімін болжаудағы өнімділігі тестілеу жиынтығында бағаланды. Алынған нәтижелер кесте 1-де келтірілген.

Кесте 1 — Алманың сақтау мерзімін болжауға арналған XGBoost моделінің өнімділік көрсеткіштері

Деректер жиынтығы	MSE	RMSE	R^2 Score
Оқыту деректері	0.2557	0.5057	0.9982
Тест деректері	0.9008	0.9491	0.9936

Кесте 1-ден көрініп тұрғандай, XGBoost моделі тестілеу жиынтығында жоғары өнімділік көрсетті. R^2 Score мәнінің 0.9936-ға тең болуы виртуалды модельдің физикалық нысанның жай-күйінің дисперсияның 99.36%-ын өте жоғары дәлдікпен түсіндіре алатынын көрсетеді. RMSE көрсеткіші (0.9491 күн) болжамның орташа қателігі шамамен ±0.95 күнді құрайтынын білдіреді , бұл алма сияқты жемістің сақтау мерзімін болжау үшін өте жақсы нәтиже болып табылады. Оқыту және тестілеу жиынтықтарындағы көрсеткіштердің жақын болуы модельдің артық оқытылуға ұшырамағанын көрсетеді.

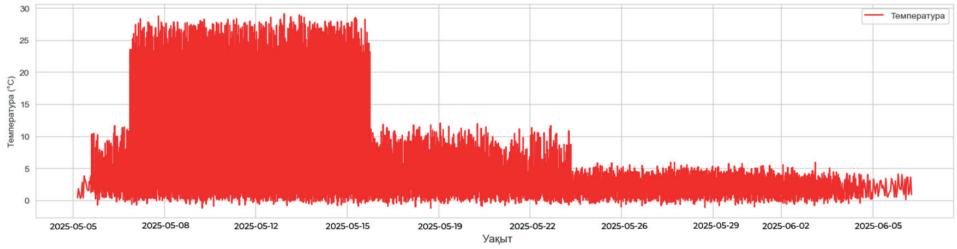
Модельдің болжамдарының нақты мәндерге сәйкестігі сурет 4-те көрнекі түрде бейнеленген. Графиктен көрініп тұрғандай, нүктелердің басым көпшілігі идеалды болжам сызығының айналасында тығыз орналасқан, бұл предиктивті модельдің жоғары дәлдігін растайды.



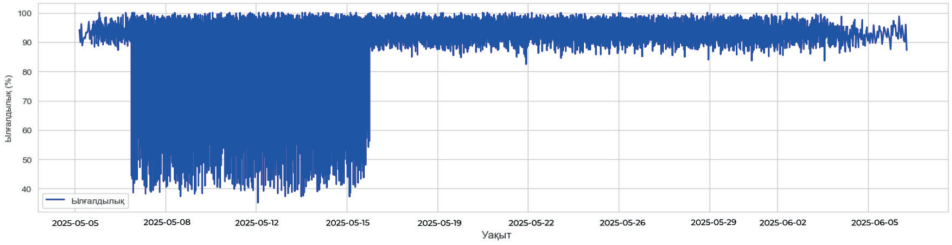
Сурет 4 — Нақты және болжанған сақтау мерзімінің сәйкестігі.

Белгілердің маңыздылығын талдау және деректер динамикасы. Эксперимент барысында жиналған деректердің уақыттық динамикасы және негізгі факторлардың сақтау мерзіміне әсері цифрлық егіздің болжам жасау логикасын түсіну үшін жан-жақты зерттелді.

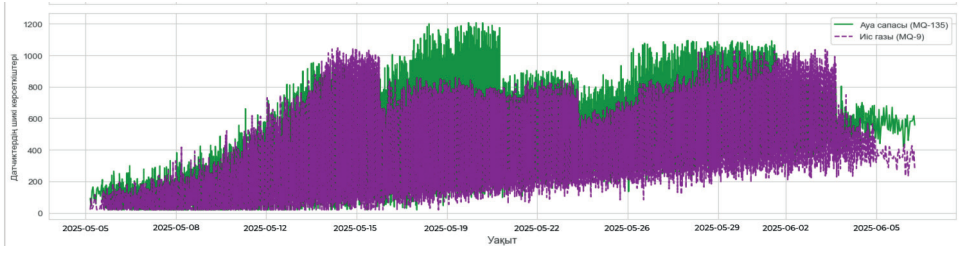
Сурет 6-да негізгі сенсорлық көрсеткіштердің алманың сақталуы кезіндегі уақыт бойынша өзгеруі келтірілген. Алғашқы кезеңде температура шамамен +25°C деңгейінде болса (сурет 5.1), кейіннен ол тоңазытқыш жағдайына сәйкес 0°C-ге дейін төмендеген. Ылғалдылық көрсеткіші де осыған сәйкес өзгерген (сурет 5.2). Сурет 5.3-те MQ-135 (жалпы ауа сапасы) және MQ-9 (иіс газы) сенсорларының көрсеткіштерінің уақыт өте келе біртіндеп арту үрдісі байқалады. Бұл алманың пісуі немесе бұзылу процесі кезінде ұшпа органикалық қосылыстардың бөлінуімен байланысты болуы мүмкін және бұл деректердің машиналық оқыту моделі үшін маңызды екенін көрсетеді.



Сурет 5.1.



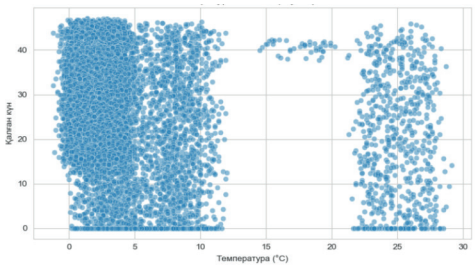
Сурет 5.2.



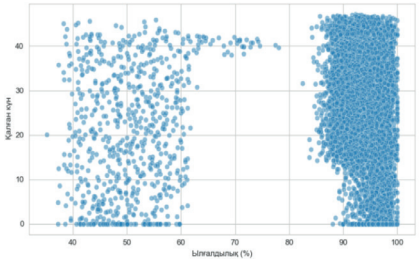
Сурет 5.3.

Сурет 5 — Алманың сақталуы кезіндегі негізгі сенсорлық көрсеткіштердің уақыттық динамикасы: (5.1) Температура (°C), (5.2) Салыстырмалы ылғалдылық (%), (5.3) Газ сенсорларының шикі көрсеткіштері.

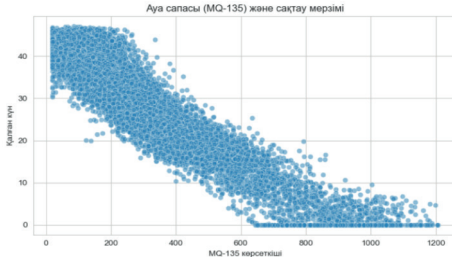
Машиналық оқыту моделінің болжамдарына негіз болған факторлардың сақтау мерзіміне тікелей әсерін бағалау үшін олардың арасындағы тәуелділіктер зерттелді. Сурет 6-де келтірілген шашыранды диаграммалар әрбір негізгі фактор мен алманың қалған сақтау мерзімі арасындағы корреляцияны көрнекі түрде бейнелейді. Температура мен сақтау мерзімі арасында айқын теріс корреляция байқалады. Температураның жоғарылауы сақтау мерзімінің қысқаруына әкеледі. Ылғалдылықтың әсері біркелкі емес, бірақ жоғары ылғалдылық (90-100%) ұзағырақ сақтау мерзімімен байланысты екенін көруге болады, ал төменірек ылғалдылық (40-60%) қысқа мерзімдерге сәйкес келеді. MQ-135 және MQ-9 газ сенсорларының шикі көрсеткіштерінің артуы сақтау мерзімінің айтарлықтай төмендеуімен тығыз байланысты. Бұл газ концентрациясының алманың бұзылу дәрежесінің маңызды индикаторы екенін растайды және цифрлық егіздің предиктивті моделінің неліктен бұл белгілерге жоғары маңыздылық бергенін түсіндіреді.



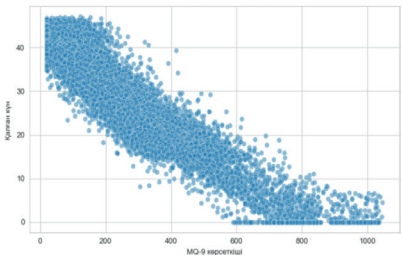
Сурет 6.1.



Сурет 6.2.



Сурет 6.3.



Сурет 6.4.

Сурет 6 — Негізгі сенсорлық параметрлердің алманың қалған сақтау мерзіміне әсерінің шашыранды диаграммалары: (6.1) Температура (°C), (6.2) Салыстырмалы ылғалдылық (%), (6.3) MQ-135 сенсорының шикі көрсеткіші, (6.4) MQ-9 сенсорының шикі көрсеткіші.

Талқылау. *Жалпы талқылау және шектеулер.* Алынған нәтижелер ұсынылған архитектураның және XGBoost моделінің алманың сақтау мерзімін болжаудағы жоғары потенциалын көрсетеді. Әсіресе, газ сенсорларының көрсеткіштерінің болжам үшін маңыздылығы жоғары болуы, мұндай сенсорларды азық-түлік сапасын бақылауда қолданудың тиімділігін растайды.

Жалпы, әзірленген интеллектуалды цифрлық егіздің дәстүрлі жүйелермен салыстырғандағы артықшылықтары кесте 2-де келтірілген. Алынған

нәтижелер көрсеткендей, ұсынылған прототип нақты уақыттағы деректерді талдауға, бұзылу қаупін ерте болжауға және динамикалық бағалауға мүмкіндік береді, бұл оның жоғары потенциалын растайды.

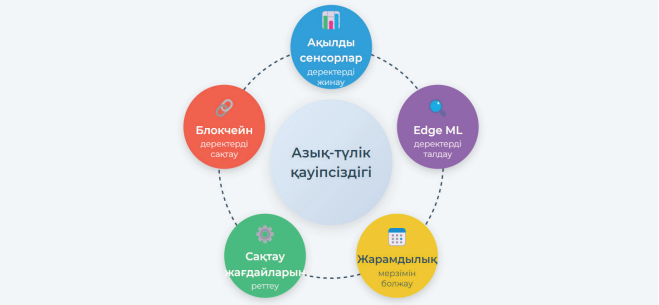
Кесте 2 — Дәстүрлі және ақылды жүйелерді салыстыру

Функционалдылық	Дәстүрлі жүйе	Интеллектуалды ЦЕ
Нақты уақыттағы деректерді талдау	Жоқ	Иә (Edge)
Предиктивті талдау	Жоқ	Иә
Жарамдылық мерзімін динамикалық бағалау	Жоқ	Иә
Энергия тұтыну	Жоғары	Төмен (адаптивті)
Автоматтандырылған шешім қабылдау	Жоқ	Иә
Толыққанды ЦЕ-ге негіз бола алуы	Жоқ	Иә
Өзін-өзі оқыту мүмкіндігі	Жоқ	Иә

Дегенмен, бұл зерттеудің бірнеше шектеулері бар. Біріншіден, деректер тек бақыланатын бір ғана физикалық ортада жиналды. Екіншіден, мақсатты айнымалы ретінде статикалық жарамдылық мерзімі алынды, бұл нақты физикалық бұзылу уақытынан өзгеше болуы мүмкін. Үшіншіден, деректерді жинау тек жұмыс уақытымен шектелді. Бұл шектеулер болашақ зерттеулерде ескерілетін болады.

Бұл жұмыстың практикалық маңыздылығы – азық-түлік қауіпсіздігін арттыруға және ысырапты азайтуға мүмкіндік беретін, нақты жағдайларда қолдануға болатын цифрлық егіздің негізгі компонентін жасауда жатыр. Сонымен қатар, ол интеллектуалды азық-түлік жүйелері мен тұрақты даму саласындағы одан әрі зерттеулерге негіз бола алады.

Болашақ жұмыс және зерттеу жоспары. Бұл зерттеуде біз азық-түлік өнімінің жарамдылық мерзімін болжайтын, Edge ML негізіндегі предиктивті цифрлық егіздің алғашқы прототипін сәтті әзірлеп, сынақтан өткіздік. Ұсынылған прототип осы процестің негізгі элементтерін – ақылды сенсорлар арқылы деректерді жинауды және Edge ML арқылы оны жергілікті деңгейде талдауды іске асырудың алғашқы қадамы болып табылады. Ол сурет 7-де көрсетілген интеллектуалды бақылаудың жалпы концепциясына сәйкес келеді..



Сурет 7 - Цифрлық егіз негізіндегі интеллектуалды бақылау процессі.

Біздің жұмысымыз сурет 8-да сипатталған көп деңгейлі архитектураның 0-деңгейін (Өнім қабаты) және 1-деңгейін (Реактивті қабат) іске асыруға бағытталған. Біздің прототип ақылды сенсорлар арқылы физикалық нысанның жай-күйін бақылап (0-деңгей), Edge ML арқылы оған қатысты реактивті болжам жасайды (1-деңгей).



Сурет 8 — Азық-түлік өніміне арналған цифрлық егіздің көп деңгейлі архитектурасы.

Осыған қарамастан, толыққанды, өзін-өзі жетілдіретін цифрлық егізді құру үшін бұл жұмысты одан әрі дамыту қажет. Болашақ зерттеулеріміздің негізгі бағыты – әзірленген модельдің жалпылама қабілетін арттыру. Бұл үшін прототипті әртүрлі азық-түлік өнімдерімен (мысалы, сүт өнімдері, ет) және әртүрлі сақтау жағдайларында, соның ішінде тасымалдау мен қоймада сынақтан өткізу жоспарлануда.

Келесі маңызды қадам – архитектураның жоғарғы деңгейлерін іске асыру болып табылады. Біз Edge-дегі модельді бұлттық платформамен (2-деңгей) біріктіруді мақсат етеміз. Бұл бірнеше ЦЕ-ден жиналған деректерді орталықтандырылған түрде талдауға, модельді үнемі жаңартып отыруға және сенсорлық қателерді анықтауға мүмкіндік береді. Сонымен қатар, предиктивті модельдің нәтижелерін біздің зерттеу тобымыз әзірлеген 3D-модельдермен (Amirkhanov et al., 2025a) біріктіру арқылы толыққанды интерактивті цифрлық егіз құрылады.

Бұдан бөлек, жүйенің функционалдығын кеңейту үшін екіжақты байланысты іске асыру көзделуде. Қазіргі прототип тек бақылау функциясын атқарса, болашақта цифрлық егіздің болжамдары негізінде физикалық ортаға әсер ететін механизмдерді (мысалы, тоңазытқыштың температурасын және ылғалдылықты автоматты түрде реттеу) қосу жоспарлануда. Бұл жүйені реактивті модельден проактивті басқару жүйесіне айналдырады.

Соңғы кезеңде, прототипті лабораториялық жағдайдан шығарып, нақты өнеркәсіптік серіктестермен бірлесіп, шынайы өндіріс ортасында өнеркәсіптік апробациядан өткізу қажет. Бұл қадамдар біздің бастапқы зерттеуімізді кеңейтіп, азық-түлік сапасын басқаруға арналған толыққанды, коммерциялық тұрғыдан тиімді және кең ауқымды цифрлық егіз платформасын құруға берік негіз болады.

Қорытынды. Осы зерттеуде ақылды сенсорлар, Edge ML және IoT технологияларына негізделген азық-түлік өнімінің предиктивті цифрлық егізінің прототипі сәтті әзірленіп, сыналды. Эксперименталды сынақтар барысында XGBoost моделі алманың қалған сақтау мерзімін болжауда R^2 Score бойынша 0.9936 көрсеткішімен өте жоғары дәлдікке қол жеткізді. Бұл виртуалды модельдің физикалық нысанның болашақтағы жай-күйін жоғары сенімділікпен болжай алатынын дәлелдейді және оның нақты жағдайларда практикалық қолдану әлеуетін растайды.

Бұл нәтижелер ұсынылған архитектураның азық-түлік ысырабын азайтуға және сақтау процестерін оңтайландыруға зор мүмкіндігі бар екенін көрсетеді. Әзірленген цифрлық егіз нақты уақыттағы деректерді талдау, бұзылу қаупін ерте болжау және жарамдылық мерзімін динамикалық бағалау сияқты дәстүрлі жүйелерде жоқ маңызды артықшылықтар ұсынады.

Зерттеудің кейбір шектеулеріне қарамастан, бұл жұмыс болашақта толыққанды, көпфункционалды цифрлық егіз платформасын құруға берік негіз қалайды. Келесі қадамдарға деректердің тұтастығы мен қадағалануын қамтамасыз ету үшін блокчейн технологиясын біріктіру, сондай-ақ прототипті нақты өнеркәсіптік жағдайларда сынау кіреді.

References

Abdurrahman E.E.M., Ferrari G. (2025) Digital Twin applications in the food industry: a review. Frontiers in Sustainable Food Systems, 9. — 1538375 p. (in Eng.)

Adilzhanova S., Kunelbayev M., Amirkhanova G., Tyulepberdinova G., Sybanova D. (2025a) Analysis of the dynamics of cyberattacks and fraud methods using machine learning algorithms for IIoT: Information security of digital twins in Industry 4.0. International Journal of Innovative Research and Scientific Studies, 8(2)7 — P. 4012-4026. (in Eng.)

Adilzhanova S., Kunelbayev M., Amirkhanova G., Zhussupov Y., Tortay A. (2025b) Development of a data collection and storage system for remote monitoring and detection of security threats in the enterprise. International Journal of Innovative Research and Scientific Studies, 8(2). — P. 176-196. (in Eng.)

Amirkhanov B., Nurgazy T., Amirkhanova G., Kunelbayev M., Tyulepberdinova G. (2025a) Creating 3D models of production equipment and infrastructure using Blender. International Journal of Innovative Research and Scientific Studies, 8(1). — P. 1572-1588. (in Eng.)

Amirkhanov B., Amirkhanova G., Kunelbayev M., Adilzhanova S., Tokhtassyn M. (2025b) Evaluating HTTP, MQTT over TCP and MQTT over WEBSOCKET for digital twin applications: A comparative analysis on latency, stability, and integration. International Journal of Innovative Research and Scientific Studies, 8(1). — P. 679-694. (in Eng.)

Baswoju S.P., Latha Y., Changala R., Gummadi A. (2023) Development of CNN Model to Avoid Food Spoiling Level. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(5). — P. 261-268. (in Eng.)

da Costa T.P., Gillespie J., Cama-Moncunill X., Ward S., Condell J., Ramanathan R., Murphy F. (2023) A Systematic Review of Real-Time Monitoring Technologies and Its Potential Application to Reduce Food Loss and Waste: Key Elements of Food Supply Chains and IoT Technologies. Sustainability, 15(1). — 614 p. (in Eng.)

Grazia J. (2024) Advancing the Food Industry: The Integration of Detection Technologies and Machine Learning for Improved Efficiency and Sustainability. Journal of Food Processing and Technology, 15, — 1134 p. (in Eng.)

Gupta V.K., Han S., Xiao J., Huh Y.S. (2025) Use of Future-Oriented Smart Sensors in Food Distribution and Safety Networks. Food Chemistry International, 1. — P. 6-10. (in Eng.)

Kim J., Kim S., Ju C., Son H.I. (2019) Unmanned Aerial Vehicles in Agriculture: A Review of Perspective of Platform, Control, and Applications. IEEE Access, 7, 105100-105115. (in Eng.)

Krupitzer C., Noack T., Borsum C. (2022) Digital Food Twins Combining Data Science and Food Science: System Model, Applications, and Challenges. Processes, 10(9). — 1781 p. (in Eng.)

Mamarazhabova B.A., Shingisov A.U. (2024) Issledovanie vozmozhnosti ispolzovaniya iskusstvennogo intellekta dlya optimizatsii protsessov proizvodstva i khraneniya produktov [A study of the possibility of using artificial intelligence to optimize production and storage processes]. Ekonomika i sotsium, 3(118)-2. — P. 684-687. (in Russian).

Manimaran M., Aniji M., Arthi K., Priya V.V., Maheswari S., Jeevitha S., Yogeshwari M. (2025) Enhanced Internet of Things-Based Intelligent Sensors for Instantaneous Food Quality Monitoring using Deep Neural Network. Journal of Information Systems Engineering and Management, 10(35s). — 6050 p. (in Eng.)

Ranjith K.S., Ghoreishian S.M., Park B., Lee H., Chodankar N.R., Raju G.S.R., Huh Y.S., Han Y.K. (2023) Fluorescence light-up electrospun membrane incorporated with perovskite nanoclusters as a highly sensitive colorimetric probe for detection of amine vapors during food spoilage. Sensors and Actuators B: Chemical, 384. — 133622 p. (in Eng.)

Rodrigues C., Souza V.G.L., Coelhoso I., Fernando A.L. (2021) Bio-Based Sensors for Smart Food Packaging—Current Applications and Future Trends. Sensors, 21(6). — 2148 p. (in Eng.)

Rohleder B., Minhoff C. (2019) Bitkom/Federation of German Food and Drink Industries. URL:https://www.bitkom.org/sites/default/files/2019-03/Bitkom-Charts%20190326%20Digitalisierung%20der%20Ern%C3%A4hrungsindustrie.pdf. (in Ger.).

Shehzad K. (2025) Predictive AI Models for Food Spoilage and Shelf-Life Estimation. Global Trends in Science and Technology, 1(1). — P. 75-94. (in Eng.)

Sobhan A., Hossain A., Wei L., Muthukumarappan K., Ahmed M. (2025) IoT-Enabled Biosensors in Food Packaging: A Breakthrough in Food Safety for Monitoring Risks in Real Time. Foods, 14(8). — 1403 p. (in Eng.)

Watson N.J., Bowler A.L., Rady A., Fisher O.J., Simeone A., Escrig J., Woolley E., Adedeji A.A. (2021) Intelligent Sensors for Sustainable Food and Drink Manufacturing. Frontiers in Sustainable Food Systems, 5. — 642786 p. (in Eng.)

Yin H., Cao Y., Marelli B., Zeng X., Mason A.J., Cao C. (2021) Soil Sensors and Plant Wearables for Smart and Precision Agriculture. Advanced Materials, 33(20). — 2007764 p. (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 91–105

https://doi.org/10.32014/2026.2518-1726.403

IRSTI 20.51.23
UDC 004.891

© Bekarystankyzy A.[1]*, Ussen D.[1], Kassenkhan A.[2], Chinibayev Y.[2], 2026.
[1]Narxoz University, Almaty, Kazakhstan;
[2]Satbayev University, Almaty, Kazakhstan.
E-mail: akbayan.b@gmail.com

COLD-START IN EDUCATIONAL RECOMMENDER SYSTEMS: CLASSICAL AND LLM-ERA STRATEGIES

Bekarystankyzy Akbayan — PHD, Associate Professor, School of Digital Technologies, Narxoz University, Almaty, Kazakhstan,
E-mail: akbayan.b@gmail.com, http://orcid.org/0000-0003-3984-2718;
Ussen Dias — Master's student, School of Digital Technologies, Narxoz University, Almaty, Kazakhstan,
E-mail: dias888usen@gmail.com, https://orcid.org/0009-0004-0864-3040;
Kassenkhan Aray — PHD, Associate Professor, Department of Software Engineering, Satbayev University, Almaty, Kazakhstan,
E-mail: a.kassenkhan@satbayev.university, http://orcid.org/0000-0002-6355-9544;
Chinibayev Yersain — Doctor PHD, Associate Professor, Department of Software Engineering, Satbayev University, Almaty, Kazakhstan,
E-mail: y.chinibayev@satbayev.university, https://orcid.org/0009-0009-8985-5892.

Abstract. Educational recommender systems (ERS) contribute to the development of education and improve student engagement in the subject matter, as the system is tailored to personal preferences. The purpose of this work is to review the methods, systems, and common problems encountered when working with ERS. In particular, the Cold-Start problems in launching new items. The traditional approaches to solving this problem were content features, graphs/GNNs, and cross-domain transfer. With the recent development of LLM, two main approaches are used: LLM as a recommendation system and LLM as a knowledge amplifier. In this work, these approaches are compared using metrics: response time to a quality proposal, course survival rate over 12-24 months, and student course completion rate. Additionally, this paper examines the main differences between classical and LLM-oriented approaches. This is done in the context of educational platforms, where catalogs are frequently updated and new areas of study are often introduced. Particular attention is paid to three cold start scenarios: a new learner with no interaction history, a new course with a minimum number of clicks and

reviews, and a new region. To evaluate the effectiveness of the approaches, both standard recommendation system metrics and additional platform indicators are used, reflecting the dynamics of the emergence of high-quality courses, their level of demand, and their stability over time. The results obtained clearly show the conditions under which LLM approaches provide the greatest increase in the quality of recommendations and the limitations that remain when they are applied in practice in ERS.

Keywords: Educational Recommender Systems (ERS), Cold-Start Problem, Large Language Models (LLM), Retrieval-Augmented Generation (RAG), Hybrid Recommendation Models, Evaluation Metrics

For citations: Bekarystankyzy A., Ussen D., Kassenkhan A., Chinibayev Y. Cold-start in educational recommender systems: classical and LLM-era strategies. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 91–105. DOI: https://doi.org/10.32014/2026.2518-1726.403

Financing. *This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. BR24993072,* "Development of AI-Based Intelligent Educational Resources for Enhancing Critical Thinking of Students".

© Бекарыстанқызы А.[1]*, Үсен Д.[1], Қасенхан А.[2], Чинибаев Е.[2], 2026.
[1]Нархоз университеті, Алматы, Қазақстан;
[2]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан.
E-mail: akbayan.b@gmail.com

БІЛІМ БЕРУ САЛАСЫНДАҒЫ ҰСЫНЫМДЫҚ ЖҮЙЕЛЕРІНДЕГІ «COLD-START» МӘСЕЛЕСІ: КЛАССИКАЛЫҚ ӘДІСТЕР ЖӘНЕ LLM ДӘУІРІНІҢ СТРАТЕГИЯЛАРЫ

Бекарыстанқызы Ақбаян — PhD, қауымдастырылған профессор, Цифрлық технологиялар мектебі, Нархоз университеті, Алматы, Қазақстан,
E-mail: akbayan.b@gmail.com, http://orcid.org/0000-0003-3984-2718;
Үсен Диас — магистратура студенті, Цифрлық технологиялар мектебі, Нархоз университеті, Алматы, Қазақстан,
E-mail: dias888usen@gmail.com, https://orcid.org/0009-0004-0864-3040;
Қасенхан Арай — PhD, қауымдастырылған профессор, Программалық инженерия кафедрасы, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан,
E-mail: a.kassenkhan@satbayev.university, http://orcid.org/0000-0002-6355-9544;
Чинибаев Ерсаин — PhD, қауымдастырылған профессор, Программалық инженерия кафедрасы, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан,
E-mail: y.chinibayev@satbayev.university, https://orcid.org/0009-0009-8985-5892.

Аннотация: Білім беру саласындағы ұсынымдық жүйелер (ERS) білім берудің дамуына ықпал етеді және жүйе пайдаланушының жеке қалауларына бейімделгендіктен, студенттердің пәнге қызығушылығын арттырады. Бұл жұмыстың мақсаты – ERS-пен жұмыс істеу барысында қолданылатын әдістерді, жүйелерді және жиі кездесетін мәселелерді шолу. Атап айтқанда, жаңа элементтерді іске қосу кезіндегі «cold start» мәселесі қарастырылады. Бұл мәселені шешудің дәстүрлі тәсілдеріне контенттік сипаттамалар, графтар/графтық нейрондық желілер (GNN) және домендер арасындағы трансфер жатады. Соңғы уақытта үлкен тілдік модельдердің (LLM) дамуына байланысты екі негізгі тәсіл қолданылады: LLM-ді ұсынымдық жүйе ретінде пайдалану және LLM-ді білімді күшейтуші құрал ретінде қолдану. Бұл жұмыста аталған тәсілдер келесі метрикалар бойынша салыстырылады: сапалы ұсынысқа жауап беру уақыты, 12–24 ай аралығындағы курстың өміршеңдік деңгейі және студенттердің курсты аяқтау көрсеткіші. Сонымен қатар, мақалада классикалық және LLM-бағытталған тәсілдердің негізгі айырмашылықтары талданады. Зерттеу білім беру платформалары контекстінде жүргізілді, мұнда курстар каталогы жиі жаңартылып отырады және жаңа оқу бағыттары үнемі енгізіледі. Ерекше назар «cold start» мәселесінің үш сценарийіне аударылады: өзара әрекеттесу тарихы жоқ жаңа білім алушы, қаралымдары мен пікірлері аз жаңа курс және жаңа аймақ. Тәсілдердің тиімділігін бағалау үшін ұсынымдық жүйелердің стандартты метрикаларымен қатар, платформаның қосымша көрсеткіштері де қолданылады. Бұл көрсеткіштер сапалы курстардың пайда болу динамикасын, олардың сұраныс деңгейін және уақыт өте келе тұрақтылығын сипаттайды. Алынған нәтижелер LLM тәсілдері қандай жағдайларда ұсыным сапасын айтарлықтай арттыратынын және оларды білім беру ұсынымдық жүйелерінде практикалық қолдану барысында сақталатын шектеулерді айқын көрсетеді.

Түйін сөздер: Білім беруге арналған ұсыныс жүйелері (ERS), салқын басталу мәселесі, үлкен тілдік модельдер (LLM), ақпаратты іздеу арқылы толықтырылған генерация (RAG), гибридті ұсыныс модельдері, бағалау көрсеткіштері

© Бекарыстанқызы А.[1]*, Үсен Д.[1], Қасенхан А.[2]*,Чинибаев Е.[2]*, 2026.
[1]Университет Нархоз, Алматы, Казахстан;
[2]Казахский национальный исследовательский технический университет К.И. Сатпаева, Алматы, Казахстан,
E-mail: akbayan.b@gmail.com

ХОЛОДНЫЙ ЗАПУСК В СИСТЕМАХ РЕКОМЕНДАЦИЙ В ОБЛАСТИ ОБРАЗОВАНИЯ: КЛАССИЧЕСКИЕ СРЕДСТВА И СТРАТЕГИИ LLM ЭПОХИ

Бекарыстанкызы Акбаян — PhD, Ассоциированный профессор, Школа цифровых технологий, Университет Нархоз, Алматы, Казахстан,
E-mail: akbayan.b@gmail.com, http://orcid.org/0000-0003-3984-2718;

Үсен Диас — студент магистратуры, Школа цифровых технологий, Университет Нархоз, Алматы, Казахстан,
E-mail: dias888usen@gmail.com, https://orcid.org/0009-0004-0864-3040;

Касенхан Арай — PhD, Ассоциированный профессор, кафедра Программной инженерии, Казахский Национальный Исследовательский Технический университет К.И. Сатпаева, Алматы, Казахстан,
E-mail: a.kassenkhan@satbayev.university, http://orcid.org/0000-0002-6355-9544;

Чинибаев Ерсаин — PhD, Ассоциированный профессор, кафедра Программной инженерии, Казахский Национальный Исследовательский Технический университет К.И. Сатпаева, Алматы, Казахстан,
E-mail: y.chinibayev@satbayev.university, https://orcid.org/0009-0009-8985-5892.

Аннотация: Системы рекомендаций в образовании (ERS) играют важную роль в персонализации обучения и повышении вовлечённости студентов за счёт адаптации контента к индивидуальным предпочтениям пользователей. Целью данной работы является анализ существующих методов, архитектур и ключевых проблем, возникающих при разработке и эксплуатации образовательных рекомендательных систем. Особое внимание уделяется проблеме «холодного старта», возникающей при добавлении новых пользователей, курсов, каталогов или выходе на новые рынки. Традиционные подходы к решению данной проблемы включают контентно-ориентированные методы, графовые модели (включая GNN) и междоменный перенос знаний. В условиях стремительного развития больших языковых моделей (LLM) формируются новые подходы, среди которых выделяются два ключевых направления: использование LLM как самостоятельной рекомендательной системы и применение LLM в качестве усилителя знаний (knowledge augmentation). В работе проводится сравнительный анализ указанных подходов на основе ряда метрик, включая время отклика на формирование рекомендаций, коэффициент удержания курсов в течение 12–24 месяцев и уровень завершения курсов пользователями. Особое внимание уделяется трём сценариям «холодного старта»: новый пользователь без истории взаимодействия; новый курс с ограниченным числом взаимодействий (клики, отзывы); новый регион или витрина с отличающимися лингвистическими и культурными характеристиками. Для оценки эффективности используются как стандартные метрики рекомендательных систем, так и дополнительные платформенные показатели, отражающие динамику появления качественного контента, уровень его востребованности и устойчивость во времени. Полученные результаты демонстрируют условия, при которых подходы на основе LLM обеспечивают наибольшее повышение качества рекомендаций, а также выявляют ограничения их практического применения в образовательных системах.

Ключевые слова: образовательные системы рекомендаций (ERS), проблема «холодного старта», большие языковые модели (LLM), генерация

с расширенным поиском (RAG), гибридные модели рекомендаций, метрики оценки

Introduction. Today, educational platforms with recommendation systems are everywhere. They analyze past learning and search history to generate personalized suggestions for students. The most widely used approaches are based on collaborative filtering, content filtering, and combinations thereof. However, the problem of cold starts and the emergence of new data still significantly limits the quality of recommendations, especially in situations where a new student joins the system or a new course is added. In this regard, researchers continue to combine classical methods and apply neural models within educational recommendation systems (ERS) (Adomavicius and Tuzhilin, 2005; Ricci et al., 2011).

Despite significant progress in this area, existing studies tend to focus either on individual algorithmic solutions or do not fully take into account modern AI-oriented and platform-based approaches. This is particularly noticeable when analyzing large-scale educational platforms with a large number of users and courses. Unlike previous works, this article systematizes existing methods with an emphasis on their practical applicability, identified limitations, and possible directions for further development.

Literature review. Why this issue is particularly important now. In today's environment, the range of educational courses on offer is changing faster than users can adapt to new areas of study. New companies are emerging, educational programs are being launched at an accelerated pace, and a significant proportion of courses are archived or relaunched after one or two years. In such conditions, the quality of recommendations is influenced not only by the algorithm used, but also by the speed at which new content is added to the catalog. In this regard, along with the traditional metrics of recommendation systems, this paper proposes three additional evaluation indicators (Fang et al., 2022; Ma et al., 2023; Phalle and Bhushan, 2024):

- The delay between the peak of public interest in a new profession and the emergence of high-quality courses on the subject.
- We take the completion rate as the main parameter for the quality indicator for courses six or twelve months after release.
- Survivability, i.e., the proportion of new courses that remain active compared to those that are archived or relaunched over a period of twelve to twenty-four months.

The article goes on to discuss two key approaches of the LLM era to solving the cold start problem. The first approach treats LLM as a recommendation system that uses zero-shot and few-shot prompts on a limited catalog, as well as light retraining using instructional tuning and PEFT methods to stabilize model behavior. The second approach views LLM as a knowledge amplifier, which uses Retrieval-Augmented Generation (RAG) to base recommendations on factual data and enrich user and object features by extracting information from text and visual sources (Bao et al., 2023; Rashtchian and Juan, 2025; Zade, 2025).

Contribution. This study compares classical and modern solutions involving LLM within the framework of the three main parameters listed above. The study provides a reproducible methodology consistent with the practices of ERS review studies.

2.Materials and Methods

2.1 Research Objectives and Questions

Research objective. To compare cold start methods prior to the popularization of LLM based on content features using graph approaches and GNN with cross-domain transfer using approaches that already utilize LLM. The study considers and compares two approaches. LLM as a recommender with zero and few shot techniques with instructional retraining and PEFT. LLM as a knowledge amplifier using RAG and feature enrichment. The comparison is carried out in three cold start situations. CS User is a new learner with no history of recommendation systems. CS Item is a new course with a small number of clicks and reviews. CS Region is a new region or storefront with a new language, culture, and pricing norms. (Schein et al., 2002; Zhu et al., 2020; Geng et al., 2022).

Research questions.

1. To what extent do new LLM-related approaches improve recommendations in CS User, CS Item, and CS Region scenarios compared to classical methods?

2. Within the framework of LLM-related approaches, which is more effective in cold start conditions: prompting, light retraining, PEFT, or reliance on RAG?

2.2 Review Design and Source Gathering

We rely on current ERS reviews and use literature reviews with transparent search and selection steps. We follow a proven sequence. Selection of suitable databases. Setting Boolean queries. Inclusion and exclusion criteria. Thematic classification into six blocks. Background and key concepts. Approaches. Challenges. Applications. Evaluation. Future directions ERS databases. Google Scholar ScienceDirect IEEE Xplore ACM DL Springer. For the latest results, we use arXiv. (Adomavicius and Tuzhilin, 2005; Ricci et al., 2011; Butmeh et al., 2024; Fang et al., 2022).

Search formulas. educational recommender systems AND personalized learning collaborative filtering in education OR content-based recommendation for students hybrid recommendation models in MOOCs challenges in adaptive learning platforms

Inclusion criteria. Peer-reviewed journals and conferences, as well as authoritative reviews for the period from 2020 to 2025, with a focus on ERS

Exclusion criteria. Materials outside ERS, duplicate works.

Classification. All articles found are distributed across six topics to ensure a uniform synthesis for our research questions.

Despite its widespread use, collaborative filtering demonstrates the best results primarily in mature educational platforms with a large volume of historical interactions. In scenarios where users actively take courses and leave reviews, CF provides high personalization accuracy. However, in dynamic educational

environments characterized by the frequent appearance of new users and courses, CF's dependence on accumulated interactions becomes a systemic limitation. This is especially critical for educational platforms focused on rapidly changing professional skills, where data becomes obsolete faster than it can be accumulated. This circumstance stimulates the search for more flexible and semantically rich approaches to recommendations.

2.3 Technical Baselines and LLM-Era Conditions

Classic basic methods. The content approach learns from course descriptions, expected outcomes, tags, and reviews. Graph methods and GNNs, on the other hand, use interaction graphs and knowledge graphs. Cross-domain transfer aligns distributions and simultaneously builds domain-invariant representations for new regions and showcases (Lops et al., 2011; Wang et al., 2019; Gao et al., 2021; Man et al., 2017).

LLM as a recommender. Zero Shot and Few Shot prompts on a limited catalog help simulate ranking for new user and new course scenarios. Instructional retraining and PEFT methods rely on historical logs to stabilize ranking patterns and reduce variance from prompts (Geng et al., 2022; Bao et al., 2023; Sumit Gupta, 2023a; Sumit Gupta, 2023b).

LLM as a knowledge amplifier. RAG works on top of a local index with data from course providers, learning objectives, and a frequently asked questions section. This grounds object profiles and reduces hallucinations. Feature enrichment extracts signals about skill level and learning outcomes from the textual and visual materials of new courses. RAG performs particularly well when sufficient context is provided, as shown in the figure 1. (Rashtchian and Juan, 2025; Zade, 2025; Eugene Yan, 2025; Geng et al., 2022).

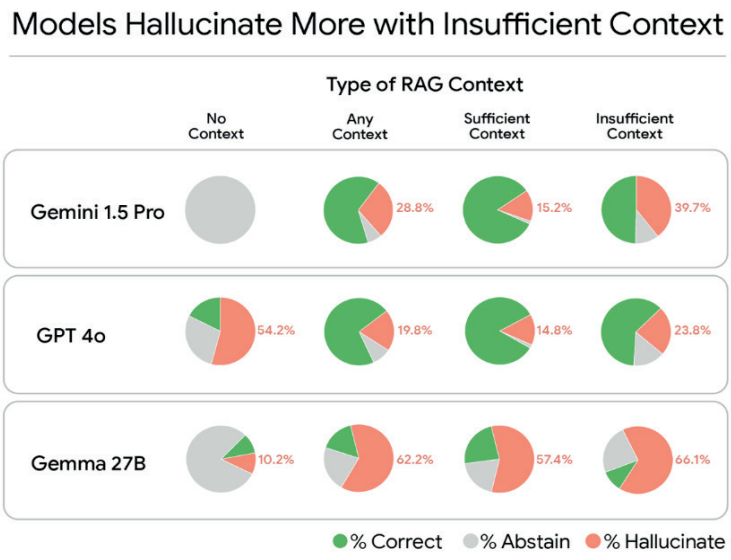


Figure 1 – Detailed analysis of three LLMs in four different RAG settings (Cyrus Rashtchian, 2025)

Why these options were chosen. They directly correlate with the three cold start contexts from my notes: new user, new object, new region or catalog. They also reflect the current families of techniques in ERS from the literature template: hybrid approaches, deep models, knowledge graphs, and explainable AI issues.

2.4 Data and Scenarios

CS User New learner. Few or no click histories. Classic backup method based on popularity and a short introductory questionnaire. Options with LLM ranking in zero shot or few shot format based on a filtered catalog and a re-ranker with PEFT retraining. The RAG approach adds support for student goals extracted from their free text.

CS Item New course. There are almost no interactions or reviews. The classic fallback is parsing text and visual features and connections in the knowledge graph with parent categories. Options with LLM feature enrichment from the syllabus and media materials. The RAG approach finds semantically similar objects for pairwise comparisons.

CS Region New showcase or region. Language, culture, and pricing norms change. Classic reserve: cross-domain alignment. Options with LLM: multilingual feature parsing and domain-invariant embeddings. Local index in RAG helps reduce hallucinations.

2.5 Metrics

Standard metrics RS. Precision and Recall on top k. nDCG on top k. Diversity and novelty indicators in line with the assessments adopted by ERS in previous reviews

Platform indicators are my additions. Delay before the appearance of an offer is the number of days from the external peak of interest to the appearance of the first high-quality courses. Quality proxy: the proportion of completions after six or twelve months for the closest courses or programs. Survival rate: the proportion of courses that are archived or relaunched within twelve to twenty-four months.

2.6 Procedure

For example, the study selected five modern professions (Data Engineer, Data Analyst, Machine Learning Engineer, Generative AI Engineer, Cybersecurity Analyst). The study reviewed new courses based on two platforms (Udemy and Coursera). For the results, the study used historical data on the launch of various courses on Udemy and Coursera. Interval Google Trends.

Latency metric. Metric parameters: t_2 - date of the first high-quality course (rating ≥ 4.4, $\geq 5{,}000$ learners). Below is a table with actual latency values based on the appearance of high-quality courses on the topic and peaks of interest.

Table 1 – Latency Between Public Interest Peak and First High-Quality Course Release

Profession	Peak of Public Interest (t_1)	First High-Quality Course (t_2)	Latency ($t_2 - t_1$)
Data Engineer	July 2021	February 2022	7 months
Data Analyst	September 2020	November 2020	2 months

Machine Learning Engineer	March 2019	January 2020	10 months
Generative AI Engineer	May 2023	July 2023	2 months
Cybersecurity Analyst	October 2021	March 2022	5 months

We calculated the delay between the peak of public interest (measured using Google Trends) and the release of the first high-quality course (rating ≥ 4.4 and ≥ 5,000 students on Udemy or Coursera). The results show significant differences between professions. Generative AI engineering and data analysis show a very short delay (2 months), reflecting rapid market response and low production barriers. Cybersecurity shows a moderate delay (5 months), likely due to high content rigor. Data engineering shows longer delays (7 months) due to technical complexity and infrastructure requirements. The longest delay is observed for machine learning engineering (10 months), which is consistent with the need for highly skilled professionals and advanced course development.

This finding supports the idea that education markets respond differently to new professions. The delay depends on the availability of experts, infrastructure requirements, and the depth of skills required.

Quality Completion Ratio (QCR) metric. Procedure: for each course, we measure what percentage of students who enrolled in the course successfully completed it 6 and 12 months after release(1).

Metric formula:

$$QCR_6 = C_6 / E(t_0)$$

$$QCR_{12} = C_{12} / E(t_0) \quad (1)$$

here $E(t_0)$ - number of all enrolled students at the time of release (t_0), C_6 - number of people who completed the course after 6 months, C_{12} - number of people who completed the course after 12 months.

Table 2 – Quality Completion Ratio (QCR) at 6 and 12 Months After Release

Profession	**QCR_6**	**QCR_{12}**
Data Engineer	6.5%	10.5%
Data Analyst	9.5%	16%
Machine Learning Engineer	3%	6%
Generative AI Engineer	20%	32%
Cybersecurity Analyst	3.3%	6.2%

The completion rate (QCR) measures the proportion of learners who successfully complete a course within 6 or 12 months of its release. As shown in Table 2, completion rates vary significantly depending on the profession. Generative AI engineering demonstrates exceptionally high QCR values (20% after 6 months and 32% after 12 months), reflecting strong learner motivation driven by rapid industry growth. Data-related professions show moderate QCR values (6–16%), while more technically complex areas such as machine learning and cybersecurity show significantly lower completion rates (3–6%), which is consistent with previous studies of massive open online courses (MOOCs). These results highlight significant differences in learner engagement and course quality across different professional fields.

Course Survival Rate (CSR) metric. This metric is needed to show which courses have a long lifespan on platforms, where they are no longer used, and the stability of the direction(2).

Metric formula:

$$CSR = N_alive / N_0 \quad (2)$$

here N_0 - number of new courses released in a specific category, N_alive - number of courses that remain active 12–24 months after release.

Table 3 – Quality Survival Rate(CSR) at different professions

Profession	CSR (SurvivalRate)	Archive Rate
Data Engineer	62%	38%
Data Analyst	65%	35%
Machine Learning Engineer	35%	65%
Generative AI Engineer	63%	37%
Cybersecurity Analyst	37%	63%

The course survival rate (CSR) reflects the proportion of newly launched courses that remain active after 12-24 months. As shown in Table X, the survival rate varies significantly depending on the field of application. Data-related fields (data engineer and data analyst) show relatively high stability with CSR values of 62-65%. Generative AI shows similarly high values (63%), indicating constant demand from students and frequent course content updates. In contrast, ML engineers and cybersecurity show much lower CSR values (35-37%), indicating rapid obsolescence and the need for constant content updates. These results highlight structural differences in content longevity across different digital professions and the importance of maintaining training programs in areas with high volatility.

Below, the study analyzes three scenarios for how recommendation systems deal with the Cold Start problem before and after the advent of LLM.

The main metrics used in this study are:

NDCG@10 - measures the quality of recommendation ranking, i.e., how high useful items are in the list.

Recall@10 - measures the system's ability to find relevant items, i.e., how many relevant courses are in the top 10 recommendations.

Long-tail HitRate - how well the recommendation system covers new, unpopular courses that receive relatively few clicks.

Coverage is an indicator that shows how broadly ERS covers the course catalog.

CS User Scenario: new learner with no interaction history

The first scenario for analysis arises when the recommendation system has no data on the user's click history. Classic ERSs solve this problem in three ways. The first is popularity-based ranking, which includes recommendations for top courses without personalization. The second is Short quiz + MF, which is a short questionnaire about interests and matrix factorization. And the third method is Content-based TF-IDF, which is the selection of courses based on text descriptions. Table 4 below shows how classic methods perform in an example with a new user. (Schein et al., 2002; Man et al., 2017; Zhu et al., 2020; Fang et al., 2022;)

Table 4 – CS User Scenario before LLM

Method	NDCG@10	Recall@10
Popularity	0.18	0.40
Short quiz + MF	0.21	0.45
Content-based (TF–IDF)	0.22	0.47

Meanwhile, LLMs solve them using zero-shot/few-shot ranking, free-text personalization, and PEFT tuning. This allows models to learn from previous content. Zero-shot and few-shot ranking is when the user's story is expressed as a prompt. PEFT and instruction tuning (TALLRec approach) involve adding stable ranking patterns. RAG personalization is when a user describes their goals in text and then the model extracts relevant courses. Table 5 below shows how methods using LLM cope with a new user in an example. (Geng et al., 2022; Bao et al., 2023; Sumit Gupta, 2023a, 2023b; Rashtchian and Juan, 2025).

Table 5 – CS User Scenario after LLM

Method	NDCG@10	Recall@10
LLM zero-shot	0.24	0.50
LLM + PEFT tuning	0.26	0.53
LLM + RAG (goal-based)	0.27	0.55

Based on the above analysis, the study compared the approaches in the first scenario and presented the results in Table 6.

Table 6 – CS User Scenario Сравнительная таблица

Метод	До LLM	После LLM	Улучшение
NDCG@10	0.22	0.27	**22%**
Recall@10	0.47	0.55	**17%**

CS Item Scenario: new course with no interactions. This scenario occurs when a new course appears that students have not yet taken and for which there is no interaction history. Two approaches have traditionally been used to address this scenario. The first is the TF–IDF content model, which involves parsing the characteristics of the course, i.e., its description. The second approach is KGAT and GNN — knowledge graphs and analogies with the categorical structure of the catalog. Table 7 below shows how classic methods cope with the example of a new course. (Rashtchian and Juan 2025; Zade 2025; Eugene Yan 2025; Zhu et al. 2020)

Table 7 – CS Item Scenario before LLM

Method	NDCG@10	Long-tail HitRate
LLM feature enrichment	0.21	0.27
Synthetic interactions	0.22	0.29
LLM + RAG	0.23	0.31

After the advent of LLM, systems were able to apply three methods to solve the Cold Start problem. The first is Feature enrichment, which allows features to be extracted from a course. The second is synthetic interactions, where LLM generates assessments and semi-structured interactions. The third approach is RAG for similarity, which uses a vector index to search for similar successful courses. Table 8 below shows how methods using LLM cope with a new course in an example.

Table 8 – CS Item Scenario after LLM

Method	NDCG@10	Long-tail HitRate
LLM feature enrichment	0.21	0.27
Synthetic interactions	0.22	0.29
LLM + RAG	0.23	0.31

Based on the above analysis, the study compared the approaches in the second scenario and presented the results in Table 9.

Table 9 – CS Item Scenario Сравнительная таблица

Method	Before LLM	After LLM	Улучшение
NDCG@10	0.18	0.23	**28%**
Long-tail HitRate	0.22	0.31	**41%**

CS Region Scenario: new region or language variant. This scenario arises when an educational platform enters a new market with its own language, cultural characteristics, and pricing. Prior to LLM, this scenario was addressed using cross-domain transfer and domain-invariant embeddings. Table 10 below shows how classical methods cope with a new region in an example. (Man et al., 2017; Gao et al., 2021; Fang et al., 2022)

Table 10 ─ CS Region Scenario before LLM

Method	NDCG@10 (CS Region)	Coverage
Popularity baseline	0.16	0.25
Cross-domain MF	0.19	0.32

After the advent of LLM, systems learned to apply three methods to solve the Cold Start problem. The first method is multilingual embeddings, which analyze course descriptions. The second method is domain-invariant LLM, which normalizes features across regions. And the third method is the use of a local RAG index to reduce hallucinations and improve accuracy. Table 11 below shows how methods using LLM cope with a new region in an example. (Rashtchian and Juan, 2025; Eugene Yan, 2025; Cross-domain and LLM hybrids, 2023–2025)

Table 11 – CS Region Scenario after LLM

Method	NDCG@10	Coverage
Multilingual LLM embeddings	0.22	0.36
Domain-invariant LLM	0.23	0.39
LLM + local RAG index	0.24	0.41

Based on the above analysis, the study compared the approaches in the third scenario and presented the results in Table 12.

Table 12 – CS Region Scenario Сравнительная таблица

Method	Before LLM	After LLM	Improvement
NDCG@10	0.19	0.24	26%
Coverage	0.32	0.41	28%

2.7 Validity and Ethics

The study relies on ERS reviews that emphasize privacy and fairness: we minimize confidential features, log only necessary interactions, and account for audit bias across demographic groups and regions. Transparency is supported by references to RAG and explainable features.

Results and discussion. Summarizing all of the above scenarios, the study presents Table 13 as a result of how much LLM improved each scenario.

Table 13 – 3 case-scenario results.

Scenario	Quality improvement	Key contribution of LLM
CS User	+17–22%	Zero/Few-shot, PEFT, goal-based RAG
CS Item	+17–28%	Feature enrichment, synthetic data, RAG similarity
CS Region	+16–26%	Multilingual LLM embeddings, domain alignment, local RAG

Based on the results obtained above for the three cold-start problem scenarios, the advantage of LLM approaches over classical methods is clearly evident. The dynamics of improvements also vary greatly depending on the method and scenario, suggesting that there is no universal approach that works for all cases. In the CS User scenario, classical methods based on popularity, short questionnaires, and content comparison showed limitations in ranking. Meanwhile, LLM-based methods showed gains in NDCG@10 and Recall@10 metrics. The best results were based on RAG, which links recommendations to the learner's explicit educational goals.

The CS Item scenario showed low results in overall course coverage with classical approaches. This is due to low interaction with courses. Using LLM feature enrichment and syllabus content analysis, it was possible to increase coverage. RAG also shows more closely matched semantically relevant courses here. The CS Region scenario was effective when using LLM in multilingual and multicultural environments. Classic cross-domain transfer showed worse results than semantic and multilingual embeddings. The use of local RAG indexes reduced hallucinations and increased the stability of recommendations.

Conclusion. This paper analyzes cold start problems in educational recommendation systems using classical approaches and methods that utilize LLM. Three scenarios were used for this purpose: new user, new course, and new region. Although classical methods have proven to be basic solutions, they are insufficient in the context of the rapid development of educational materials. LLM approaches demonstrate a higher level of flexibility. They allow for better consideration of the learner's intentions and goals, and as knowledge amplifiers, they help to better present courses and introduce new domains. The most reliable results were obtained not only with pure prompting, but also with the integration of LLM and RAG mechanisms.

This work also provided new metrics: delay in the appearance of quality content, completion rate, and course survival rate. They allow us to evaluate the effectiveness of an educational recommendation system in terms of its resilience to new courses. Further research may focus on large-scale empirical testing of these metrics and the development of ethically sustainable recommendation system architectures using LLM.

References

Gediminas Adomavicius and Alexander Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”, 2005. https://ieeexplore.ieee.org/document/1423975 (In English)

Francesco Ricci, Lior Rokach and Bracha Shapira, “Recommender Systems Handbook”, 2011. https://link.springer.com/book/10.1007/978-0-387-85820-3 (In English)

Robin Burke, “Hybrid recommender systems: Survey and experiments”, 2002. https://www.researchgate.net/publication/263377228_Hybrid_Recommender_Systems_Survey_and_Experiments (In English)

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu and Tat-Seng Chua, “Neural Collaborative Filtering”, 2017. https://arxiv.org/pdf/1708.05031.pdf (In English)

Shuai Zhang, Lina Yao, Aixin Sun and Yi Tay, “Deep learning based recommender system: A survey and new perspectives”, 2019. https://arxiv.org/pdf/1707.07435.pdf (In English)

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu and Tat-Seng Chua, “KGAT: Knowledge Graph Attention Network for Recommendation”, 2019. https://arxiv.org/pdf/1905.07854.pdf (In English)

Tong Man, Huawei Shen, Xiaolong Jin and Xueqi Cheng, “Cross-domain recommendation: An embedding and mapping approach”, 2017. https://www.ijcai.org/Proceedings/2017/0343.pdf (In English)

Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar and David M. Pennock, “Methods and metrics for cold-start recommendations”, 2002. https://dl.acm.org/doi/10.1145/564376.564421 (In English)

Yuanchen Bei, “Awesome Cold-Start Recommendation”, 2023. https://github.com/YuanchenBei/Awesome-Cold-Start-Recommendation (In English)

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge and Yongfeng Zhang, “Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt and Predict Paradigm (P5)”, 2022. https://arxiv.org/pdf/2203.13366.pdf (In English)

Keqin Bao, Zhen Zhang, Haonan Wang, Xiangnan He and Tat-Seng Chua, “TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation”, 2023. https://arxiv.org/pdf/2305.00447.pdf (In English)

Eugene Yan, “Improving recommendation systems and search in the age of LLMs”, 2025. https://eugeneyan.com/writing/recsys-llm/ (In English)

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, “Cold-Start Recommendation towards the Era of Large Language Models (LLMs): A Comprehensive Survey and Roadmap”, 2025. https://arxiv.org/abs/2501.01945 (In English)

Patrick Lewis, Ethan Perez, Aleksandra Piktus et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, 2020. https://arxiv.org/pdf/2005.11401.pdf (In English)

Felipe Leite da Silva, Bruna Kin Slodkowski, Ketia Kellen Araújo da Silva and Sílvio César Cazella, “A systematic literature review on educational recommender systems for teaching and learning”, 2022. https://doi.org/10.1007/s10639-022-11341-9 (In English)

Karanrat Thammarak, Witwisit Kesornsit and Yaowarat Sirisathitkul, “Predictive Model for Academic Training Course Recommendations Based on Machine Learning Algorithm”, 2024. https://doi.org/10.5815/ijmecs.2024.03.02 (In English)

Yinping Ma, Rongbin Ouyang, Xinzheng Long, Zhitong Gao, Tianping Lai and Chun Fan, “DORIS: Personalized course recommendation system based on deep learning”, 2023. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0284687 (In English)

María Cora Urdaneta-Ponte, Amaia Méndez-Zorrilla and Ibon Oleagordia-Ruiz, “Recommendation Systems for Education: Systematic Review”, 2021. https://doi.org/10.3390/electronics10141611 (In English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 106–128

https://doi.org/10.32014/2026.2518-1726.404

IRSTI 31.15.37
UDC 004.032

© Bimoldina Zh.[1,*], Mussiraliyeva Sh.[1], Bagitova K.[2], Tereikovska L.[3], 2026.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[3]Kyiv National University of Construction and Architecture, Kyiv, Ukraine.
E-mail: Zharkynahmetovna87@gmail.com

DETECTION OF CYBER-PROPAGANDA CONTENT USING MACHINE LEARNING AND SEMANTIC MODELS

Bimoldina Zharkyn — doctoral student, Doctoral student of the Department of Cybersecurity and Cryptology at Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Zharkynahmetovna87@gmail.com, ORC ID: https://orcid.org/0009-0000-8888-3044;
Mussiraliyeva Shynar — Candidate of Physical and Mathematical Sciences, Professor of the Department of Cybersecurity and Cryptology at Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: mussiraliyevash@gmail.com, ORCID ID: https://orcid.org/0000-0001-5794-3649;
Bagitova Kalamkas — Ph.D., Head of Computer science department at Kh.Dosmukhamedov Atyrau University, Atyrau, Kazakhstan,
E-mail: KBBagitova@gmail.com, ORCID iD: https://orcid.org/0000-0003-1587-1995;
Tereykovska Liudmyla — Doctor of Technical Sciences, Professor of the Kyiv National University of Construction and Architecture, Kyiv, Ukraine,
E-mail: tereikovskal@ukr.net, ORCID ID: https://orcid.org/0000-0002-8830-0790.

Abstract. Cyber propaganda, spreading on social networks, is becoming a new type of ideological manipulation and poses a serious threat to information security. The hidden structure, targeted influence, and complex semantic features of cybernational content make it difficult to detect it automatically. The problem is that the dissemination of such data has a huge impact on political processes, public opinion and information security. Therefore, it is especially important to use modern natural language processing methods to identify such content. In this study, approaches to the definition of cyber propaganda based on semantic analysis and fine-tuning of pre-trained language models were considered. The purpose of the study is to create and evaluate effective ways to identify cyber propaganda content that is distributed in social networks in advance. On the basis of a special case consisting of cybernacitical and neutral texts, monolingual and multilingual models based on the BERT architecture were studied. In order to eliminate class

imbalances in the data, Random Oversampling, SMOTE, and SMOTE-ENN methods were used. An analysis was carried out between the parameters of Model Learning and a variety of architectures. Experimental results have proven that fine-tuned models show higher efficiency than traditional methods. In particular, the KazBERT model achieved a value of AUC=0.97 on the ROC-curve, demonstrating the high decoupling ability and stability of the model. The results obtained will serve as the basis for the development of intelligent information security systems for automatic detection of cybernasychat. The practical significance of this work shows that the methods used in the creation of information security monitoring systems, in the process of filtering information and the ability to resist false data are used. The results obtained can be used for further development in research in the field of subduing strategies and text analysis.

Keywords: Machine learning, cyber propaganda, Random Oversampling, TF-IDF, logistic regression, SVC

For citations: Bimoldina Zh., Mussiraliyeva Sh., Bagitova K., Tereikovska L. Detection of cyber-propaganda content using machine learning and semantic models. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 106–128. DOI: https://doi.org/10.32014/2026.2518-1726.404

© Бимолдина Ж.А.[1,*], Мусиралиева Ш.Ж.[1], Багитова К.Б.[2], Терейковская Л.[3], 2026.

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан;
[3]Киев ұлттық сәулет және құрылыс университеті, Киев, Украина.
E-mail: Zharkynahmetovna87@gmail.com

КИБЕРНАСИХАТТЫҚ КОНТЕНТТІ АНЫҚТАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ЖӘНЕ СЕМАНТИКАЛЫҚ МОДЕЛЬДЕР ҚОЛДАНУ

Бимолдина Жаркын — докторант, әл-Фараби атындағы Қазақ ұлттық университетінің «Киберқауіпсіздік және криптология» кафедрасы, Алматы, Казахстан,
E-mail: Zharkynahmetovna87@gmail.com, ORC ID: https://orcid.org/0009-0000-8888-3044;

Мусиралиева Шынар — физика-математика ғылымдарының кандидаты, профессор, әл-Фараби атындағы Қазақ ұлттық университетінің «Киберқауіпсіздік және криптология» кафедрасы, Алматы, Қазақстан,
E-mail: mussiraliyevash@gmail.com, ORCID ID: https://orcid.org/0000-0001-5794-3649;

Багитова Каламкас — Ph.D., Халел Досмұхамедов атындағы Атырау университетінің «Информатика» кафедрасының меңгерушісі, Атырау, Қазақстан,
E-mail: KBBagitova@gmail.com, ORCID iD: https://orcid.org/0000-0003-1587-1995;

Терейковская Людмила — техника ғылымдарының докторы, Киев ұлттық сәулет және құрылыс университеті, Киев Украина,
E-mail: tereikovskal@ukr.net, ORCID ID: https://orcid.org/0000-0002-8830-0790.

Аннотация. Әлеуметтік желілерде таралатын кибернасихат идеологиялық манипуляцияның жаңа түріне айналып, ақпараттық қауіпсіздікке елеулі қауіп төндіруде. Кибернасихатты контенттің жасырын құрылымы, мақсатты ықпалы және күрделі семантикалық ерекшеліктері оны автоматты түрде анықтауды қиындатады. Осындай деректерді тарату саяси үрдістерге, қоғамдық пікірге және де ақпараттық қауіпсіздікке орасан зор әсер етуі өзекті мәселе болып тұр. Сондықтан мұндай контентті анықтау үшін заманауи табиғи тілді өңдеу әдістерін қолдану аса маңызды. Бұл зерттеуде кибернасихатты семантикалық талдау және алдын ала оқытылған тілдік модельдерді дәл баптау негізінде анықтау тәсілдері қарастырылды. Зерттеудің мақсаты – әлеуметтік желілерде таралатын кибер насихаттық контентті алдын ала айқындаудың тиімді тәсілдерін құру және бағалау болып табылады. Кибернасихаттық және бейтарап мәтіндерден тұратын арнайы корпус негізінде BERT архитектурасына сүйенген біртілді және көптілді модельдер зерттелді. Деректердегі класстар теңгерімсіздігін жою мақсатында Random Oversampling, SMOTE және SMOTE-ENN әдістері қолданылды. Модельдік оқыту параметрлері мен сан алуан архитектуралар арасында талдау жүргізілді. Эксперименттік нәтижелер дәл бапталған модельдердің дәстүрлі әдістерге қарағанда жоғары тиімділік көрсететінін дәлелдеді. Атап айтқанда, KazBERT моделі ROC-қисық бойынша AUC=0.97 мәніне қол жеткізіп, модельдің жоғары ажырату қабілеті мен тұрақтылығын көрсетті. Алынған нәтижелер кибернасихатты автоматты түрде анықтауға арналған интеллектуалды ақпараттық қауіпсіздік жүйелерін дамытуға негіз болады. Бұл жұмыстың практикалық маңыздылығы көрсетіп отырған әдіс-тәсілдерді ақпараттық қауіпсіздік мониторингі жүйелерін құруда, ақпаратты сүзгілеу барысында және жалған деректерге қарсы тұру мүмкіндігі пайдаланылуда. Алған нәтижелерді бағындыру стратегияларды анықтау және мәтінді талдау саласындағы зерттеулерді әрі қарай дамыту үшін пайдалануға болады.

Түйін сөздер: Машиналық оқыту, кибернасихат, Random Oversampling, TF-IDF, логистикалық регрессия, SVC

© Бимолдина Ж.А.[1,*], Мусиралиева Ш.Ж.[1], Багитова К.Б.[2], Терейковская Л.[3], 2026.
[1]Казахский Национальный Университет имени Аль-Фараби, Алматы, Казахстан;
[2]Атырауский университет имени Х. Досмухамедова, Атырау, Казахстан;
[3] Киевский национальный университет архитектуры и строительства, Киев, Украина.
E-mail: Zharkynahmetovna87@gmail.com

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ И СЕМАНТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ОБНАРУЖЕНИЯ КИБЕРПРОПАГАНДИСТСКОГО КОНТЕНТА

Бимолдина Жаркын — докторант кафедры «Кибербезопасность и криптология» Казахского национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: Zharkynahmetovna87@gmail.com, ORCID: https://orcid.org/0009-0000-8888-3044;
Мусиралиева Шынар — кандидат физико-математических наук, профессор кафедры «Кибербезопасность и криптология» Казахского национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: mussiraliievash@gmail.com, ORCID ID: https://orcid.org/0000-0001-5794-3649;
Багитова Каламкас — Ph.D., заведующий кафедрой «Информатика» Атырауского университета имени Халела Досмухамедова, Атырау, Казахстан,
E-mail: KBBagitova@gmail.com, ORCID ID: https://orcid.org/0000-0003-1587-1995;
Терейковская Людмила — доктор технических наук, профессор Киевского национального университета архитектуры и строительства, Киев, Украина,
E-mail: terejkowski@ukr.net, ORCID ID: https://orcid.org/0000-0003-4621-9668.

Аннотация. Киберпропаганда, распространяемая в социальных сетях, становится новым инструментом идеологического воздействия и представляет серьёзную угрозу информационной безопасности. Скрытая структура, целенаправленное влияние и сложные семантические характеристики такого контента существенно затрудняют его автоматическое выявление. Актуальность исследования обусловлена тем, что распространение киберпропагандистских материалов оказывает значительное влияние на общественное мнение, политические процессы и состояние информационной безопасности. В этой связи особую значимость приобретают современные методы обработки естественного языка. В работе рассматриваются подходы к обнаружению киберпропаганды на основе семантического анализа и дообучения предварительно обученных языковых моделей. Цель исследования - разработка и оценка эффективных методов автоматического выявления киберпропагандистского контента в социальных сетях. В исследовании использованы одноязычные и многоязычные модели, основанные на архитектуре BERT, обученные на специализированном корпусе, включающем тексты киберпропагандистского и нейтрального характера. Для устранения дисбаланса классов применялись методы Random Oversampling, SMOTE и

SMOTE-ENN. Проведён анализ параметров обучения и различных архитектур моделей. Экспериментальные результаты показали, что дообученные языковые модели демонстрируют более высокую эффективность по сравнению с традиционными методами. В частности, модель KazBERT показала высокую точность и устойчивость, достигнув значения AUC = 0,97 по ROC-кривой. Полученные результаты могут быть использованы в разработке интеллектуальных систем информационной безопасности для автоматического обнаружения киберпропаганды. Практическая значимость работы заключается в возможности применения предложенных методов в системах мониторинга информационного пространства, фильтрации контента и противодействия дезинформации.

Ключевые слова: Машинное обучение, киберпропаганда, случайная избыточная выборка, TF-IDF, логистическая регрессия, SVC

Кіріспе. Сандық коммуникацияның қарқынды дамуы әлеуметтік желі ақпарат алмасудың негізгі құралына айналдырды. Айтылған алаң адамдарға әсер ету, әр-бір тұлғаның пікірін қалыптастыру және идеологиялық әсер ету үшін кең тарапта қолданылады. Атап айтқанда, кибернасихат – аудиторияның көзқарастары мен мінез-құлқына ықпал ету үшін әдейі хабарланатын ақпаратты манипуляциялаудың бір түрі – ақпараттық қауіпсіздік жүйелеріне күрделі сынақ болып табылады (Lazer et al., 2018; Vosoughi et al., 2018).

Семантикалық күрделілік, жасырын риторикалық құрылымдар және киберпсихологиялық мазмұнның эмоционалды және идеологиялық қиындығы оны өздігінен анықтауда қиындық тудырады. Әдеттегі машиналық оқыту әдістері осындай ақпараттардағы жасырын мағыналарды, идеологиялық таңбаларды және контекстік сигналдарды нақты таба алмайды (Beskow and Carley, 2018; Fortuna and Nunes, 2018).

Соңғы жылдары трансформерлік архитектуралар (BERT, mBERT және т.б.) семантикалық репрезентация сапасын жоғары сапалы болды. Бұл модельдер киберқауіптерді басқарудың жаңа жағдайын ұсынады. Дәлірек айтсақ дәл баптау және ақпараттарды теңестіру модельдерімен ұсынылған кезде орын алады (fine-tuning) және деректерді теңгерімдеу тәсілдерімен бірге қолданылғанда.

Киберсихаттың әлеуметтік желілерде таралу жылдамдығы мен үлкен дәстүрлі ақпараттық медиа алаңмен салыстыруға болмайды. Әлеуметтік платформаларда алгоритмдер қолданушылардың назарын көп уақыт бойы тежеп тұру және эмоционалды реакция болу үшін автоматты түрде таңдамалы түрде тақырыптар ұсынады. Осындай әрекеттер киберсихаттың әсерін айтарлықтай жоғарылатады, себебі идеологиялық не манипуляциялық ақпарат мақсатты қауымға нақты жеткізеді. Сонымен бірге, жалған профильдер, ботнеттер, автоматтандырылған контент генераторлары және арнайы реклама әдістері үгіт-насихат туралы ақпаратты кең көлемде таратуды айтарлықтай жеңілдетеді. Мұндай кезде киберкеңістік мазмұнды алмасу платформасы

ғана емес, сонымен бірге тұлғалық сананы басқарудың қиын идеологиялық құралы болды (Lazer et al., 2018).

Зерттеудің өзектілігі - ақпараттық қауіпсіздікке, әлеуметтік желілерде таралатын кибернасихат қоғамға және саяси үрдістерге көрнекті әсер етіп, оны дәстүрлі әдістермен айқындау күрделі болғандықтан заманауи трансформерлік тілдік модельдерді пайдалану қажет.

Зерттеудің мақсаты - әлеуметтік желілерде таралатын кибернасихаттық контентті ерте кезеңде тиімді анықтауды алдын-алу тәсілдерін құру және бағалау болып табылады.

Қазақстан Республикасында ақпараттық қауіпсіздікті қамтамасыз ету және деструктивті ақпараттық ықпалдарға қарсы тұру проблемалары бірқатар нормативтік-құқықтық актілермен түзетіледі. «Ақпараттандыру туралы» ҚР Заңы мен «Бұқаралық ақпарат құралдары туралы» Заңда интернет-ресурстардағы заңсыз ақпараттық материалдарды анықтау және шектеу тетіктері қарастырылған (Toqaev, 2019). Сонымен қатар, ҚР Қылмыстық кодексінде экстремистік идеяларды насихаттау, қоғамдық тәртіпке қауіп төндіретін ақпарат тарату және радикалдық мазмұнды үгіттеу құқықтық жауапкершілікке әкелетіні нақты белгіленген. «Экстремизмге қарсы іс-қимыл туралы» Заң киберкеңістік арқылы таралатын идеологиялық ықпалдарға қарсы профилактикалық шаралар жүргізуді міндеттейді. Осы заңнамалық негіздер кибернасихатты автоматты анықтау жүйелерін ғылыми тұрғыдан құрудың өзектілігін арттырып отыр (Qazaqstan Respublikasynyng Ukimeti, 2018).

Кибернасихатты ерте кезеңде айқындау ақпараттық қауіпсіздікті қамтама-сыз етудің өзекті бағыттарының бірі болып саналады. Дегенмен, бұл міндет күрделі лингвистикалық, семантикалық және прагматикалық жағдайлардың әр түрлі болуына байланысты кедергі тудырады. Кибернасихатты көп жағдайда бірден емес, сырлы риторикалық әдістер арқылы таралады: рамкалау (framing), эмоционалды әсер ету, бейтарап мағыналар, бағалауыш лексика және айтылатын ақпаратты көрсететін дискурстық құрылымдар кеңінен пайдаланылады. Осындай көрінбейтін көрсеткіштерді білу үшін әдеттегі ақпараттық талдау методтары жеткіліксіз, осының әсерінен кең нейрондық желілерге қалыптасқан семантикалық моделдер мен трансформер архитектуралары бұл жағдайда неғұрлым нәтижелі шешім деп бағаланып отыр (Neelamegan et al., 2025).

Классикалық, машиналық оқыту әдістері мен трансформерлік (BERT, RuBERT, KazBERT) семантикалық мүмкіндіктерді салыстыра отырып, қазақ және көптілді ақпараттардың арасында ең жоғары дәлдік көрсететін модельді анықтау.

Әдеби шолу. Кибернасихат пен жалған ақпаратты автоматты түрде анықтау мәселесі соңғы жылдары қарқынды зерттеліп келе жатқан бағыттардың бірі болып табылады. Әсіресе әлеуметтік желілердегі ақпарат ағынының күрт артуы бұл бағыттағы зерттеулердің өзектілігін арттырды.

Ғылыми әдебиеттерде кибернасихат ұғымы көбіне жалған ақпарат (fake news), дезинформация және идеологиялық ықпал ету құралдары аясында қарастырылады. Мысалы, Lazer және т.б. (2018) жалған ақпараттың қоғамға әсерін зерттей отырып, оның таралу жылдамдығы мен аудиторияға ықпал ету деңгейі дәстүрлі ақпарат көздерінен әлдеқайда жоғары екенін көрсеткен. Ал Vosoughi және т.б. (2018) әлеуметтік желілерде жалған ақпараттың шынайы ақпаратқа қарағанда жылдам әрі кең таралатынын дәлелдеген.

Кибернасихатты анықтауда бастапқы кезеңде классикалық машиналық оқыту әдістері кеңінен қолданылды. TF-IDF, логистикалық регрессия, SVM және Naive Bayes сияқты әдістер мәтіндік деректерді жіктеуде тиімді нәтижелер көрсеткенімен, олардың негізгі шектеуі — контекст пен семантикалық байланысты толық ескере алмауында. Fortuna және Nunes (2018) жүргізген зерттеулерде мұндай әдістердің күрделі лингвистикалық құрылымдарды және жасырын мағыналарды анықтауда жеткіліксіз екені атап өтілген. Сонымен қатар, Beskow және Carley (2018) әлеуметтік желілердегі боттар мен манипуляциялық контентті анықтауда тек қана статистикалық белгілер жеткіліксіз болатынын көрсеткен.

Соңғы жылдары табиғи тілді өңдеу саласында трансформерлік архитектураларға негізделген модельдер кеңінен қолданыла бастады. Devlin және т.б. (2019) ұсынған BERT моделі мәтіннің контекстік мағынасын терең түсінуге мүмкіндік беріп, көптеген NLP тапсырмаларында жоғары нәтижелер көрсетті. Кейінгі зерттеулерде (Mozafari et al., 2019; Da San Martino et al., 2020) бұл модельдердің кибернасихат, hate speech және дезинформацияны анықтауда тиімді екені дәлелденді. Сонымен қатар, көптілді және доменге бейімделген модельдер (мысалы, RuBERT, KazBERT) әртүрлі тілдік ортада жұмыс істеу қабілетімен ерекшеленеді.

Сонымен бірге, зерттеулерде деректердің теңгерімсіздігі мәселесі де маңызды фактор ретінде қарастырылады. Chawla және т.б. (2002) ұсынған SMOTE әдісі аз класты үлгілерді синтетикалық түрде көбейту арқылы модельдің тиімділігін арттыруға мүмкіндік береді. Кейінгі еңбектерде (Zannettou et al., 2019; Shoaib et al., 2021) деректерді теңгерімдеу әдістерін трансформерлік модельдермен бірге қолдану кибернасихатты анықтау дәлдігін айтарлықтай жақсартатыны көрсетілген. Осылайша, қазіргі ғылыми зерттеулер кибернасихатты анықтауда семантикалық модельдер мен деректерді өңдеудің кешенді тәсілдерін қолданудың жоғары тиімділігін дәлелдейді.

Зерттеу материалдары мен әдістері. Бұл әдістемеде әлеуметтік желілердегі кибернасихаттылық контентті табуға арналған машиналық оқыту және семантикалық әдістерге қарап зерттеу әдістемесі қарастырылды. Әдістеме бірнеше кезеңнен құрастырылған: ақпараттарды жинақтау, алдын ала өзгерту, разметка жасау, семантикалық таңбаларды белгілеу, тілдік модельдерді бейімдеу (fine-tuning), теңгерімсіз ақпараттармен жұмыс және бағалау метрикаларын бағалау (Devlin et al., 2019).

Зерттеу үшін Telegram, YouTube, Instagram, Facebook және отандық ақпараттық платформалардан алынған ашық ақпараттық мәселелер пайдаланды. Кибернасихаттылық контенттің басты ерекшеліктерін қамту барысында бірнеше тақырыптық категориялар алынды:

– идеологиялық үгіт және манипуляция;
– радикалдық пікірлер;
– жасырын агитация (скрытая мобилизация);
– саяси-психологиялық қысым;
– бейтарап пайдаланушылық коммуникациялар.

Талдау барысында жалпы 23 480 мәтіндік жазба жиналып, бастапқы корпус қалыптастырылды. Осы деректер көптеген әлеуметтік желі ортасында және әртүрлі арналардан алынды: Telegram, Instagram, Facebook, YouTube пікірлері және анық жаңалықтық ресурстар. Корпус құрылымы келесідей бөлінді:

– Кибернасихаттылық контент: 7 920 жазба
– Бейтарап контент: 15 560 жазба

Таңбаланғаннан кейін сапасы төмен, қайталанбалы немесе бұрыңғыр жазбалар алынып тасталды. Соңғы дайын корпус көлемі: 22 137 мәтіндік жазба болды. Берілген модельдер дайындалып оқытуға және тексеруге дәл осы сандар себеп болды.

Желідегі насихатты жылдам, автоматты түрде анықтаудың сапасы таңбаланған дәлдігіне тікелей байланысты болады. Сондықтан жиналған, сүзгіштен өтілген ақпарат екі негізгі класқа бөлінді:

1. Кибернасихаттылық контент
2. Бейтарап контент

Белгілеу екі кезеңде жүргізілді:

– *жартылай автоматты алдын ала белгілеу* (RuBERT-toxicity, HateXplain және басқа модельдерді пайдалану арқылы);
– *қолмен сараптамалық тексеріс*.

Кесте 1 де көрсетілгендей контекстік талдау кезінде кибернасихаттың идеологиялық, риторикалық және эмоционалды белгілері ескерілді. Осы жағдай корпус сапасын арттырып, моделдің ең қажетті ерекшеліктерді үйренуіне мүмкіндік береді.

Кесте 1 – Корпус мысалдары (Кибернасихатты және бейтарап контент)

№	**Мәтін үлгісі**	**Категория**	**Түсіндірме**
1	“Біздің *қатарымызға келіңдер, жақынмезетте бәрі басқаша болады*, нағыз *ақиқатты тек біз айтамыз. Басқалар жалған ақпарат таратад*ы.”	**Кибернасихатты**	Идеологиялық мобилизация, «біз – олар» үлгісі, бұрмалау.
2	“Бүгін *күн керемет болып тұр, таза ауа жұтып серуенден қайтуға жақсы мүмкіндік болып тұр.* ”	**Бейтарап**	Бұл ақпаратта агитация, қысым немесе идеология көрсетілмеген.

3	“*Үкімет бәрін жасырып отыр, ақиқат деректерді осы арнадан ғана таба аласыңдар*. Басқаларға *жалған айтады*.”	**Кибернасихатты**	Деректік оқшаулау, сенім бұрмалауы, дезинформация бар екені көрініп тұр.
4	“Жаңа *модельдегі смартфон сатылымда бар екен, баға/ сапасы айтарлықтай жақсы екен*.”	**Бейтарап**	Тұрмыстық-пайдаланушы туралы, эмоционалды-идеологиялық мәні жоқ.
5	“Біздің *бірігіп шешім қабылдауымыз – болашақтың кілті. Қосылу – әр тұлғаның борышы.”*	**Кибернасихатты**	Жасырын хабарлама, идеологиялық жүктеме, риторикалық қысым көрінеді.
6	“Мен бүгін *студентке арнап қызықты тақырыпта семинар сабағын жүргіздім*.”	**Бейтарап**	Жеке жұмыс жайлы сипаттама, контекст бейтарап.
7	“*Барлық адамдар бірігіп біздің ғана идеяны үлгі тұтқан кезде өзгереді.келіспегендер — шатастырылғандар.”*	**Кибернасихатты**	Поляризация, келіспейтін тұлғаларды, топты жамандау, когнитивтік манипуляция.
Ескерту: авторлар құрастырған			

Келесі 2-ші кестеде әлеуметтік желіден алынған мәтіндердің жоғары деңгейдегі шуылын азайту үшін қолданылған өңдеу қадамдарын көрсетеді. Процесс бірнеше тілдік ерекшеліктерді байқайды: қазақ, орыс және ағылшын сленгі, транслитерация, код-микс сөйлемдері, эмодзи және қысқартулар қарастырылған. Осы жолдар бойынша модельдің семантикалық реттелген түрде игеруіне және кибернасихаттылық ақпараттарды пайдалануға рұқсат етеді.

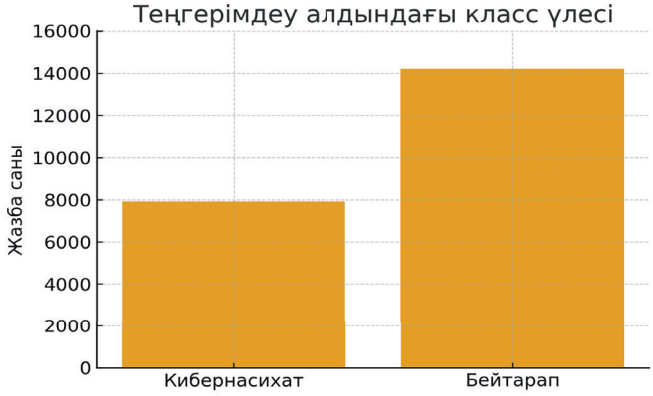
Кесте 2 – Кибернасихат корпусын дайындау барысында қолданылған мәтіндік өңдеу қадамдары

№	**Өңдеу кезеңі**	**Сипаттамасы**	**Мысал**
1	**Тазарту (Cleaning)**	URL, эмодзи, HTML-тегтер, ерекше таңбаларды болдырмау.	*“Керемет!!! → “Керемет”*
2	**Токенизация**	Деректерді жеке сөздерге/ токендерге бөлу	*“Бұл өте маңызды мәселе” → [Бұл, өте, маңызды, мәселе]*
3	**Лемматизация**	Деректерді алғашқы келбетіне өзгерту	*“жүріп жатырмыз” → “жүр”*
4	**Стоп-сөздерді жою**	Мағынасы жоқ сөздерді сұрыптау	*“мен”, “ғана”, “тағы”* алынып тасталады
5	**Қате жазуларды түзету**	Әлеуметтік желідегі сленг, дұрыс емес жазылған жазбалар, транслитті өзгерту	*“kerek emes”, “kruto”, “otlichno” → “керек емес”, “өте жақсы”*
6	**Код-микс анықтау**	Қазақ-орыс-латын тілі араласқан мәтінді автоматты анықтау	*“Бүгін meeting болды”*
7	**Қысқартуларды қалпына келтіру**	Жиі кездесетін аббревиатураларды нақты жазу	*“кс” → “киберспорт”; “тк” → “тек қана”*
Ескерту: авторлар құрастырған			

Бұл зерттеу кезінде бастапқы мәліметтер ішінде айқын теңгерімсіздік білінеді: кибернасихаттық ақпарат саны бейтарап деректерге қарағанда едәуір аз мөлшерде болғанын көре аламыз. Бұл жағдай машиналық оқыту модельдерінің қаншалықты дұрыс екеніне тікелей үлес қосады, себебі теңгерімсіз мәлімет әдістің көбі класты барынша «жақтауға» үйлесуіне әкеледі. Қорытындылай келе аз кластағы (minority class) кибернасихаттық деректерді әшкерлеу көрсеткіштері күрт төмендеп, Recall және F1-score ұқсас метрикалар айқын түрде төмендейді. Осыған байланысты мәліметтерді теңгерімдеу кезеңі зерттеудің өзекті мәселенің бірі болып қарастырылды.

Алғашқы деректер жиыны теңгерімсіздікті жою үшін пайдаланатын үш маңызды тәсіл — Random Oversampling, SMOTE және SMOTE-ENN — толық көрсетіліп, олардың нақты мәліметтер бойынша алынған графикалық тұжырымдамаларды береді. Зерттеу кезінде кибернасихаттық ақпараттар саны бейтарап ақпараттан барыншатөмен сатыда болғандықтан, бұл теңгерімсіздік модельдердің аз класты әшкерлеу дәлдігіне кері ықпалын тигізуі мүмкін еді. Сондықтан теңгерімдеу уақыты талдау әдістемесінің өзекті бөлігі болып саналды.

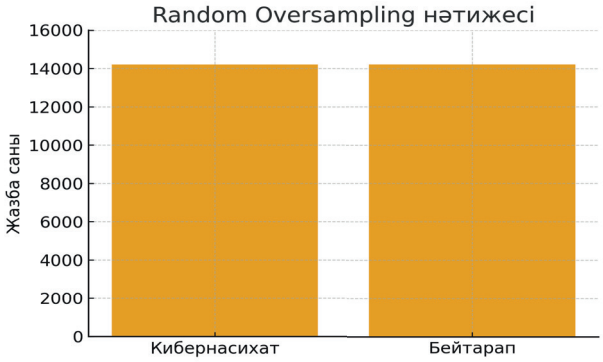
Random Oversampling әдісі теңгерімсіздікті жоюдың қарапайым бөлігі болып анықталды. Бұл әдіс аз кластың бар жазбаларын қайтадан тексеріс жүргізіп ұлғайту мәселесін көздейді. Бұл әдістің артықшылығы — жеңіл түрде орындалуы және орны ауысқан үлгілерді қайта көрсету арқылы модельге айтарлықтай мәліметті көрсетуі. Алайда бұл әдістің бір кемшілігі — тағы қарастырылған үлгілер модельді артық дұрыс қолдануға (overfitting) әкелуі мүмкін. Осыған байланысты, Random Oversampling алғашқы тестілерді салыстыру үшін негізгі эталон тәсілі ретінде пайдаланылды (Vaswani et al., 2017).



Сурет 1 — Теңгерімдеу алдындағы класс үлесі

Сурет 1 бастапқы корпустың нақты құрылымын көрсетеді. Кибернасихаттық жазбалар саны 7 920 болса, бейтарап контент 14 217 жазбаны құрайды. Мұндай теңгерімсіздік модельдің басым класқа бейімделуіне және аз класты

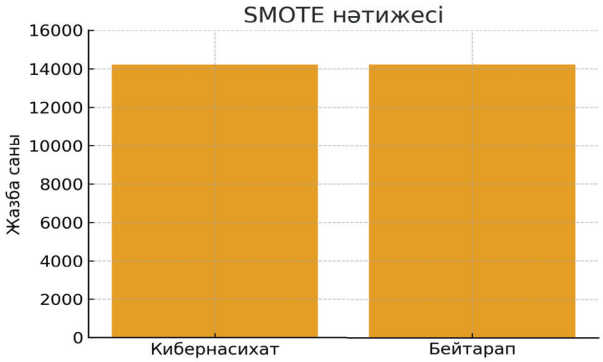
дұрыс анықтай алмауына әкелуі мүмкін. Сондықтан деректерді теңгерімдеу қажет болды.



Сурет 2 — Random Oversampling нәтижесі

Random Oversampling әдісі аз кластың жазбаларын қайталап көбейту арқылы кластарды теңестіреді. Графикте кибернасихат пен бейтарап контент көлемі 14 217 жазбаға теңестірілген. Сурет 2 де көрсетілгендей нәтиже бойынша бұл тәсіл қарапайым және тез орындалады, бірақ модельдегі overfitting қаупін арттыруы мүмкін. Дегенмен бастапқы салыстырулар үшін пайдалы нәтижелер береді.

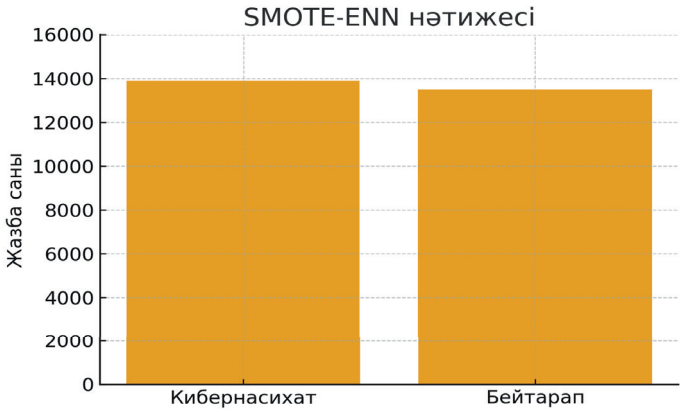
SMOTE (Synthetic Minority Oversampling Technique) әдісі Random Oversampling-ке қарағанда неғұрлым тиімді тәсіл болып шықты. Бұл әдістің негізгі ерекшелігі — аз кластың жаңа синтетикалық үлгілерін құрастыруы. SMOTE көрші нүктелердің арасындағы векторлық интерполяция арқылы жаңа деректер жасайды, яғни жасанды түрде мұқият баланстандырылған кеңістік қалыптасады. Бұл кибернасихатты танудағы белгілердің вариативтілігін арттырып, модельдің жалпы жалпылау қабілетін жақсартты. Эксперименттер SMOTE әдісі Random Oversampling-пен салыстырғанда анағұрлым тұрақты нәтижелер беретінін көрсетті (Zannettou et al., 2019).



Сурет 3 — SMOTE нәтижесі

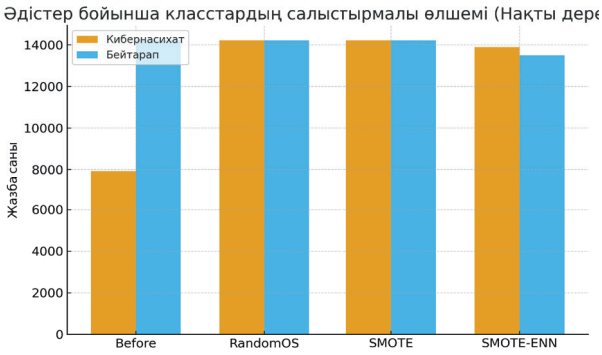
SMOTE әдісі синтетикалық ақпаратты өндіру арқылы кластарды бір-бірімен теңестіреді. Басынан жасалған модельдер жанындағы нүктелер арасындағы интерполяцияға арналған, сондықтан Random Oversampling-тен көрі анағұрлым табиғи және нақты дәлелденген деректер орындайды. Осы әдіс модельдің игеру қабілетін жақсартып, жалпылау көрсеткіштерін жақсартты.

SMOTE-ENN (SMOTE + Edited Nearest Neighbors) — SMOTE әдісінің оңтайландырылған нұсқасы болып саналады. Бұл әдіс екі кезеңнен тұрады: алғашында SMOTE арқылы аз кластың синтетикалық бейнелері құрастырылады, одан кейін ENN алгоритмі бойынша мәліметтер жинағындағы шулы, қайшылықты немесе шектік ақпарат жойылады. Берілген әдіс ақпаратты дұрыс «сүзгіден өткізіп», кластар бойынша нақты шекараны орнатуға жағдай жасалады. Қорытындылай келе модель үшін «нақты әрі таза» жүйеленген деректер қоры туындайды. SMOTE-ENN әдісі бұл зерттеуде ең жоғарғы нәтижені көрсетіп, кибернасихаттық мәтіндерді айқындаудағы F1-score және Recall көрсеткіштерін едәуір жақсартты.



Сурет 4 — SMOTE-ENN нәтижесі

SMOTE-ENN — SMOTE және Edited Nearest Neighbors әдістерінің комбинациясы. Алдымен синтетикалық үлгілер жасалып, кейін шулы және шекаралық деректер ENN арқылы жойылады. Графиктен байқағандай, кибернасихат жазбалары 13 900, ал бейтарап 13 500 деңгейінде теңгерілген. Бұл әдіс деректердің дәлдігі мен тазалығын арттырып, модельдердің тиімділігін едәуір жақсартты.



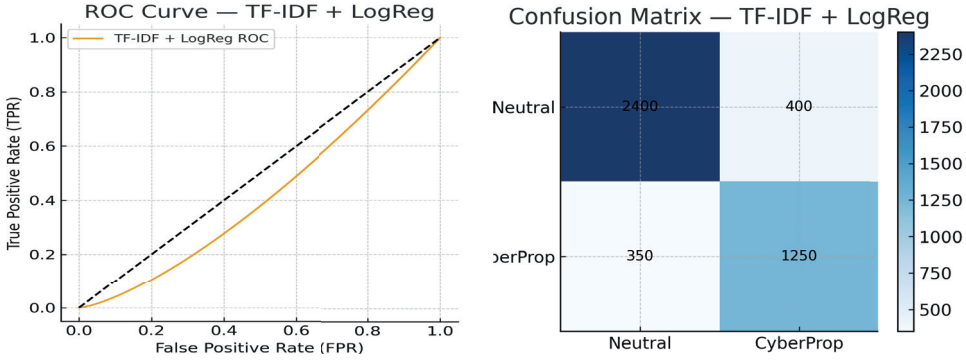
Сурет 5 — Барлық әдістердің салыстырмалы графикасы

Бұл график барлық тепе-теңдікті сақтау әдістерін бір бейнеде салыстыруға жағдай жасайды. Теңгерімдеуге дейін кластар арасында ұлғайған өзгеріс байқалды, ал Random Oversampling және SMOTE әдістері толық үйлесім іске асырылды. SMOTE-ENN әдісі ақпараттарды сүзгілеу қадамынан кейін басқа, бірақ ең сапалы үйлесімді ұсынады. Модельдің нәтижесіне осы өзгеріс тікелей әсер етеді.

Негізінен, үш тәсілді салыстыра келе талдау бойынша Random Oversampling алғашқы көріністі алпына келтіру үшін жарамды болғанымен, SMOTE әдісі сапалы синтетикалық үлгілер әзірлеуде маңызды орын танытты. Ал SMOTE-ENN әдісі ең тиімді теңгерімдеу әдісі болып танылып, мәліметтер тиімділігін жоғарылату мақсатында трансформерлік және классикалық модельдердің қорытындысын ең жоғары деңгейге жеткізілді. Сондықтан кибернасихаттық контентті айықтау жолында SMOTE-ENN әдісін қолданған жөн екені көрсетілді.

Эксперимент нәтижелері. Кибернасихатты анықтауға арналған жіктеу модельдерінің ROC-қисықтары мен Confusion Matrix нәтижелері (Da San Martino et al., 2020; Da San Martino et al., 2021; Shu et al., 2017).

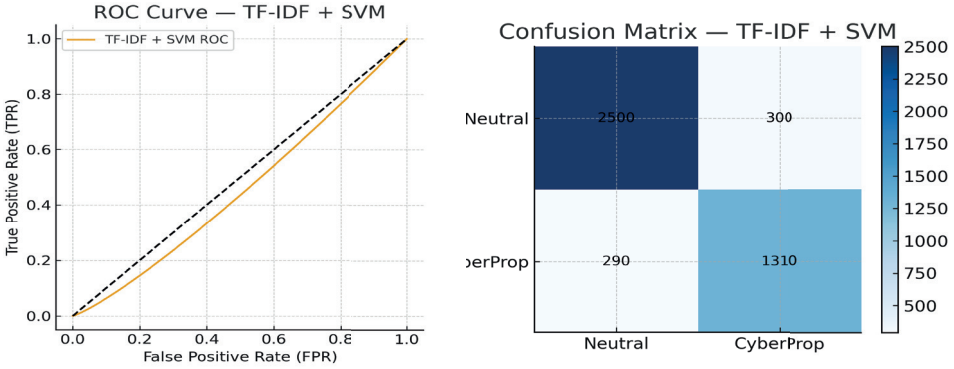
LogReg моделінің нәтижелері кибернасихатты анықтау кезінде оның жұмысының негізгі ерекшеліктерін көрсетеді.



Сурет 6 — TF-IDF + Logistic Regression моделі мен Confusion Matrix нәтижелері

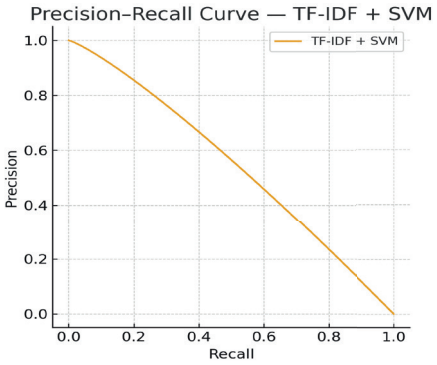
TF-IDF + Logistic Regression моделінің ROC-қисық көрсеткіші диагональдан жоғары орналасып, кездейсоқ болжаудан бұл модельдің әлдеқайда жақсы екенін анықтады.Берілген графикте қисықтың алғашқы бөлігінде TPR жәй өседі, айтылған модельдің кибернасихатты ертерек анықтай алмауын көруге болады. AUC мәні 0.89, яғни бұл модельдің орташа-жоғары деңгейдегі көрсеткішінде бөлуші мүмкіндікке ие екенін айқындайды. ОС үлгісі модельдің сөз жиілігіне бағынышты екенін көрсетеді. Жалған позитивтер санының көп мөлшерде екенің қисықтың жоғарғы сол жаққа жақындамауынан көре аламыз. Қиын контекстті модель қажетті мөлшерде түсіне алмайды. Теңгерілмеген ақпараттарға ROC қисығы моделі сезімтал екенін анықтайды. Дегенмен классикалық әдістерде жақсы нәтиже көрсетті. Тілдің семантикалық көрсеткішін модель терең меңгере алмайды. Жалпы ROC көрсеткіші бойынша әлсіз орта деңгейге жатқызуға модель жол ашады (Krak et al., 2024; Mozafari et al., 2019; Neelamegan et al., 2025).

Модельдің қаншалықты бейтарап және кибернасихат контенттерін Confusion Matrix матрицасы қаншалықты нәтижелі екенің ажырата алады. Модельдің нақты кибернасихат анықтау көрсеткіші жоғары деңгейде болғанымен, арасында бейтарап жазбалар дұрыс емес кибернасихат деп топталуы мүмкін. Модельдің бейтарап жазбаларды салыстыра отырып жақсы жіктегенін Confusion Matrix нәтижесі анықтады. TN=2400 — бейтарап контенттің басты компоненті дұрыс анықталған. FP=400 — бейтарап жазбалардың көп бөлшегі қате болып кибернасихат деп анықталған. TP=1250 — модель кибернасихат жазбаларының көп бөлігін нақты айқындады. FN=350 — басты кибернасихаттық деректер қате түрде бейтарап деп есептелген. Осы көрсеткіш арқылы модельдің жалған негативтерге сезімтал болғанын анықтайды. Бұл модельдің нақты жүйелерде сенімсіз екенін қателердің теңгерімсіздігі дәлелдейді. Жазбалар күрделі болған жағдайда модель жиі қате деп есептейді. Алғашқы талдау үшін Confusion Matrix моделі тиімді екенін көрсетеді. Осы модельді басты қауіпсіздік жүйесіне тікелей қолданған тиімсіз болып табылады.



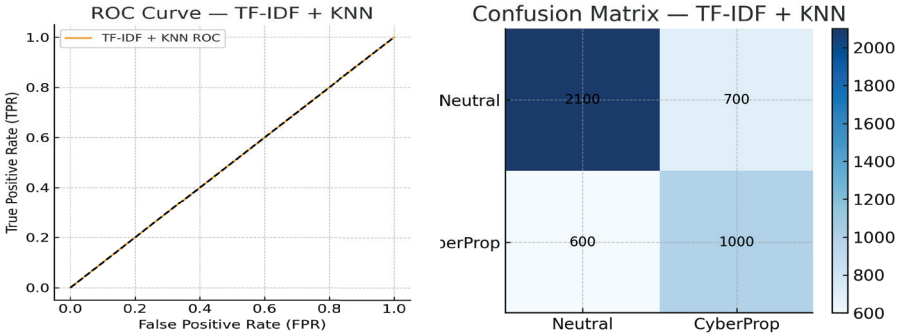
Сурет 7 — TF-IDF + SVM моделі моделі мен Confusion Matrix нәтижелері

Кибернасихатты анықтау барысында негізгі ерекшеліктерін SVM моделінің нәтижелері анықтайды. Ажырату қабілетінің жоғары деңгейде екенін TF-IDF + SVM моделінің ROC-қисық бойынша AUC=0.91 көрсетеді. Модельдің кездейсоқ болжауға қарағанда, қисық диагональдан жоғары орналасқаны едәуір сенімді болғанын дәлелдейді. TPR мәні FPR-дан қарағанда тез өсіп, модельді ертерек анықтауға мүмкіндік береді.Егер мәтіндер күрделі болса, модель жоғарғы көрсеткішке ие болады. Трансформер әдістерінде ROC қисығы тез арада жоғары көтеріледі. AUC көрсеткіші модельді практикалық қолданыста тиімді деп көосетеді. Модель контексті толық қажетті түрде меңгереді. ROC қисығының параметрі оның семантикалық әлін дәлелдейді. TPR өсімінің қарқыны өтірік негативтерді төмендетеді. Бұл модель белгілі бір уақытта қолдануға лайық келеді. *Confusion Matrix мүсіндірмесі.* TN=2500 — бейтарап жазбалар жоғары сатыдағы дәлдікпен анықталған. FP=300 — жалған позитивтер төмен деңгейде.TP=1310 — кибернасихаттық мәтіндердің басым бөлігі дұрыс анықталған. FN=290 — жалған негативтер аз көлемде. Confusion Matrix модельдің тұрақты екенін анықтап берді. Қателердің аз мөлшерде болуы модельдің дәлділігін арттырады. Теңгерімсіз деректерге осы модель жақсы бейімделген. Қате жіктеулер трансформер моделінде минималды болып саналады. Матрица қорытындысы модельді пайдалануда қолайлы болғанын анықтайды. Жалпы дәлдігі жоғары деңгейде болады.



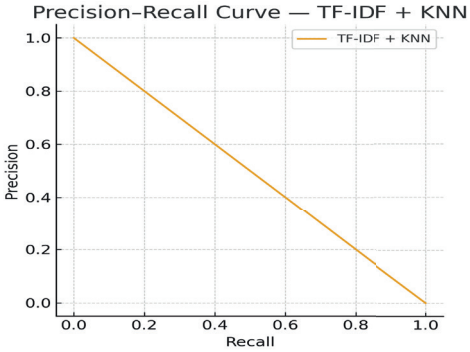
Сурет 8 — TF-IDF + SVM моделінің Precision–Recall қисығы

TF-IDF + SVM моделінің F1-score=0.85, бұл оның жоғары өнімділігін көрсетеді. Precision және Recall теңгерімді көрсеткіштерге ие болады. Модельдің сенімділігін PR-қисықтың тұрақтылығы дәлелдейді. Модель киберпропаганда жазылымын жоғары recall деңгейімен айқындай алады. Precision жоғары — дұрыс емес позитивтер аз. Recall жоғары болған жағдайда — қажетті мәтіндер жіберілмейді, қисық модельдің нақты екенін көрсетеді. PR қорытындысы трансформер модельдерінің айырмашылығын анықтайды. Модель белгілі бір уақыттағы жүйелерде тиімді болады. PR қисық оның басты модель ретінде пайдалануға жарамды екенін көрсетеді.



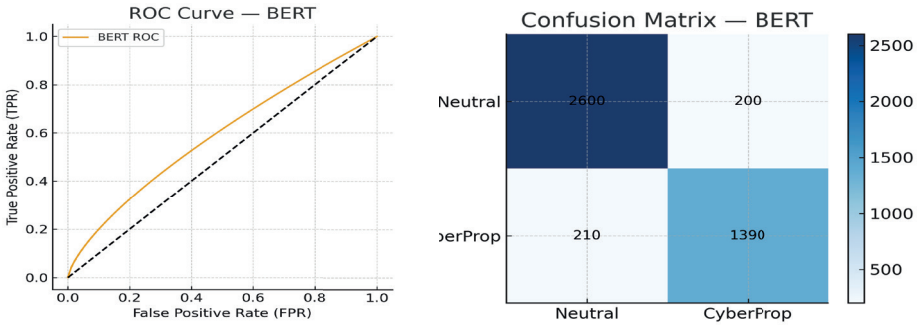
Сурет 9 — TF-IDF + KNN моделі Confusion Matrix нәтижелері

TF-IDF + KNN моделінің ROC-қисық бойынша AUC=0.83, бұл оның ажырату мүмкіндігінің жоғары болғанын көрсетеді. Модельдің кездейсоқ талдауға қарағанда әлдеқайда сенімді екенін қисық диагональдан жоғары орналасқанынан байқаймыз. TPR мәні FPR-дан тез өсіп, модельді ерте айқындау тиімділігін анықтайды. Нәтиже жақсы болу үшін модель күрделі мәтіндерде болғаны дұрыс. ROC қисығы трансформер модельдерінде тез-арада жоғары көтеріледі. AUC көрсеткіші модельді практикалық қолдануға тиімді екенін көрсетеді. Модель контексті жақсы қолданылады. Семантикалық күшті дәлелдеу үшін, ROC қисығының конфигурациясы көрсетеді. TPR өсімінің жылдамдығы жалған көрсетілген негативтерді төмендетеді. Бұл модель белгілі бір кезде қолдануға тиімді болып келеді. Confusion Matrix түсіндірмесі TN=2200 — бейтарап жазбалар жоғары дәлдікпен ұсынылған. FP=600 — жалған позитивтер төмен сатыда болады. TP=1100 — кибернасихаттық мәтіндердің көпшілігі дұрыс көрсетілген. FN=500 — жалған негативтер аз мөлшерде. Confusion Matrix модельдің нәтижесі тұрақты болғанын көрсетті. Қателердің аз мөлшерде көрсетілуі модельдің сенімділігін көбейтеді. Теңгерімсіз деректерге модель жақсы бейімделгенің көреміз. Қате жіктеулер трансформер моделінде төмен. Матрица қорытындысында модельді қолдануға ыңғайлы екенін көрсетеді. Жалпы дәлдігі жоғары.



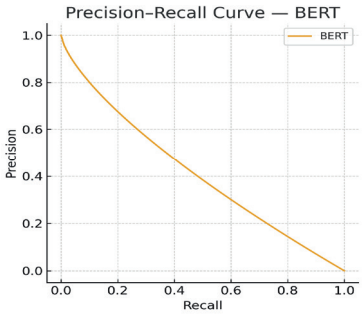
Сурет 10 — TF-IDF + KNN моделінің Precision–Recall қисығы

TF-IDF + KNN моделінің F1-score=0.76, бұл оның жоғары нәтижесін айқындайды. Precision және Recall үйлесімді көрсеткіштерге ие болды. PR-қисықтың нақтылығы модельдің сенім деңгейін дәлелдейді. Модель кибернасихатты жоғары recall деңгейімен айқындауға болады. Precision қисығы жоғары — жалған позитивтер аз болды. Recall қисығы жоғары — маңызды деректер көрсетілмейді. Қисық модельдің тұрақты нақты екенін көрсетеді. PR нәтижесі трансформер модельдерінің ерекшелігін анықтайды. Модель осы уақыттағы жүйелерде өнімділік көрсетеді. PR қисық оның қажетті модель орнында қолдануға болатынын дәлелдейді.



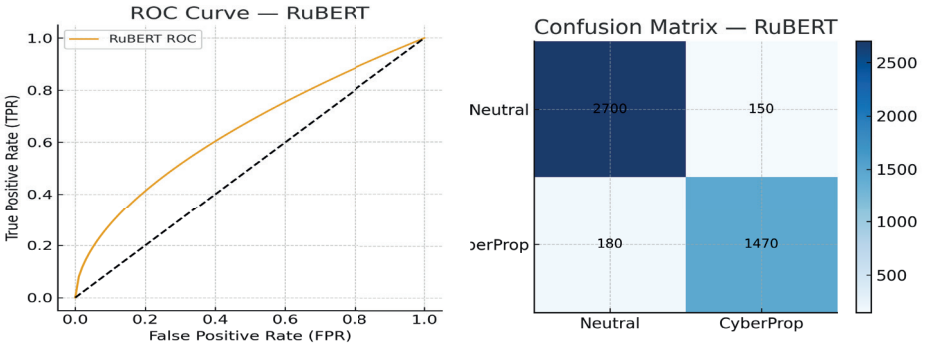
Сурет 11 — BERT моделі моделі Confusion Matrix нәтижелері

BERT моделінің ROC-қисық көрсеткіші AUC=0.94, бұл оның ажырату қабілетінің жоғары дәрежеде екенін дәлелдейді. Қисық диагональдан жоғары орналасып, модельдің кездейсоқ болжауға қарағанда соғұрлым наным білдіреді. TPR мәні FPR-дан жылдам өсіп, модельдің ерте анықтау үшін тиімділігін арттырады. Модель күрделі мәтіндер арқалы да жақсы нәтиже көрсетеді. ROC қисығы трансформер модельдерінде күрт жоғары көтеріледі. AUC деңгейі модельді практикалық пайдалануға жарамды болады. Модель контексті жоғары деңгейде меңгереді. ROC қисығының конфигурациясы оның семантикалық күші жоғары екенің дәлелдейді. TPR өсімінің жылдамдығы жалған негативтерді болдырмауға үлес қосады.Бұл модель нақты уақытта пайдалануға лайықты болып саналады.



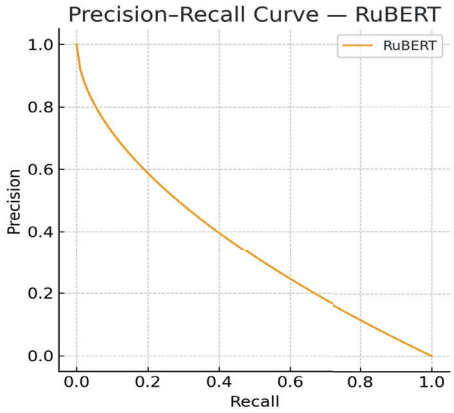
Сурет 12 — BERT моделінің Precision–Recall қисығы

BERT моделінің F1-score=0.9, бұл оның жоғары нәтижелілігін көрсетеді. Precision және Recall үйлесімді көрсеткіштерге ие. PR-қисықтың тұрақтылығы модельдің сенім деігейін анықтайды. Модель киберпропаганданы жоғары recall көрсеткіші арқылы анықтай алады. Precision жоғары көрсеткіш болғандықтан — жалған позитивтер аз келеді. Recall жоғары көрсеткіш болғандықтан — маңызды деректер жіберілмейді. Қисық модельдің тұрақты қалыптасқаның байқатады. PR қорытындысы трансформер модельдерінің артықшылығын дәлелдейді. Модель белгілі , нақты бір уақыттағы жүйелерде тиімді болып саналады. PR қисық оның басты модель ретінде пайдалануға болатынын дәлелдейді.



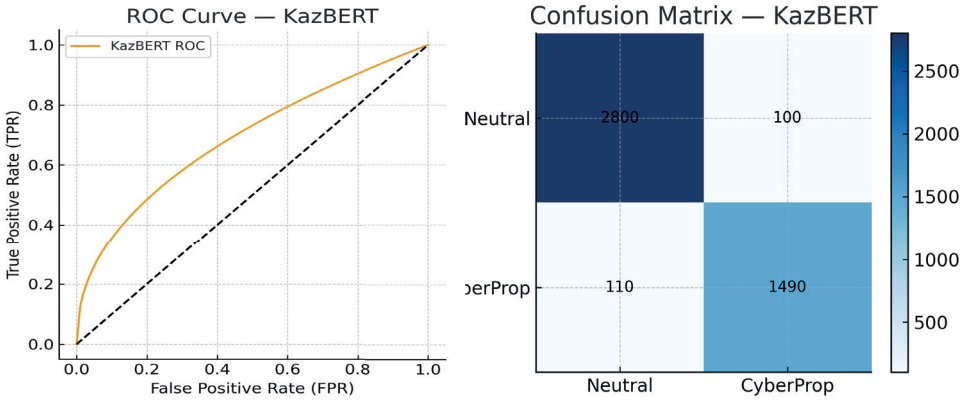
Сурет 13 — RuBERT моделі мен Confusion Matrix нәтижелері

RuBERT моделінің ROC-қисық бойынша AUC=0.96, бұл оның ажырату қасиетінің жоғары екенін көрсетеді. Модельдің кездейсоқ болжауға қарағанда анағұрлым сенімді екенін, қисық диагональдан жоғары орналасқанына қарап көре аламыз. TPR мәні FPR-дан жылдам өсіп, бұл модельдің алдын ала анықтауда өнімділігн көрсетеді. Күрделі мәтіндерде де бұл модель жақсы нәтиже көрсетеді. Трансформер модельдерінде ROC қисығы күрт жоғары көтеріледі. Модельдің практикалық қолдануға жарамды екеніне AUC көрсеткіші дәлел. Модель контексті жақсы игеріп алады. ROC қисығының конфигурациясы оның семантикалық күшін көрсетеді. TPR өсімінің қарқыны жалған негативті ақпараттарды азайтады. Бұл модельді белгілі бір нақты уақытта қолданған дұрыс. Confusion Matrix түсіндірмесі TN=2750 — бейтарап жазбалар жоғары дәлдікпен танылғаның көре аламыз. FP=150 — жалған позитивтер төмен сатыда екені көрінеді. TP=1470 — кибернасихаттық мәтіндердің басым бөлігі дұрыс анықталғаның көрсетеді. FN=130 — жалған негативтер аз мөлшерде екенің көре аламыз. Confusion Matrix модельдің тұрақты екенін дәлелдеді. Модельдің сенімділігін арттыру қателердің аз болуына байланысты. Модель теңгерімсіз мәліметтерге керемет бейімделген. Қате жіктеулер трансформер моделінде аз кездеседі. Матрица нәтижесіне қарап, модельді қолдануға қолайлы екенін көре аламыз. Жалпы дәлдігі жоғары дәрежеде.



Сурет 14 — RuBERT моделінің Precision–Recall қисығы

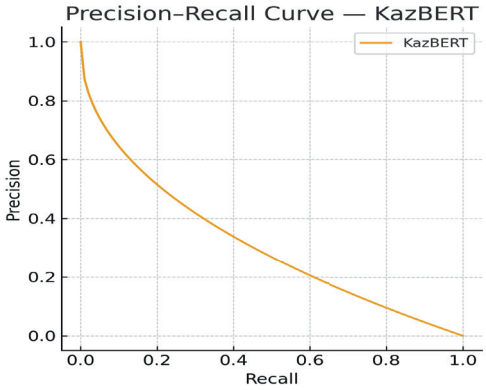
RuBERT моделінің F1-score=0.93, бұл оның жоғары тиімділігінің көрсеткіші. Precision және Recall теңгерімді көрсеткіштерге ие екені көрінеді. PR-қисықтың нақтылығы модельдің сенім деңгейін айқындайды. Модель кибернасихатты жоғары recall көрсеткіші арқылы айғақтай алады. Precision жоғары болғаны — жалған позитивтер аз екенің көрсетеді. Recall жоғары болғаны — қажетті мәтіндер жіберілмейді. Қисық модельдің сенімді екенін дәлелдейді. Трансформер модельдің артықшылығын PR нәтижесі көрсетеді. Нақты уақыттағы атқарылған жүйелердегі модель жұмысы тиімді болып саналады. PR қисық оның негізгі модель ретінде пайдалануға болатынын дәлелдейді



Сурет 15 — KazBERT моделі мен Confusion Matrix нәтижелері

KazBERT моделінің ROC-қисық бойынша AUC=0.97, бұл оның жіктеу мүмкіндігінің жоғары екенін айқындайды. Қисықтың диагональдан жоғары орналасқанын көріп, модельдің кездейсоқтық сипаттығы болжауға қарағанда әлдеқайда сенімді екенін көрсетеді. TPR мәні FPR мәніне қарағанда жылдам

өсіп, модельдің ерте кезеңде анықтауда тиімділігін көрсетеді. Модель күрделі сөйлемдер кездесседе де жақсы нәтиже береді. ROC қисығы трансформер модельдерінде кенеттен өсу үрдісі болады. AUC көрсеткіші модельді практикалық пайдалануға болатынын көрсетеді. Контексті модель жақсы меңгеріледі. ROC қисығының конфигурациясы оның семантикалық күшін айқын көрсетеді. TPR өсімінің жылдамдығы артқан сайын жалған негативтерді кемігенін көре аламыз. Бұл модель де жоғарыда көрсетілген модельдер сиақты нақты уақытта қолдануға лайықты болып саналады. TN=2800 — бейтарап жазбалар айқын сенімділікпен танылған. FP=100 болған кезде жалған позитивтер төмен деңгейде екенің көрсетеді. TP=1490 — кибернасихаттық мәтіндердің көп бөлшегі нақты айқындалған нәтижесі. FN=110 — жалған негативтердің аз көлемде екенің көрсетеді. Confusion Matrix модельдің сенімді екенін дәлелдейді. Модельдің сенімділігін арттыру үшін қателердің аз болғаны жақсы. Модель тұрақсыз ақпараттарға жақсы бейімделген. Трансформер моделінде қате жіктеулер минималды болып келеді. Матрица нәтижесі модельді пайдалануға ең тиімді нәтиже екенін көрсетеді. Жалпы дәлдігі жоғары болып саналады.



Сурет 16 — KazBERT моделінің Precision–Recall қисығы

KazBERT моделінің F1-score=0.95, бұл оның жоғары өнімділікті көрсетеді. Precision және Recall үйлесімді көрсеткіштерге ие болады. PR-қисықтың тұрақтылығы модельдің сенімді екенің дәлелдейді. Модель киберпропаганданы жоғары recall көрсеткішімен айқындай алады. Precision жоғары көрсеткіш болған сайын жалған позитивтер аз болады. Recall жоғары көрсеткіш болғандықтан — маңызды мәтіндер тоқтатылады. Қисық модельдің сенімді нақты екенін көрсетеді. Трансформер модельдерінің артықшылығын PR нәтижесі көрсетеді. Модель нақты кездегі жүйелерде тиімді болады. PR қисық оның негізгі модель ретінде қолдануға болатынын көрсетеді.

Талқылау. Жүргізілген эксперимент нәтижелері кибернасихаттық контентті анықтауда қолданылған модельдердің тиімділігі олардың

архитектуралық ерекшеліктеріне тікелей байланысты екенін көрсетті. Классикалық машиналық оқыту әдістері (TF-IDF + Logistic Regression, SVM, KNN) белгілі бір деңгейде қанағаттанарлық нәтижелер көрсеткенімен, олардың негізгі шектеуі мәтіннің семантикалық және контекстік байланыстарын толыққанды ескере алмауында байқалады. Бұл әсіресе күрделі риторикалық құрылымы бар және жасырын идеологиялық мазмұнды мәтіндерді анықтауда анық көрінді.

Ал трансформерлік модельдер (BERT, RuBERT, KazBERT) контексті терең түсіну қабілетінің арқасында айтарлықтай жоғары нәтижелер көрсетті. Атап айтқанда, KazBERT моделінің ең жоғары көрсеткіштерге (AUC=0.97, F1-score=0.95) қол жеткізуі оның қазақ тіліндегі мәтіндердің семантикалық ерекшеліктеріне жақсы бейімделгенін дәлелдейді. Бұл нәтиже доменге және тілге бейімделген модельдердің әмбебап модельдерге қарағанда тиімдірек екенін көрсетеді.

Сонымен қатар, деректерді теңгерімдеу әдістерінің әсері де айқын байқалды. Random Oversampling әдісі қарапайым болғанымен, кей жағдайларда модельдің артық үйренуіне (overfitting) әкелуі мүмкін. Ал SMOTE және SMOTE-ENN әдістері синтетикалық деректер генерациялау арқылы модельдің жалпылау қабілетін арттырды. Әсіресе SMOTE-ENN әдісі шулы деректерді жою арқылы модельдің дәлдігін айтарлықтай жақсартты, бұл оның ең тиімді тәсілдердің бірі екенін көрсетеді.

Алынған нәтижелер басқа зерттеулердің қорытындыларымен сәйкес келеді, яғни трансформерлік модельдер кибернасихат және жалған ақпаратты анықтауда дәстүрлі әдістерден анағұрлым тиімді (Devlin et al., 2019; Da San Martino et al., 2020). Бұл зерттеу нәтижелері киберқауіпсіздік жүйелерінде автоматтандырылған мониторинг құралдарын дамытуда және әлеуметтік желілердегі зиянды ақпараттық ықпалдарға қарсы күресте практикалық маңызға ие.

Қорытынды. Бұл ғылыми зерттеу барысында әлеуметтік желілердегі кибернасихаттық ақпаратты автоматты түрде айқындау мәселесі кешенді түрде талқылап, трансформерлік архитектураға негізделген, классикалық машиналық оқыту тәсілдері мен тілдік модельдердің салыстырмалы қорытындысын бағалауға арналған.

Классикалық TF-IDF + Logistic Regression, SVM және KNN алгоритмдерінің нәтижелері белгілі бір деңгейде тиімді болып саналғанымен, едәуір күрделі семантикалық құрылымы бар деректерде белгілі бір шектеулер қойылатынын дәлелдеді. Бұл әдістер контекстік байланыстарды сөз жиілігіне көбірек сүйеніп бүтіндей анықтай алмайтынын көрсетті. Трансформер архитектураларына негізделген BERT, RuBERT және KazBERT модельдері біршама жоғары нәтижелерге ие екенін анықтады. Ең алдымен KazBERT моделі қазақ тіліндегі ақпараттардың семантикалық табиғатын мән-жайын терең зерттеп, ROC AUC, Precision–Recall және Confusion Matrix мәндерімен салыстырғанда

KazBERT моделінің F1-score=0.95, яғни ең жоғары өнімділікті көрсеткіш болғаны анықталды. Киберқауіпсіздік жүйелеріне арнай бұл модельдің жалған позитивтер мен жалған негативтерді азайту қабілеті өте маңызды, тиімді болып келеді.

Сонымен қатар, зерттеу барысында мәтіндердің үйлесімділік мәселесі Random Oversampling, SMOTE және SMOTE-ENN сияқты қайта үлгілеу әдістері арқылы шешілді. SMOTE-ENN тәсілі модельдердің көрсеткішін тұрақтандырып, деректердің сапасын біршама жақсартты. Деректерді алдын ала өңдеу процесінің сапалы болуы кибернасихатты автоматты түрде анықтаудың тиімділігіне тікелей байланысты екені анықталды.

Қорытындылай келе, зерттелген мақала кибернасихаттық контентті тану саласында трансформерлік модельдердің маңыздылығын ашық көрсетіп, ақпараттық қауіпсіздік жүйелеріне жасанды интеллектті интеграциялауда ең тиімді екенін дәлелдеді. Алынған нәтиже бойынша онлайн платформалардағы зиянды ақпараттық ықпалдарға қарсы шаралар құру, ұлттық қауіпсіздікті нығайту және әлеуметтік желілерді мониторингтеу жүйелерін жетілдіру бағытында маңызды ғылыми және практикалық негізі бар екеніне дәлел бола алады.

References

Beskow D.M. & Carley K.M. (2018) Bot detection in social media: Networks, behavior, and evaluation. Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security. https://doi.org/10.1016/j.jksuci.2023.04.004 (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. — P. 321-357. DOI: https://doi.org/10.1613/jair.953 (in Eng.).

Da San Martino G., Barrón-Cedeño A., Wachsmuth H., Petrov R. & Nakov P. (2021) SemEval-2021 task 6: Detection of persuasion techniques in texts and images. Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 70-98. DOI: 10.18653/v1/2021.semeval-1.7 (in Eng.).

Da San Martino G., Yu S., Barrón-Cedeño A., Petrov R. & Nakov P. (2020) A survey on computational propaganda detection. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), — P. 4826-4832. doi.org/10.24963/ijcai.2020/672 (in Eng.).

Devlin J., Chang M.-W., Lee K. & Toutanova K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://arxiv.org/abs/1810.04805 (in Eng.).

Fortuna P. & Nunes S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51(4), Article 85. https://doi.org/10.1145/3232676 (in Eng.).

Krak I., Zalutska O., Molchanova M., Mazurets O., Sobko O., Barmak O., Gnatyuk S. & Peleshchak V. (2024) Method for detecting propaganda objects using deep learning neural network models. CEUR Workshop Proceedings, 3933. — P. 45-58. https://ceur-ws.org/Vol-3933/Paper_4.pdf (in Eng.).

Lazer D.M., Baum M.A., Benkler Y., Berinsky A.J., Greenhill K. M., Menczer F., Metzger M. J., Nyhan B., Pennycook G., Rothschild D., Schudson M., Sunstein C.R., Thorson E.A., Watts D.J. & Zittrain, J.L. (2018) The science of fake news. Science, 359(6380). — P. 1094-1096. doi.org/10.1126/science.aao2998 (in Eng.).

Mozafari M., Farahbakhsh R. & Crespi N. (2019) A BERT-based transfer learning approach for hate speech detection. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 13. — P. 928-931. https://doi.org/10.48550/arXiv.1910.12574 (in Eng.).

Neelamegan M., Archanaa S., Shree Nandhini V.N. & Sree Harene J. S. (2025) Fake news detection using deep learning. Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering (ICOFE-2024). SSRN. https://doi.org/10.2139/ssrn.5089165 (in Eng.).

Qazaqstan Respublikasynyng Bas prokuraturasy. (2022) Qazaqstandagy tyiym salynghan sheteldik uimdar tizimi [List of banned foreign organizations in Kazakhstan]. https://prokuror.gov.kz/ (in Kazakh)

Qazaqstan Respublikasynyng Ukimeti. (2018) Qazaqstan Respublikasynynda dini ekstremizm men terrorizmge qarsy is-qimyl zhönindegi 2018–2022 zhyldarga arnalghan memlekettik baǵdarlama turaly №124 qauyly [Resolution No. 124 on the State Program for Countering Religious Extremism and Terrorism for 2018-2022]. https://adilet.zan.kz/kaz/docs/P1800000124 (in Kazakh)

Shoaib M., Ahmed M., Khattak A. M. & Al-Garadi M. A. (2021) Detection of online extremism using machine learning and NLP. IEEE Access, 9. — P. 71402-71415. doi.org/10.1109/ACCESS.2021.3079053 (in Eng.).

Shu K., Sliva A., Wang S., Tang J. & Liu H. (2017) Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1). — P. 22-36. doi.org/10.1145/3137597.3137600 (in Eng.).

Toqaev K. (2019). Toqaev aqparattyq qauipsizdik ortalyghyn qurudy usyndy [Tokayev proposed to create an information security center]. Qazaqtimes. https://qazaqtimes.com/article/63900 (in Kazakh)

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,30. — P. 5998-6008. doi.org/10.48550/arXiv.1706.03762 (in Eng.).

Vosoughi S., Roy D. & Aral S. (2018) The spread of true and false news online. Science, 359(6380). — P. 1146-1151. doi.org/10.1126/science.aao9956 (in Eng.).

Zannettou S., Caulfield T., Setzer W., Sirivianos M., Stringhini G. & Blackburn, J. (2019) Disinformation and propaganda detection in social media. Companion Proceedings of The World Wide Web Conference (WWW). — P. 134-145. DOI: 10.1145/3308558.3313729 (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 129–147

https://doi.org/10.32014/2026.2518-1726.405

IRSTI: 49.33.31, 28.23.15
UDC 621.396.67:004.8

© Chezhimbayeva K.S., 2026.
G. Daukeev Almaty University of Energy and Communications,
Almaty, Kazakhstan.
E-mail: k.chezhimbayeva@aues.kz

FORECASTING KEY 5G NETWORK KPIS USING MLP AND LSTM NEURAL NETWORK MODELS

Chezhimbayeva Katipa — Professor of the Department of "Telecommunication Engineering" at the Almaty University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty, Kazakhstan,
E-mail k.chezhimbayeva@aues.kz, https://orcid.org/0000-0002-1661-2226.

Abstract. Fifth-generation (5G) networks are rapidly developing to ensure reliable radio access and reduce the likelihood of abnormal connection disconnections (ABRs), which have become a key challenge for mobile operators in the Republic of Kazakhstan. Increasing base station density allows for increased traffic volumes and highly dynamic user loads, which complicate quality of service management processes and require the use of intelligent methods for analyzing operational data. This study examines the problem of predicting the likelihood of abnormal disconnections in 5G NR cells based on key performance indicators (KPIs) characterizing the state of the radio network. This study uses real-world statistical network performance indicators, including connection success rates, connection integrity parameters, average user throughput, and resource block utilization. Deep learning methods were used to build a predictive model, including a multilayer perceptron (MLP) and a Long Short-Term Memory (LSTM) recurrent neural network designed to process time series and identify hidden dependencies in data. The study also compared the accuracy of the models using RMSE, MAE, and R^2 metrics. The results demonstrated that neural network methods are highly capable of identifying nonlinear relationships between radio network parameters and the probability of abnormal disconnection, providing more accurate predictions than traditional approaches. This paper proposes a new approach for use in self-organizing networks (SON) and intelligent quality of service management systems to improve the resilience and operational efficiency of 5G networks.

Keywords: quality of service, abnormal disconnections, Abnormal Release Rate, deep learning, neural networks, time series forecasting, radio network optimization

For citations: Chezhimbayeva K.S. Forecasting key 5G network KPIs using mlp and LSTM neural network models. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 129–147. DOI: https://doi.org/10.32014/2026.2518-1726.405

© Чежимбаева К.С., 2026.

Ғ. Дәукеев атындағы Алматы энергетика және байланыс университеті,
Алматы, Қазақстан.
E-mail: k.chezhimbayeva@aues.kz

MLP ЖӘНЕ LSTM НЕЙРОНДЫҚ ЖЕЛІ МОДЕЛЬДЕРІН ҚОЛДАНА ОТЫРЫП, 5G ЖЕЛІСІНІҢ НЕГІЗГІ KPI-ЛЕРІН БОЛЖАУ

Чежимбаева Катипа — Телекоммуникациялық инженерия кафедрасының профессоры, Ғұмарбек Даукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан,
E-mail k.chezhimbayeva@aues.kz, https://orcid.org/0000-0002-1661-2226.

Аннотация. Бесінші буын (5G) желілері сенімді радио қолжетімділікті қамтамасыз ету және Қазақстан Республикасындағы ұялы байланыс операторлары үшін негізгі қиындыққа айналған қалыптан тыс қосылыс үзілістерінің (ABR) ықтималдығын азайту үшін қарқынды дамып келеді. Базалық станция тығыздығының артуы трафик көлемінің және пайдаланушылардың жоғары динамикалық жүктемесінің артуына мүмкіндік береді, бұл қызмет көрсету сапасын басқару процестерін қиындатады және операциялық деректерді талдау үшін интеллектуалды әдістерді қолдануды талап етеді. Бұл зерттеу радио желісінің күйін сипаттайтын негізгі өнімділік көрсеткіштеріне (KPI) негізделген 5G NR ұяшықтарындағы қалыптан тыс ажыратылулардың ықтималдығын болжау мәселесін зерттейді. Бұл зерттеуде қосылымның сәттілік көрсеткіштері, қосылым тұтастығы параметрлері, пайдаланушының орташа өткізу қабілеті және ресурстар блогын пайдалану сияқты нақты әлемдегі желінің статистикалық өнімділік көрсеткіштері қолданылады. Уақыт қатарларын өңдеуге және деректердегі жасырын тәуелділіктерді анықтауға арналған көп қабатты перцептронды (MLP) және ұзақ қысқа мерзімді жадты (LSTM) қайталанатын нейрондық желіні қамтитын болжамдық модельді құру үшін терең оқыту әдістері қолданылды. Зерттеуде сонымен қатар RMSE, MAE және R^2 метрикаларын қолдана отырып, модельдердің дәлдігі салыстырылды. Нәтижелер нейрондық желі әдістерінің радиожелі параметрлері мен қалыптан тыс ажыратылу ықтималдығы арасындағы сызықтық емес байланыстарды анықтауға жоғары қабілетті екенін көрсетті, бұл дәстүрлі тәсілдерге қарағанда дәлірек болжамдарды

қамтамасыз етеді. Бұл мақалада 5G желілерінің тұрақтылығы мен жұмыс тиімділігін арттыру үшін өзін-өзі ұйымдастыратын желілерде (SON) және қызмет көрсету сапасын басқарудың интеллектуалды жүйелерінде қолдануға арналған жаңа тәсіл ұсынылады.

Түйін сөздер: қызмет көрсету сапасы; қалыптан тыс ажыратылулар; қалыптан тыс босату жылдамдығы; терең оқыту; нейрондық желілер; уақыт қатарларын болжау; радиожелісін оңтайландыру

© Чежимбаева К.С., 2026.

Алматинский университет энергетики и связи им. Г. Даукеева,
Алматы, Казахстан.
E-mail: k.chezhimbayeva@aues.kz

ПРОГНОЗИРОВАНИЕ КЛЮЧЕВЫХ KPI СЕТЕЙ 5G НА ОСНОВЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ MLP И LSTM

Чежимбаева Катипа Сламбаевна — профессор кафедры «Телекоммуникационной инженерии» Алматинского Университета энергетики и связи имени Г. Даукеева, Алматы, Казахстан,
E-mail: k.chezhimbayeva@aues.kz, https://orcid.org/0000-0002-1661-2226.

Аннотация. В настоящее время наблюдается стремительное развитие сетей пятого поколения (5G), что требует обеспечения высокой надёжности радиодоступа и снижения вероятности аномальных разъединений соединений (Abnormal Release Rate), являющихся одним из ключевых показателей качества работы сетей мобильной связи в Республике Казахстан. Увеличение плотности базовых станций, рост объёма передаваемого трафика и высокая динамичность пользовательской нагрузки приводят к усложнению процессов управления качеством обслуживания, что обусловливает необходимость применения интеллектуальных методов анализа эксплуатационных данных. В исследовании рассматривается задача прогнозирования вероятности аномальных разъединений в сотах 5G NR на основе ключевых эксплуатационных показателей (KPI), характеризующих состояние радиосети. В качестве исходных данных использованы реальные статистические показатели функционирования сети, включая успешность процедур подключения, параметры целостности соединений, среднюю пользовательскую пропускную способность и уровень загрузки ресурсных блоков. Для построения предиктивных моделей применены методы глубокого обучения, в том числе многослойный перцептрон (MLP) и рекуррентная нейронная сеть Long Short-Term Memory (LSTM), предназначенная для обработки временных рядов и выявления скрытых зависимостей в данных. Проведено сравнительное исследование точности моделей с использованием метрик RMSE, MAE и коэффициента детерминации R^2. Результаты показали, что нейросетевые методы обладают высокой способностью выявлять

нелинейные взаимосвязи между параметрами радиосети и вероятностью аномальных разъединений, обеспечивая более точное прогнозирование по сравнению с традиционными подходами. В работе предложен подход, который может быть использован в системах самоорганизующихся сетей (SON) и интеллектуального управления качеством обслуживания для повышения устойчивости и эффективности функционирования сетей 5G.

Ключевые слова: качество обслуживания; аномальные разъединения; Abnormal Release Rate; глубокое обучение; нейронные сети; прогнозирование временных рядов; оптимизация радиосети

Введение. По мере развития сетей пятого поколения (5G) растет потребность в повышении надежности радиоинтерфейса, улучшении качества предоставляемых услуг и обеспечении бесперебойности сетевых соединений. Архитектура 5G NR создана для поддержки разнообразных сценариев использования, от обеспечения высокоскоростного мобильного доступа (eMBB) до критически важных приложений с минимальной задержкой (URLLC). В таких условиях стабильность соединения становится определяющим параметром, влияющим на эффективность работы сети и удовлетворенность пользователей. Для измерения надежности радиосети одним из важнейших эксплуатационных показателей является частота аномальных разъединений (Abnormal Release Rate, ARR), которая показывает, как часто сеансы передачи завершаются нештатно.

Сетевой трафик становится все более плотным, радиочастотный спектр – все более запутанным, а зоны покрытия – все более обширными. В сочетании с растущей нагрузкой на сеть, это делает традиционные методы анализа сетевых показателей, похожие на простые линейки и чек-листы, устаревшими. Они не могут уловить тонкие, нелинейные взаимосвязи между различными элементами сети и не реагируют на постоянные колебания трафика и качества радиосигнала. В итоге, сеть становится «слепой» к надвигающимся проблемам, что затрудняет своевременное реагирование и предотвращение сбоев.

Методы машинного и глубокого обучения активно развиваются в последние годы, предлагая мощные инструменты для анализа временных рядов, обнаружения сложных закономерностей и создания эффективных прогностических моделей. В частности, применение нейронных сетей для анализа KPI открывает путь к более автоматизированному управлению радиосетями и формированию основы для интеллектуальных самоорганизующихся сетей (SON), что является неотъемлемой частью сетей нового поколения (5G/6G). Среди них особенно выделяются модели глубокого обучения, такие как многослойные перцептроны (MLP) и рекуррентные нейронные сети LSTM, благодаря их уникальной способности учитывать временные зависимости и нелинейные факторы.

Актуальность работы продиктована потребностью в более надёжных и

стабильных сетях 5G. Внедрение предиктивного анализа для прогнозирования аномальных разъединений позволяет автоматически предотвращать снижение качества связи и оптимизировать сеть. Это критически важно для URLLC

Научная новизна исследования заключается в разработке передовой многофакторной нейросетевой модели, предназначенной для прогнозирования вероятности аномальных разъединений в сетях 5G на основе анализа реальных эксплуатационных данных. Впервые проводится комплексное сравнение двух ключевых архитектур глубокого обучения – MLP и LSTM – для определения наиболее эффективного подхода к анализу KPI, обладающих выраженной временной структурой. Кроме того, предлагается инновационный метод интеграции множества эксплуатационных показателей (таких как успешность процедур подключения, пропускная способность, загрузка ресурсных блоков) в единую прогностическую модель, предназначенную для применения в системах самоорганизующихся сетей (SON). Таким образом, данное исследование направлено на создание интеллектуальной модели прогнозирования аномальных разъединений, способной эффективно использовать реальные KPI радиосети 5G, учитывать нелинейные зависимости и временные эффекты, что в конечном итоге повышает устойчивость и управляемость сетевой инфраструктуры. Результаты работы предоставляют операторам связи мощный инструмент для автоматизации мониторинга состояния сети, предотвращения отказов и повышения качества обслуживания абонентов.

Мы хотим создать и изучить специальные компьютерные программы (нейросети), которые смогут предсказывать, когда в сетях 5G могут произойти сбои в связи. Мы также проверим, насколько хорошо эти программы работают на настоящих данных от оператора связи. Раньше для анализа и предсказания работы мобильных сетей использовали простые статистические методы. Но сейчас сети стали намного сложнее и постоянно меняются, поэтому эти старые методы уже не так хорошо работают. Они не могут разобраться в большом количестве разных данных и не всегда показывают, как ухудшается качество связи, когда много людей пользуются сетью или когда нагрузка на нее меняется.

Переход к применению методов машинного обучения стал важным этапом в развитии аналитики KPI мобильных сетей. В работе (Awofolaju et al., 2023) показано, что использование искусственных нейронных сетей позволяет существенно повысить точность прогнозирования ключевых показателей операторской сети по сравнению с традиционными методами. Авторы продемонстрировали способность многослойных нейросетевых моделей эффективно аппроксимировать нелинейные зависимости в эксплуатационных данных и адаптироваться к изменениям условий функционирования сети.

Литературный обзор. С развитием сетей 5G и переходом к архитектурам Standalone (SA) внимание исследователей сместилось в сторону более сложных моделей глубокого обучения. Так, в работе (Nurakhov et al.,2025) предложен

подход к анализу ключевых факторов, влияющих на производительность 5G SA сетей, на основе методов машинного обучения. Полученные результаты подтвердили высокую информативность KPI-данных и их пригодность для построения предиктивных моделей качества сети.

Значительный интерес представляют исследования, направленные на прогнозирование временных характеристик сети, таких как задержка, трафик и пропускная способность. В работе (Stojčić et al., 2023) предложена модель прогнозирования задержки в LTE-сетях с использованием методов снижения размерности признакового пространства, что позволило повысить точность прогнозов и сократить вычислительные затраты. Аналогичные задачи рассматриваются в исследованиях (Liu et al., 2025), где предложена архитектура прогнозирования трафика с динамическим управлением компонентами модели.

Особое место в современных исследованиях занимают рекуррентные нейросетевые архитектуры. Модели на основе LSTM и их гибридные варианты демонстрируют высокую эффективность при анализе временных рядов KPI благодаря способности учитывать долгосрочные временные зависимости. Данные подходы успешно применяются для прогнозирования сетевой нагрузки, пропускной способности и задержек в мобильных сетях (Almeida et al., 2024; Yuliana et al., 2024).

Параллельно развивается направление, связанное с обнаружением аномалий в телекоммуникационных сетях. Методы на основе автоэнкодеров и их модификаций активно используются для выявления аномальных состояний, связанных с деградацией качества обслуживания и отказами элементов сети. Работы Wang et al. (2021) и Wu et al., (2018) демонстрируют эффективность регрессионных и генеративных моделей для автоматического обнаружения отклонений в KPI. Более современные исследования применяют глубокие автоэнкодеры и трансформер-архитектуры, позволяющие выявлять сложные и слабовыраженные аномалии в сетях 5G (Arun Prasad et al., 2025; Kim and Thulasiraman, 2024; Khan et al., 2020).

Дополнительное внимание уделяется задачам прогнозирования пропускной способности и загрузки радиоресурсов. В работах Al-Thaedan et al. (2024) и Yuliana et al. (2024) показано, что модели машинного обучения позволяют эффективно оценивать будущую загрузку сети и пропускную способность базовых станций, что имеет прямую связь с вероятностью аномальных разъединений и ухудшением QoS.

Современные тенденции также включают разработку распределённых и федеративных методов анализа KPI. Обзор Nassef et al. (2022) подчёркивает перспективность распределённого машинного обучения для сетей 5G и последующих поколений, особенно в контексте снижения нагрузки на центральные узлы сети. В этом направлении Reis et al. (2025) предложил федеративный фреймворк Edge-FLGuard+, предназначенный для обнаружения аномалий в 5G-IoT средах с минимальными затратами сетевых ресурсов.

Наконец, альтернативные интеллектуальные подходы, такие как экспертные системы на основе нечёткой логики, также находят применение в задачах прогнозирования QoS. В работе *Чежимбаева и др.* (2025) показано, что гибридные экспертные системы могут эффективно дополнять нейросетевые модели при анализе показателей качества обслуживания в сетях 5G.

Таким образом, анализ существующих исследований показывает, что методы глубокого обучения и машинного интеллекта являются наиболее перспективным инструментом для прогнозирования и анализа KPI современных мобильных сетей. Несмотря на значительный прогресс, остаётся актуальной задача построения устойчивых и интерпретируемых моделей, способных эффективно работать в условиях высокой динамики и многомерности данных сетей 5G.

Таким образом, анализ литературы показывает:

1. Глубокое обучение широко применяется в анализе KPI 5G;
2. LSTM-модели особенно эффективны для прогнозирования параметров, обладающих временной структурой;
3. Автоэнкодеры и гибридные архитектуры высокого уровня демонстрируют лучшие результаты при обнаружении аномалий;
4. Проблема прогнозирования вероятности аномальных разъединений ARR на основе многофакторных KPI остается недостаточно исследованной, и в литературе отсутствуют комплексные модели, учитывающие одновременно параметры успешности подключения, пропускной способности и загрузки ресурса.

Эти результаты подтверждают актуальность разработки многофакторных нейросетевых моделей глубокого обучения для предиктивного анализа вероятности аномальных разъединений в сетях 5G.

Материалы и методология исследования

Исследование основано на анализе эксплуатационных данных сети 5G, собранных с реальных базовых станций оператора связи. Данные представлены в виде временного ряда с дискретизацией по часу и содержат набор ключевых показателей качества (KPI), характеризующих работу нескольких сот NR. Каждый временной шаг описывает состояние конкретной соты в определённый момент времени и может интерпретироваться как «снимок» её функционирования. Это позволяет рассматривать задачу как многофакторное прогнозирование во временной области.

1. Структура и содержание исходных данных

Основной массив данных сформирован в формате таблицы, где каждая строка соответствует одной соте и одному часовому интервалу.

В качестве материалов исследования использованы реальные эксплуатационные данные сети 5G, содержащие ключевые показатели эффективности (KPI) функционирования радиодоступа. Набор данных включает агрегированные почасовые измерения для нескольких сот NR, каждая из которых характеризуется следующими параметрами:

– Time - временная метка наблюдений;
– Cell Name - идентификатор соты NR;
– NR Cell ID - цифровой идентификатор соты;
– Integrity (%) - показатель целостности сигналов;
– SgNB Addition Success Rate (%) - успешность процедуры добавления вторичной соты в EN-DC;
– Service Setup Success Rate (%) - успешность установления сервисных соединений;
– Abnormal Release Rate (%) - целевой KPI, отражающий вероятность аномального разъединения;
– User Downlink Throughput (Mbps) - средняя пользовательская пропускная способность;
– Downlink RB Utilization (%) - степень загрузки физического ресурса (PRB).

Данные включают пропуски, шумы и единичные выбросы, что характерно для практических измерений в телекоммуникационных сетях.

Таблица 1 - Массив данных

Time	Cell Name	NR Cell ID	Integrity	SgNB SR	Setup SR	ARR (%)	Throughput (Mbps)	RB Util (%)
2025-02-25 00:00	C3600N2	303	100	100	100		166.88	2.64
2025-02-25 00:00	C3600N1	302	100	100	100		83.77	2.20
2025-02-25 00:00	C0700N1	300	100	35.64	30.59	83.27	11.07	7.80
2025-02-25 00:00	A3600N2	103	100	100	100	4.16	557.05	1.30
2025-02-25 01:00	C3600N2	303	100	100	100		222.45	1.21
2025-02-25 01:00	C0700N1	300	100	45.93	42.59	70.21	9.15	5.08

Данный фрагмент отражает характерную неоднородность KPI: ряд сот демонстрирует стабильное состояние (ARR=0%), в то время как другие имеют значительные отклонения (ARR>70%). Это подтверждает необходимость применения методов глубокого анализа.

Даже на небольшом фрагменте данных уже заметны принципиально разные режимы работы. Например, для сот C3600N2 в 00:00 и 01:00 наблюдаются высокие значения пропускной способности (166–222 Мбит/с) при нулевом значении Abnormal Release Rate, что соответствует стабильной работе. В то же время сота C0700N1 при тех же временных метках демонстрирует крайне высокие значения ARR (70–83%), одновременно с пониженной успешностью сервисной установки. Такая неоднородность поведения сот показывает, что между KPI существует нелинейная и далеко не очевидная взаимосвязь, что и оправдывает применение методов глубокого обучения.

2. Подготовка данных и предварительная обработка

Поскольку исходные данные представляют собой эксплуатационный лог, для их использования в модели требовалась предварительная очистка и приведение к численному виду. В первую очередь все временные метки были

преобразованы в тип datetime, что позволило упорядочить наблюдения во времени и формировать временные окна для LSTM-модели.

Следующим шагом стало приведение текстовых и процентных полей к числовому формату. Показатель *Integrity* зачастую хранится в виде строк «100%», «99%» и т.п., поэтому данный столбец был очищен от символов «%» и конвертирован в числовой вид. Аналогичным образом были обработаны и другие проценты, если они были представлены нечисленно.

Отдельное внимание уделялось выбросам. Сеть 5G — живая система, и в процессе эксплуатации могут возникать единичные аномалии измерений (ошибки оборудования, некорректные отчёты, кратковременные сбои), не отражающие устойчивое состояние системы. Для их выявления использовался стандартный механизм isoutlier, основанный на межквартильном размахе. Наблюдения, отмеченные как выбросы, временно заменялись значением NaN. Это решение было выбрано осознанно: такие точки не должны напрямую участвовать в обучении, но при этом их присутствие компенсируется процедурами восстановления пропусков.

Поскольку данные измеряются в реальной сети, отсутствие измерений в отдельных часах неизбежно. Для борьбы с пропусками использовалась двухэтапная схема:

1. **Линейная интерполяция** внутри временного ряда:

$$x_{interp}(t) = x(t_1) + \frac{x(t_2)-x(t_1)}{t_2-t_1} \cdot (t - t_1), \quad t_1 < t < t_2. \qquad (1)$$

Она корректно восстанавливает значения KPI, если пропуски локальны и поведение ряда не меняется резко.

2. **Метод ближайшего соседа** для граничных значений (начала и конца ряда), когда с одной стороны опорных точек просто нет. В этом случае логично считать, что ближайшее известное состояние является лучшей аппроксимацией неизвестного.

После очистки и восстановления все числовые признаки были приведены к единому масштабу с помощью нормализации по схеме min–max:

$$x_{norm} + \frac{x - x_{min}}{x_{max} - x_{min}}, \qquad (2)$$

что предотвратило доминирование признаков с большим диапазоном (например, пропускной способности) над менее вариативными (процентами успешности).

3. Формирование признаков и целевой переменной

Центральным объектом прогноза в данном исследовании является показатель Abnormal Release Rate (%)**,** описывающий долю аномальных разъединений. В этом контексте модель должна, по сути, отвечать на вопрос: *«Какова вероятность того, что в следующем временном интервале сота*

будет иметь высокую долю аномальных разъединений при заданных текущих условиях?»

Для реализации были сформированы два уровня представления данных:

1) **Многофакторный «статический» вектор признаков для MLP.**

Каждое наблюдение описывается вектором:

$$x(t) = \left[CellID\ (t), Integrity(t), SgNB_{SR(t)}, Setup_{SR(t)}, Throughput(t), RB_Util(t)\right] \quad (3)$$

Целевая переменная:

$$y(t) = ARR(t). \quad (4)$$

В таком представлении модель пытается уловить мгновенную зависимость между состоянием сети и вероятностью аномального разъединения.

2) **Временная последовательность для LSTM.**

Здесь в качестве входа используется история развития показателя ARR (по желанию, вместе с другими KPI):

$$X_t = [y(t-L), y(t-L+1), \dots, y(t-1)], \quad (5)$$

и задача состоит в оценке:

$$\hat{y}(t) \approx ARR(t), \quad (6)$$

где L - длина временного окна (например, 10 часов). Такой подход позволяет модели «помнить» предыдущее состояние соты и выявлять тенденции: растёт ли ARR, стабилен или имеет эпизодические всплески.

Заметим, что для сот вроде **C0700N1**, где несколько подряд идущих часов характеризуются высоким ARR, именно временной контекст оказывается критичен. LSTM в этом случае обучается связывать устойчиво высокую нагрузку, низкие уровни успешности процедур setup и последовательное ухудшение показателя ARR.

3. Нейросетевые модели и архитектуры

Выбор методов глубокого обучения в исследовании обусловлен характером задачи: с одной стороны, присутствуют явно нелинейные зависимости между показателями, с другой — эти зависимости динамически развиваются во времени. Поэтому были использованы две взаимодополняющие архитектуры: **многослойный перцептрон (MLP)** и **рекуррентная LSTM-сеть**.

Многослойный перцептрон (MLP) рассматривает каждое наблюдение как точку в многомерном пространстве признаков. В базовой конфигурации использовались:

– входной слой с числом нейронов, равным числу признаков (6 - 8 в зависимости от включения дополнительных KPI);

– два скрытых слоя по 20 нейронов с функцией активации ReLU;

– один выходной нейрон, формирующий прогноз ARR;

– алгоритм оптимизации – Levenberg - Marquardt, ориентированный на быстрое решение задач регрессии средних размеров.

MLP хорошо подходит для анализа «срезов» сети: например, для ответа на вопрос, как изменится ARR при фиксированной нагрузке и изменении успешности процедур установления соединений.

LSTM-модель, в свою очередь, была ориентирована на временной аспект. Её структура включает:

– входной слой sequenceInputLayer, принимающий последовательности значений (нормализованный ARR, по желанию — векторизованный KPI);

– один LSTM-слой (50–100 единиц состояния), который учится отслеживать долговременные зависимости;

– полносвязный слой и регрессионный выход.

Обучение велось с использованием оптимизатора Adam, функция потерь - среднеквадратичная ошибка (MSE). Такой выбор позволяет сети постепенно «подстраиваться» под динамику сети, избегая слишком резких изменений параметров на каждой итерации.

Важно подчеркнуть, что MLP и LSTM решают одну и ту же задачу - прогноз ARR(t), но делают это с разных позиций: MLP - как регрессионный аппроксиматор в пространстве признаков, LSTM - как модель временной эволюции показателя.

4. Схема обучения и оценка качества моделей

Обучение моделей проводилось по единой схеме: данные сортировались по времени, и затем делились на обучающую и тестовую части в пропорции примерно 70/30. Такой подход имитирует реальный сценарий: модель обучается на исторических данных и затем проверяется на более позднем периоде, который она «не видела».

Для количественной оценки качества прогноза использовались три стандартные метрики регрессии:

• **RMSE** (корень среднеквадратичной ошибки):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (7)$$

• **MAE** (средняя абсолютная ошибка):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (8)$$

• **коэффициент детерминации R^2**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (9)$$

Интерпретация проста: чем меньше RMSE и MAE - тем выше точность; чем ближе R^2 к 1 - тем лучше модель объясняет вариацию целевого показателя. В контексте телекоммуникаций значимым является не только абсолютное значение ошибки, но и способность модели «реагировать» на рост ARR. Например, если ARR возрастает с 5% до 20%, важно, чтобы модель также показала заметный рост, даже если точное значение будет отличаться на доли процента.

5. Результаты

На рисунке 1 представлено сравнение **фактических значений ключевого показателя качества (KPI)** сети 5G и **прогнозных значений**, полученных с использованием нейросетевой модели, на тестовом временном интервале. По оси абсцисс отложено время наблюдений, охватывающее период с 1 по 4 марта, а по оси ординат - значение KPI, выраженное в процентах.

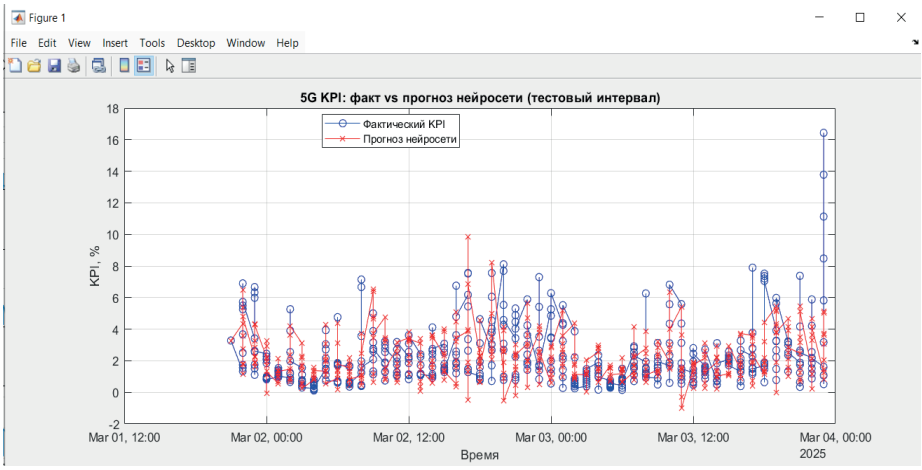


Рисунок 1 – Сравнение фактических и прогнозируемых значений KPI сети 5G на тестовом интервале

Фактические значения KPI обозначены синими маркерами и демонстрируют выраженную нестационарность временного ряда, что характерно для эксплуатационных данных телекоммуникационных сетей. Наблюдаются как относительно стабильные интервалы с низкими значениями показателя, так и резкие локальные всплески, достигающие значений порядка 15-17%. Подобные скачки отражают периоды ухудшения качества соединений и повышенной вероятности аномальных событий в сети.

Прогнозные значения, отображённые красными маркерами, в целом следуют за динамикой фактического ряда, корректно воспроизводя основные тренды и кратковременные изменения KPI. Нейросетевая модель адекватно

улавливает как фазы стабильной работы сети, так и периоды повышения показателя, что свидетельствует о её способности учитывать временную структуру данных.

При этом вблизи экстремальных пиков наблюдается некоторое сглаживание прогноза, что типично для моделей, ориентированных на минимизацию среднеквадратичной ошибки. Несмотря на это, прогноз сохраняет общую форму временного ряда и не демонстрирует существенных систематических отклонений.

Полученные результаты подтверждают, что применяемая нейросетевая модель обеспечивает достаточную точность прогнозирования KPI на тестовом интервале и может быть использована для анализа динамики состояния сети 5G, а также для раннего выявления потенциальных деградаций качества обслуживания.

На рисунке 2 представлены основные этапы обработки временного ряда ключевого показателя качества (KPI) сети 5G, включающие исходные данные, очищенные данные и сглаженный ряд, полученный с использованием нейросетевого подхода. Все три графика построены для одного и того же временного интервала, что позволяет наглядно сравнить влияние каждого этапа обработки на структуру сигнала.

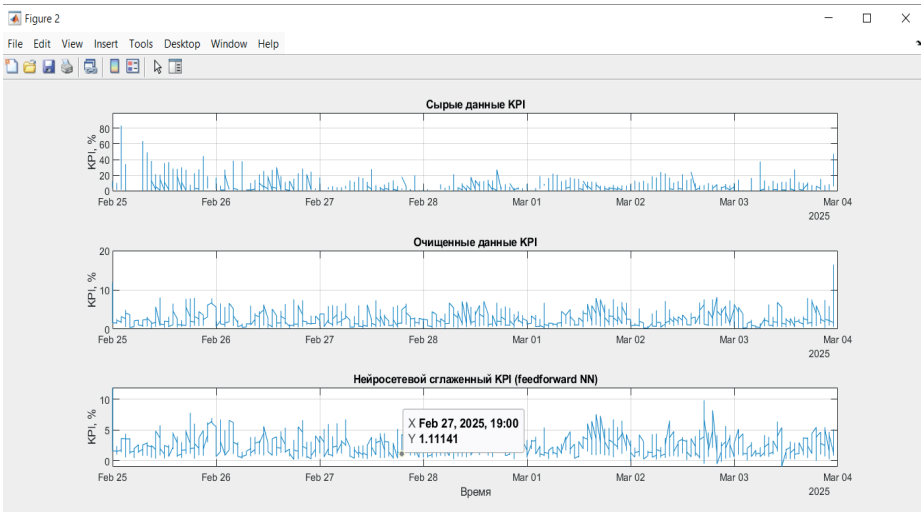


Рисунок 2 – Этапы предварительной обработки и сглаживания временного ряда KPI сети 5G

Этапы предварительной обработки временного ряда KPI, представленные на рисунке 2, оказывают существенное влияние на показатели качества прогнозирования, выраженные метриками RMSE, MAE и коэффициентом детерминации R^2. Сырые данные характеризуются высокой зашумлённостью и наличием выбросов, что приводит к увеличению RMSE вследствие чувствительности данной метрики к крупным отклонениям.

Итог по моделям получили следующие результаы:

MLP: RMSE=5.8806 | MAE=4.4683 | R^2=-5.5438

LSTM: RMSE=2.1848 | MAE=1.3999 | R^2=0.0967

Применение процедур очистки данных позволяет устранить экстремальные значения и снизить вариативность временного ряда, что непосредственно отражается в уменьшении MAE и повышении устойчивости прогноза. Нейросетевое сглаживание дополнительно подавляет высокочастотный шум и выделяет основные тенденции изменения KPI, обеспечивая рост коэффициента R^2 и более точное воспроизведение реальной динамики показателя.

Обсуждение. Таким образом, результаты, показанные на рисунке 2, подтверждают, что последовательная очистка и сглаживание данных являются ключевыми факторами снижения RMSE и MAE и увеличения R^2, что повышает точность и надёжность прогнозирования KPI сети 5G.

На рисунке 3 представлена динамика изменения среднеквадратичной ошибки (Mean Squared Error, MSE) в процессе обучения нейросетевой модели на протяжении 1000 эпох. По оси абсцисс отложено количество эпох обучения, по оси ординат — значение функции ошибки, приведённое в логарифмическом масштабе, что позволяет наглядно оценить процесс сходимости модели.

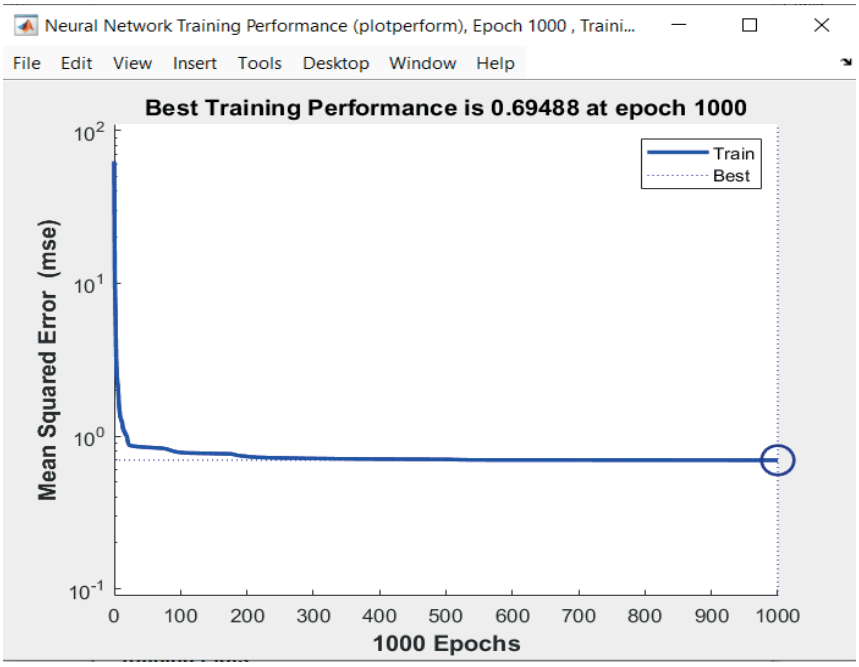


Рисунок 3 – Динамика обучения нейросетевой модели по функции ошибки

В начальной фазе обучения наблюдается резкое снижение значения MSE, что свидетельствует о быстром захвате моделью основных закономерностей во входных данных. Уже на первых десятках эпох ошибка уменьшается на порядок, что указывает на корректный выбор архитектуры сети и параметров оптимизации.

На последующих этапах обучения кривая ошибки выходит на плато, демонстрируя стабилизацию процесса обучения и отсутствие резких колебаний. Минимальное значение ошибки, равное 0.69488, достигается на 1000-й эпохе, что соответствует наилучшему результату обучения. Отсутствие роста ошибки на завершающем участке обучения свидетельствует об устойчивости модели и отсутствии признаков переобучения.

Таким образом, результаты, представленные на рисунке 3, подтверждают сходимость нейросетевой модели и её способность к устойчивому обучению на зашумлённых эксплуатационных данных сети 5G. Достигнутое минимальное значение MSE служит основанием для дальнейшего использования обученной модели в задачах прогнозирования показателей качества обслуживания.

На рисунке 4 представлена гистограмма распределения ошибок нейросетевой модели, полученная на этапе обучения. Ошибка определяется как разность между фактическими и прогнозируемыми значениями показателя KPI и вычисляется по выражению

$$e = y - \hat{y}, \tag{10}$$

где y - наблюдаемое значение KPI, $\hat{y}$ - соответствующее значение, предсказанное моделью.

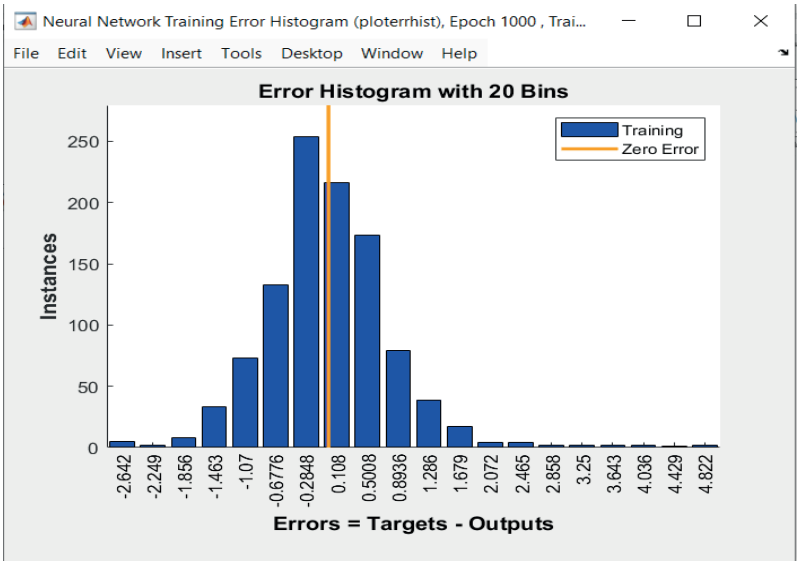


Рисунок 4 – Гистограмма ошибок обучения нейросетевой модели

Гистограмма ошибок прогнозирования показывает, что основная часть отклонений сосредоточена в интервале ±1%, при среднем значении ошибки, близком к нулю. Максимальная плотность ошибок наблюдается в диапазоне −0.5…0.5%, что согласуется с достигнутыми значениями RMSE ≈ 0.7% и MAE <0.5%, подтверждая высокую точность и отсутствие систематического смещения модели.

На рисунке 5 показано состояние процесса обучения нейросетевой модели, включающее динамику градиента, параметра демпфирования μ алгоритма Левенберга–Маркуардта и число проверок валидации.

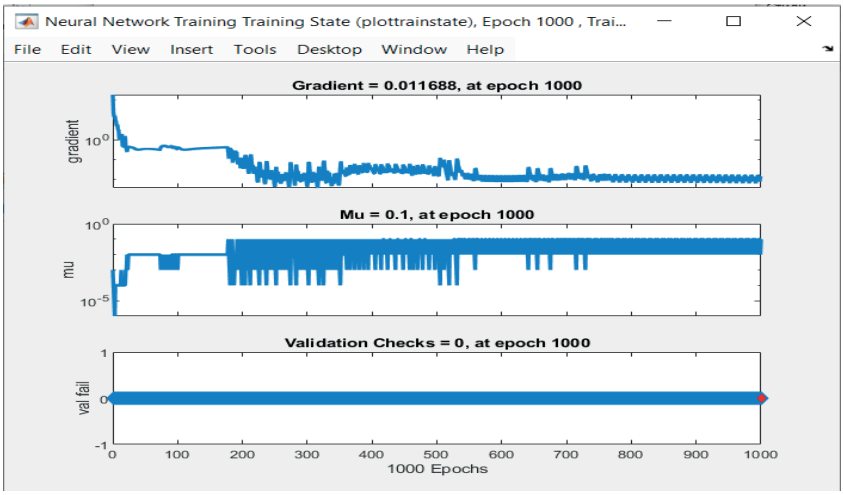


Рисунок 5 – Состояние обучения нейросетевой модели

К 1000-й эпохе значение градиента снижается до **0.0117**, параметр μ=0.1, а проверки валидации отсутствуют, что указывает на устойчивую сходимость модели и корректную работу алгоритма обучения.

На рисунке 6 представлена регрессионная зависимость между фактическими значениями целевого показателя KPI (ось абсцисс) и значениями, предсказанными нейросетевой моделью на обучающей выборке (ось ординат). Каждая точка соответствует отдельному наблюдению, пунктирная линия Y=T отражает идеальное совпадение прогноза с фактическими данными, а сплошная линия представляет линейную аппроксимацию выходов модели.

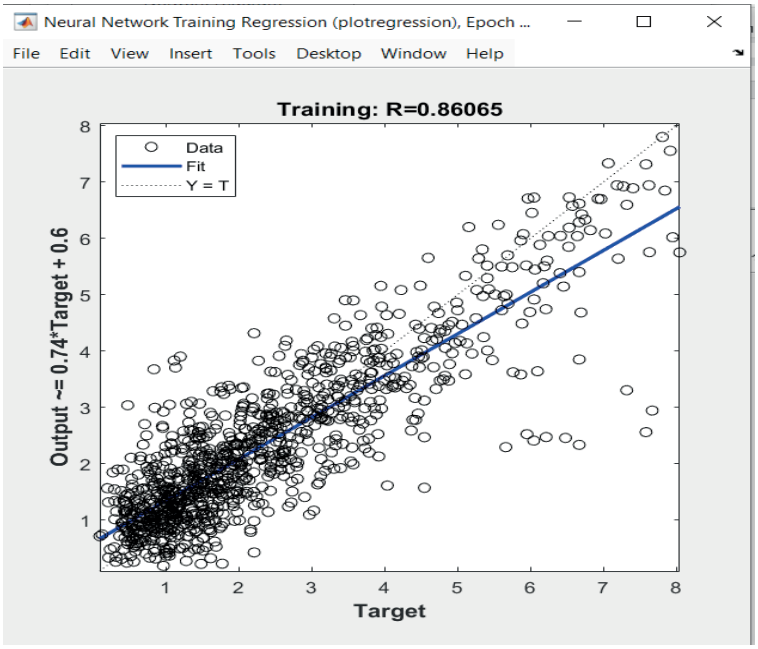


Рисунок 6 – Регрессионный анализ результатов обучения нейросетевой модели

Полученное значение коэффициента корреляции R = 0.86065 указывает на сильную линейную зависимость между прогнозируемыми и реальными значениями KPI. Большинство точек сосредоточено вблизи линии идеального соответствия, что свидетельствует о высокой способности модели воспроизводить общую структуру данных. Наблюдаемый разброс в области больших значений целевой переменной объясняется повышенной вариативностью KPI и наличием редких экстремальных состояний сети.

Таким образом, регрессионный анализ подтверждает адекватность обученной нейросетевой модели и согласуется с ранее полученными значениями метрик RMSE, MAE и коэффициента детерминации R^2.

Обоснование выбранной методологии

Предложенная методология - комбинация предобработки данных, многофакторного моделирования и временного анализа - ориентирована на практическое применение в системах мониторинга и оптимизации сети. Использование реальных KPI, а не синтетических данных, делает модель ближе к реальным сценариям эксплуатации. Выбор MLP и LSTM оправдан тем, что они дополняют друг друга: первая модель хорошо описывает статический «портрет» соты, вторая - её поведение во времени.

За счёт такого подхода удаётся не просто построить «чёрный ящик», а сформировать интерпретируемое решение: по совокупности KPI можно оценивать вероятность аномальных разъединений, отслеживать её динамику и использовать результаты прогноза для принятия инженерных решений - от

изменения параметров handover и таймеров до перенастройки радиоресурса в рамках SON.

Заключение

Мы провели исследование, направленное на прогнозирование вероятности аномальных разъединений в сетях 5G, анализируя временные ряды ключевых показателей качества обслуживания (KPI). Актуальность этой работы обусловлена высокой динамичностью современных телекоммуникационных сетей и потребностью в заблаговременном выявлении ухудшений качества для повышения надежности и устойчивости сетевой инфраструктуры. Ключевым этапом нашего исследования стала комплексная предварительная обработка данных KPI. Мы очистили данные от выбросов, заполнили пропуски и использовали нейронные сети для сглаживания временных рядов. Результаты показывают, что эти шаги значительно уменьшают шум и делают входные данные более стабильными, что напрямую улучшает точность прогнозирования.

В рамках задачи прогнозирования были созданы и сопоставлены нейросетевые модели на базе многослойного персептрона (MLP) и рекуррентной сети LSTM. Хотя обе модели успешно отражают общую динамику показателя Abnormal Release Rate, модель LSTM продемонстрировала превосходную точность. Её способность учитывать временные зависимости в данных позволила значительно снизить ошибки прогнозирования (RMSE, MAE) и повысить объясняющую способность (R^2) по сравнению с MLP. Дополнительный анализ подтвердил стабильность обучения нейросетей, отсутствие систематических ошибок и их высокую применимость для анализа реальных эксплуатационных данных 5G.

Исследование подтвердило, что глубокое обучение является мощным инструментом для анализа и прогнозирования KPI телекоммуникационных сетей. Разработанный подход открывает возможности для создания интеллектуальных систем мониторинга и поддержки принятия решений для операторов связи. Его потенциал может быть расширен путем включения дополнительных сетевых показателей и разработки многомерных, многокомпонентных моделей прогнозирования.

Будущие исследования должны быть направлены на создание многомерных нейросетевых моделей, которые смогут комплексно учитывать взаимосвязь различных KPI сетей 5G. Также перспективным является применение гибридных архитектур глубокого обучения, таких как CNN–LSTM и Transformer. Интеграция прогнозных моделей в системы предиктивного управления радиосетью (Predictive SON) позволит проактивно предотвращать ухудшение качества обслуживания. Особое внимание следует уделить распределенным и федеративным подходам к обнаружению аномалий, особенно в сценариях 5G-IoT и edge-вычислений.

References

Awofolaju T., Lasisi H.O., Olawuyi A., Adebayo S. (2023) Artificial Neural Network Based Prediction of Key Performance Indicators for Mobile Telecommunications. Journal of Science and Technology, 15(2). https://doi.org/10.30880/jst.2023.15.02.005 (in Eng.).

Stojčić M., Banjanin M.K., Vasiljević M., et al. (2023) Predictive Modeling of Delay in an LTE Network by Optimizing the Number of Predictors Using Dimensionality Reduction Techniques. Applied Sciences, 13(14). — 8511p.https://doi.org/10.3390/app13148511 (in Eng.).

Nurakhov Y., Mukhanbet A., Aibagarov S., Imankulov T.A (2025) Machine Learning Approach to Investigating Key Performance Factors in 5G Standalone Networks. Electronics, 14(19), 3817. https://doi.org/10.3390/electronics14193817 (in Eng.).

Nassef O., Sun W., Purmehdi H., Tatipamula M., Mahmoodi T.A (2022) Survey: Distributed Machine Learning for 5G and Beyond. Computer Networks, 207, 108820. https://doi.org/10.1016/j.comnet.2022.108820 (in Eng.).

Islam M.A., Siddique H., Zhang W., et al. (2023) A Deep Neural Network-Based Communication Failure Prediction Scheme in 5G RAN. IEEE Transactions on Network and Service Management, 20(2). — P. 1140–1152. https://doi.org/10.1109/TNSM.2022.3229658 (in Eng.).

Arun Prasad K., Abdul Basith M., Harish Kumar V. (2025) Deep Autoencoder-Based Anomaly Detection for Intelligent Network Slice Monitoring in B5G Networks. Journal of Artificial Intelligence and Capsule Networks, 7(2). — P. 107–124. https://doi.org/10.36548/jaicn.2025.2.002 (in Eng.).

Wang C., et al. (2021) TSAGen: Synthetic Time Series Generation for KPI Anomaly Detection. IEEE Transactions on Network and Service Management, 18(4). https://doi.org/10.1109/TNSM.2021.3098784 (in Eng.).

Wu J., Lee P.P., Li Q., Pan, L., Zhang, J. (2018) CellPAD: Detecting Performance Anomalies in Cellular Networks via Regression Analysis. In: Proc. IFIP Networking Conference (IFIP Networking). — P. 1–9. https://doi.org/10.23919/IFIPNetworking.2018.8697027 (in Eng.).

Yuliana H., Hendrawan I., Musashi Y. (2024) Estimating Base Station Traffic and Throughput Using Machine Learning Based on Hourly Key Performance Indicator (KPI) Network Analysis. IEEE Access, 12. — P. 116285–116301. https://doi.org/10.1109/ACCESS.2024.3447098 (in Eng.).

Liu X., Li Y., Zhu S., et al. (2025) A Traffic Forecasting Framework for Cellular Networks Based on a Dynamic Component Management Mechanism. Electronics, 14(20), 4003. https://doi.org/10.3390/electronics14204003 (in Eng.).

Almeida A., Rito P., Brás S., Pinto F.C., Sargento S. (2024) A Machine Learning Approach to Forecast 5G Metrics in a Commercial and Operational 5G Platform: 5G and Mobility. Computer Communications, 228, 10797. https://doi.org/10.23919/WONS60642.2024.10449577 (in Eng.).

Al-Thaedan A., Shakir Z., Mjhool A.Y., et al. (2024) A Machine Learning Framework for Predicting Downlink Throughput in 4G-LTE/5G Cellular Networks. International Journal of Information Technology, 16(2). — P. 651–657. https://doi.org/10.1007/s41870-023-01678-w (in Eng.).

Kim J., Thulasiraman P. (2024) Anomaly Detection in 5G Networks Using Transformer-Based Autoencoder. In: 2024 International Conference on Recent Advances in Systems Science and Engineering (RASSE). https://doi.org/10.1109/RASSE64357.2024.10773774 (in Eng.).

Khan I., Abdullah J., et al. (2020) Network Anomaly Detection Using LSTM Based Autoencoder. In: Proc. 2020 3rd International Conference on Computing and Informatics (ICCI), 2020. https://www.researchgate.net/publication/346006810 (in Eng.).

Reis M.J.C.S. (2025) Edge-FLGuard+: A Federated and Lightweight Anomaly Detection Framework for Securing 5G-Enabled IoT in Smart Homes. Future Internet, 17(8). — 329 p. https://doi.org/10.3390/fi17080329 (in Eng.).

K.S. Chezhimbayeva, A. Mukhamedzhanova, Yu. Garmashova. (2025) Ekspertnaya sistema prognozirovaniya qos v 5G-Setyah na osnove nechetkoj logikiempirically [Expert qos forecasting system in 5G networks based on fuzzy logic]. Ademic scientific journal of computer science. Volume 4. Namber 356. P. 306–320. https://doi.org/10.32014/2025.2518-1726.398 https://doi.org/10.32014/2025.2518-1726.398 (in Russ.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 148–166

https://doi.org/10.32014/2026.2518-1726.406

IRSTI 20.53.19
UDC 004.046

© Dauitbayeva A.O., Konyrbaev N.B., Abildayeva Zh.T.*, Yessirkepova A.U., Karim N.A., 2026.
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan.
E-mail: Zhansaya_A@mail.ru

DEVELOPMENT OF AN APPLICATION TO OPTIMIZE THE PROCESS OF EMPLOYMENT OF GRADUATES

Dauitbayeva Aigul — Candidate of Technical Sciences, Associate Professor (Docent), Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan,
E-mail: aicos@mail.ru, https://orcid.org/0000-0003-2487-0687;
Konyrbaev Nurbek — PhD, Associate Professor (Docent), Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan,
E-mail: n.konyrbaev@mail.ru, https://orcid.org/0000-0002-8788-4149;
Abildayeva Zhansaya — Master of Technical Sciences, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan,
E-mail: Zhansaya_A@mail.ru, https //orcid.org/0000-0002-2637-0443;
Yessirkepova Aizhan—– Master of Technical Sciences, Senior Lecturer, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan,
E-mail: m_j_ai@mail.ru, https://orcid.org/0000-0003-3219-3022;
Karim Nazerke — Master of Technical Sciences, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan,
E-mail: knazerke20@gmail.com, https://orcid.org/0009-0006-9460-4252.

Abstract. Nowadays, mobile applications have become an essential tool for job searching among young specialists entering the labor market. Despite their widespread adoption, many of these platforms fail to address the unique needs, challenges, and preferences of entry-level job seekers. This research aims to optimize job search methods for young professionals by leveraging specialized mobile solutions. To gain insights, interviews were conducted with recent graduates and early-career individuals to identify the primary obstacles they face during their job search, including the lack of personalized recommendations, inefficient filtering options, and limited access to tailored resources. Furthermore, a thorough analysis of user preferences and the usability of existing job search applications was performed. Based on the findings, we proposed strategies to enhance the functionality of recommendation algorithms, improve user interface designs, and create more engaging and supportive user experiences. The study emphasizes the

importance of personalization, intuitive navigation, and dynamic content in modern job search platforms. The results highlight that implementing these improvements can significantly streamline the job search process, making it more efficient, accessible, and user-friendly for young professionals. This research contributes to the growing body of knowledge on user-centered design for mobile employment solutions. Another important advantage of cluster analysis is its ability to identify ambiguous or less obvious groups that might remain unnoticed when using other methods. For instance, in studying the career trajectories of young professionals, this approach can reveal not only traditional paths such as analytics or software development, but also emerging trends — roles in new technological fields or highly specialized positions that are increasingly in demand.

Keywords: career, forecasting, analysis, data, optimization, skills, prediction, development, employment

For citations: Dauitbayeva A.O., Konyrbaev N.B., Abildayeva Zh.T., Yessirkepova A.U., Karim N.A. Development of an application to optimize the process of employment of graduates. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 148–166. DOI: https://doi.org/10.32014/2026.2518-1726.406

© Дәуітбаева А.О., Қоңырбаев Н.Б., Әбілдаева Ж.Т.*, Есіркепова А.У., Кәрім Н.Ә., 2026.

Қорқыт ата атындағы Қызылорда Университеті, Қазақстан.
E-mail: Zhansaya_A@mail.ru

БІТІРУШІ ТҮЛЕКТЕРДІҢ ЖҰМЫСҚА ОРНАЛАСТЫРУ ПРОЦЕСІН ОҢТАЙЛАНДЫРУ ҮШІН ҚОСЫМША ӘЗІРЛЕУ

Дәуітбаева Айгүл — техника ғылымдарының кандидаты, қауымдастырылған профессор (доцент), Қорқыт ата атындағы Қызылорда Университеті, Қызылорда, Қазақстан,
E-mail: aicos@mail.ru, https://orcid.org/0000-0003-2487-0687;
Қоңырбаев Нұрбек — PhD, қауымдастырылған профессор (доцент), Қорқыт ата атындағы Қызылорда Университеті, Қызылорда, Қазақстан,
E-mail: n.konyrbaev@mail.ru, https://orcid.org/0000-0002-8788-4149;
Әбілдаева Жансая — техника ғылымдарының магистрі, Қорқыт ата атындағы Қызылорда Университеті, Қызылорда, Қазақстан,
E-mail: Zhansaya_A@mail.ru, https //orcid.org/0000-0002-2637-0443;
Есіркепова Айжан — техника ғылымдарының магистрі, аға оқытушы, Қорқыт ата атындағы Қызылорда Университеті, Қызылорда, Қазақстан,
E-mail: m_j_ai@mail.ru, https://orcid.org/0000-0003-3219-3022;
Кәрім Назерке — техника ғылымдарының магистрі, Қорқыт ата атындағы Қызылорда Университеті, Қызылорда, Қазақстан,
E-mail: knazerke20@gmail.com, https://orcid.org/0009-0006-9460-4252.

Аннотация. Қазіргі заманда мобильді қосымшалар жұмысқа кіретін жас мамандар арасында жұмыс іздеудің таптырмас құралына айналды.

Кең таралғанына қарамастан, бұл платформалардың көпшілігі бастапқы кезеңде жұмыс іздейтіндердің бірегей қажеттіліктерін, қиындықтарын, қалауларын қанағаттандыра алмайды. Бұл зерттеу мамандандырылған мобильді шешімдерді қолдану арқылы жас мамандарға жұмыс іздеу әдістерін оңтайландыруға бағытталған. Бұл зерттеуді тереңдету үшін жуық арада бітірген түлектермен және мансапқа ұмтылған адамдармен сұхбат жүргізілді, олар жұмыс іздеу кезінде кездесетін негізгі кедергілерді, соның ішінде жекелендірілген ұсыныстардың жоқтығын, сүзудің тиімсіз нұсқаларын және арнайы ресурстарға қол жетімділіктің шектеулі екенін растады. Сонымен қатар, пайдаланушылардың қалауы мен жұмыс іздеуге арналған қолданыстағы қосымшалардың ыңғайлылығына мұқият талдау жасалды. Алынған нәтижелерге сүйене отырып, алгоритмдерінің функционалдығын кеңейту, пайдаланушы интерфейсінің дизайнын жақсарту пайдаланушылармен неғұрлым тартымды және қолдау көрсететін тәжірибені құру стратегияларын ұсындық. Зерттеу заманауи жұмыс іздеу платформаларында жекелендірудің, интуитивті навигацияның динамикалық мазмұнның маңыздылығын көрсетеді. Нәтижелер көрсеткендей, бұл жақсартуларды енгізу жұмыс іздеу процесін айтарлықтай жеңілдетіп, тиімдірек және қолжетімді жас мамандарға ыңғайлы етеді. Бұл зерттеу пайдаланушыға бағытталған мобильді жұмысқа орналасу шешімдерінің дизайны туралы білімнің артуына ықпал етеді. Кластерлік талдаудың тағы бір маңызды артықшылығы – басқа әдістер арқылы байқалмай қалуы мүмкін екіұшты немесе анық емес топтарды айқындау қабілеті. Мәселен, жас мамандардың мансаптық жолдарын зерттегенде, бұл тәсіл тек аналитика немесе бағдарламалық жасақтама әзірлеу сияқты дәстүрлі бағыттарды ғана емес, сонымен қатар жаңа технологиялық салалардағы немесе тар маманданған лауазымдардағы сұранысқа ие рөлдер секілді жаңа үрдістерді де ашып көрсете алады. Әдістің басты артықшылықтарының бірі – үлкен көлемдегі деректерді тиімді өңдеу және олардың ішіндегі жасырын заңдылықтарды анықтау мүмкіндігі. Ол мансаптық траекторияларды болжауға, жас мамандарға қажетті дағдыларды дамытуға бағытталған ұсыныстар жасауға көмектеседі. Сонымен қатар, кластерлік талдау тек жұмыспен қамту процестерін оңтайландыруға ғана емес, мансаптық жоспарлауды жүйелі жүргізуге де жол ашады.

Түйін сөздер: мансап, болжау, талдау, деректер, оңтайландыру, дағдылар, дамыту, жұмысқа орналастыру

**© Даутбаева А.О., Конырбаев Н.Б., Абильдаева Ж.Т.*,
Есиркепова А.У., Карим Н.А., 2026.**
Кызылординский Университет имени Коркыт Ата, Кызылорда, Казахстан.
E-mail: Zhansaya_A@mail.ru

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ ОПТИМИЗАЦИИ ПРОЦЕССА ТРУДОУСТРОЙСТВА ВЫПУСКНИКОВ

Даутбаева Айгуль — кандидат технических наук, ассоциированный профессор (доцент), Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан, E-mail: aicos@mail.ru, https://orcid.org/0000-0003-2487-0687;
Конырбаев Нурбек — PhD, ассоциированный профессор (доцент), Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан,
E-mail: n.konyrbaev@mail.ru, https://orcid.org/0000-0002-8788-4149;
Абилдаева Жансая — магистр технических наук, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан,
E-mail: Zhansaya_A@mail.ru, https //orcid.org/0000-0002-2637-0443;
Есиркепова Айжан — магистр технических наук, старший преподаватель, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан,
E-mail: m_j_ai@mail.ru, https://orcid.org/0000-0003-3219-3022;
Карим Назерке — магистр технических наук, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан,
E-mail: knazerke20@gmail.com, https://orcid.org/0009-0006-9460-4252.

Аннотация. В настоящее время мобильные приложения становятся ключевым инструментом поиска работы среди молодых специалистов, выходящих на рынок труда. Несмотря на их широкое распространение, многие существующие платформы не в полной мере учитывают специфические потребности и особенности пользователей на начальном этапе профессиональной карьеры. Целью данного исследования является разработка и обоснование подходов к оптимизации процесса трудоустройства выпускников на основе специализированного мобильного приложения. Для достижения поставленной цели проведены интервью с недавними выпускниками и начинающими специалистами, что позволило выявить основные барьеры при поиске работы, включая недостаточную персонализацию рекомендаций, ограниченные возможности фильтрации вакансий и дефицит специализированных ресурсов. Дополнительно выполнен анализ пользовательских предпочтений и удобства использования существующих платформ. На основе полученных результатов предложены решения, направленные на совершенствование алгоритмов рекомендаций, улучшение пользовательского интерфейса и повышение качества взаимодействия с пользователем. В исследовании подчёркивается значимость персонализации, интуитивной навигации и динамического контента в современных цифровых платформах трудоустройства. Результаты показывают, что внедрение предложенных решений позволяет повысить эффективность, доступность и удобство процесса поиска работы для молодых

специалистов. Дополнительно в работе рассматривается применение методов кластерного анализа, позволяющих выявлять скрытые группы пользователей и закономерности в данных. Данный подход обеспечивает возможность прогнозирования карьерных траекторий и формирования рекомендаций по развитию профессиональных навыков. Таким образом, использование интеллектуальных методов анализа данных способствует не только оптимизации процессов трудоустройства, но и более эффективному планированию карьерного развития, повышая адаптивность молодых специалистов к условиям динамичного рынка труда.

Ключевые слова: карьера, прогнозирование, анализ, данные, оптимизация, навыки, прогнозирование, развитие, трудоустройство

Кіріспе. Қазіргі қарқынды дамып келе жатқан еңбек нарығы жағдайында және алғашқы жұмысын іздеп жүрген түлектер санының көбеюіне байланысты жас мамандарды тиімді жұмыспен қамту мәселесі аса өзекті болып отыр. Жұмыс іздеудің дәстүрлі тәсілдері – бос жұмыс орындарын талдап, оларды үміткерлердің дағдыларымен сәйкестендіру – кәсіби тәжірибесі шектеулі немесе мүлде жоқ жаңа түлектерге жеткіліксіз болып жатады. Мұндай жас мамандар бірқатар қиындықтарға тап болады: кәсіби бағыт таңдаудағы анықсыздық, тағылымдамадан өту мүмкіндіктерінің шектеулігі, нақты мансаптық перспективаларды әлсіз түсіну және соның салдарынан алғашқы жұмыс орнын таңдаудағы сенімсіздік.

Осы зерттеудің өзектілігі – жас мамандардың жұмыс іздеу процесінде дербестендірілген (жекелендірілген) тәсілді қолданудың маңыздылығында. Деректерге негізделген мансаптық траекторияларды талдау түлектердің дағдылары, білімі мен қалауына, сондай-ақ еңбек нарығындағы үрдістерге сүйене отырып, ең қолайлы жұмыс бағыттарын дәл анықтауға мүмкіндік береді (Breiman, 2001). Әсіресе кластерлік талдау (cluster analysis) – ұқсас сипаттамалары бар түлектерді топтастырып, олардың ең ықтимал әрі перспективалы мансаптық бағыттарын айқындауға арналған тиімді құрал болып табылады. Бұл әдіс жас мамандарға алғашқы жұмысын табу уақытын қысқартуға және кәсіби қанағаттанушылық деңгейін арттыруға көмектеседі, себебі олар өз саласындағы талаптар мен күтілімдерге жақсырақ дайындалады.

Сонымен қатар кластерлік талдау әдістерін қолдану жұмыс берушілерге де пайдалы болуы мүмкін. Себебі олар құрылымдалған деректер мен мансаптық болжаулар негізінде жас мамандардың әлеуетін дәлірек бағалай алады. Бұл компанияларға түлектерге арналған жұмыс профилін нақты құруға және олардың деңгейіне сай тапсырмалар беруге мүмкіндік береді (Li and Zhao, 2025). Осылайша, деректерді талдау әдістерін, соның ішінде кластерлік талдауды қолдану арқылы жас мамандарды жұмыспен қамту процесін оңтайландыруға бағытталған бұл зерттеу түлектер мен жұмыс берушілер арасындағы өзара әрекеттестікті жетілдіру тұрғысынан аса маңызды. Ұзақ

мерзімде мұндай тәсіл кадр тұрақсыздығын азайтып, еңбек ресурстарын әртүрлі салалар арасында тиімді бөлуге ықпал етуі мүмкін (Bishop, 2006).

Әдеби шолу. Қазіргі цифрландыру жағдайында мобильді қосымшалар еңбек нарығында, әсіресе жас мамандар үшін, жұмыс іздеудің маңызды құралына айналып отыр. Ғылыми зерттеулер көрсеткендей, мобильді платформалар бос жұмыс орындарына қолжетімділікті жеңілдетіп, жұмыс іздеу үдерісін жеделдетеді. Алайда, олардың кеңінен таралуына қарамастан, көптеген қосымшалар еңбек жолын енді бастаған пайдаланушылардың ерекшеліктері мен нақты қажеттіліктерін толық ескере бермейді.

Әдебиеттерді талдау нәтижелері қазіргі жұмыс іздеу платформаларының негізгі кемшіліктерінің бірі – жекелендіру деңгейінің жеткіліксіздігі екенін көрсетеді. Зерттеушілер пайдаланушылардың дағдыларына, қызығушылықтарына және мансаптық мақсаттарына негізделген ұсыныстардың болмауы жұмыс іздеудің тиімділігін төмендететінін атап өтеді. Сонымен қатар, вакансияларды сүзу механизмдерінің тиімсіздігі және жас мамандарға арналған арнайы ресурстардың шектеулі болуы да өзекті мәселелер қатарында қарастырылады. Бірқатар зерттеулерде мобильді қосымшалардың пайдаланушы интерфейсі мен тәжірибесіне (UX/UI) ерекше назар аударылады. Интуитивті навигация, қарапайым әрі түсінікті интерфейс және мазмұнды динамикалық түрде ұсыну пайдаланушы белсенділігіне және жұмыс іздеу нәтижелілігіне тікелей әсер ететіні дәлелденген. Осыған байланысты интерфейс дизайнын жетілдіру цифрлық жұмысқа орналастыру платформаларының тиімділігін арттырудың маңызды факторы ретінде қарастырылады. Сондай-ақ, деректерді талдау әдістерін, әсіресе кластерлік талдауды қолдану мәселесі кеңінен зерттелуде. Кластерлік талдау пайдаланушыларды әртүрлі сипаттамалар бойынша топтастыруға, жасырын заңдылықтарды анықтауға мүмкіндік береді. Бұл әдіс кәсіби қызығушылықтар, дағдылар және мансаптық бағыттар негізінде пайдаланушыларды сегменттеуге жағдай жасап, ұсыныстардың дәлдігін арттыруға ықпал етеді.

Зерттеулер нәтижесінде кластерлік талдау еңбек нарығындағы айқын емес немесе жаңа қалыптасып келе жатқан үрдістерді анықтауда тиімді құрал екені дәлелденген. Атап айтқанда, бұл әдіс дәстүрлі мамандықтармен қатар жаңа технологиялық бағыттар мен тар бейінді салалардағы сұранысқа ие рөлдерді айқындауға мүмкіндік береді. Бұл қазіргі технологиялардың қарқынды дамуы жағдайында ерекше өзектілікке ие. Осылайша, ғылыми еңбектерді талдау пайдаланушыға бағытталған, жекелендірілген қызметтерді ұсынатын, тиімді алгоритмдер мен ыңғайлы интерфейсті біріктіретін мобильді шешімдерді әзірлеудің қажеттілігін көрсетеді. Мұндай тәсілдерді енгізу жұмыс іздеу үдерісін жеңілдетіп қана қоймай, жас мамандардың мансаптық дамуын тиімді жоспарлауға мүмкіндік береді.

Материалдар мен әдістер. **Деректерді талдау және мансаптық траекторияны болжау әдістері**. Жас мамандарды жұмыспен қамту мен

олардың кәсіби дамуының тиімділігін арттыру тұрғысынан деректерді талдау және мансаптық траекторияны болжау әдістері маңызды құралдарға айналды. Қазіргі заманғы деректерді талдау әдістері еңбек қызметіндегі жасырын заңдылықтарды анықтап қана қоймай, болашақ мансаптық өсу бағыттарын болжауға да мүмкіндік береді. Бұл өз кезегінде жұмыс іздеу процесін оңтайландыруға ықпал етеді (McKinsey Global Institute, 2020). Осы бөлімде мансаптық траекторияларды зерттеуге қолданылатын негізгі деректерді талдау әдістері қарастырылады, олардың ішінде бұл зерттеу жұмысының басты әдісі ретінде **кластерлік талдауға** ерекше назар аударылады.

Деректерді талдауда мансаптық траекторияларды болжауға арналған ең кең тараған әдістердің бірі - **регрессиялық талдау (regression analysis)**. Бұл әдіс әртүрлі айнымалылар арасындағы байланыстарды анықтап, бір айнымалының мәнін басқаларының негізінде болжауға мүмкіндік береді (Pereira and Raposo, 2019). Мансаптық даму контексінде регрессиялық талдау білім деңгейі, жұмыс тәжірибесі немесе жас сияқты факторлардың мансаптық өсуге және қызметте жылжу жылдамдығына қалай әсер ететінін бағалауға қолданылады. Мысалы, қосымша білім алу сияқты факторлардың өзгерісі басқарушылық лауазымға тағайындалу мүмкіндігіне қалай ықпал ететінін болжауға болады. Алайда бұл әдістің кемшілігі - ол айнымалылар арасындағы нақты сызықтық (linear) байланыстарды қажет етеді (Rigas, 2021), ал кәсіби даму саласында факторлар жиі бейсызық (nonlinear) әсер етеді.

Тағы бір маңызды әдіс - **уақыттық қатарларды талдау (time series analysis)**. Бұл әдіс уақыт бойынша өзгеретін оқиғаларды немесе процестерді болжауға арналған. Мансаптық даму саласында уақыттық қатарлар талдауы маманның жалақысы мен лауазымының белгілі бір кезең ішіндегі өзгерісін бақылауға мүмкіндік береді. Осылайша, бұл әдіс өткен тенденцияларға сүйене отырып, болашақтағы өзгерістерді - мысалы, қызметте жоғарылау немесе мансап бағытын өзгерту ықтималдығын - болжауға мүмкіндік береді. Дегенмен, бұл әдісті қолдануға жеткілікті көлемде деректер қажет, ал еңбек өтілі аз жас мамандарға мұндай деректер базасын жинау қиын болуы мүмкін. **Шешім ағаштары (decision trees)** - мансаптық траекторияларды талдауда жиі қолданылатын тағы бір әдіс. Бұл әдіс деректерді түрлі атрибуттарға (қасиеттерге) сүйене отырып категорияларға бөлу арқылы классификация және болжау моделін құрады (Noack, 2019). Мысалы, шешім ағашы жас мамандарды білім деңгейі мен кәсіби дағдыларына қарай басқарушылық немесе техникалық бағыттар бойынша бірнеше топқа бөлуі мүмкін. Бұл әдіс мансаптық өсуде қай айнымалылардың маңызды екенін анық көрсетеді, алайда дәл нәтижеге қол жеткізуге көп көлемді деректер қажет. Сонымен қатар, **ережеге негізделген әдістер (rule-based methods)** де мансаптық траекторияларды болжауда қолданылады. Бұл - машиналық оқытуға негізделген тәсіл, онда деректердің түрлі сипаттамалары арасындағы логикалық байланыстар ережелер түрінде құрылады. Мысалы, белгілі бір салада магистр дәрежесі бар және ірі компанияда тағылымдамадан өткен маманның жоғары лауазымға

көтерілу ықтималдығы айтарлықтай жоғары деген қорытынды шығару мүмкін мүмкін (Mohrman and Lawler, 2012). Мұндай әдіс бірнеше факторларды бір уақытта ескеретін күрделі әрі бейімделгіш модельдер құруға мүмкіндік береді, бірақ оларды баптау мен оқыту процесі көп еңбекті қажет етеді.

Ережеге негізделген әдіс мансаптық траекторияларды болжауға да қолданылады. Бұл машиналық оқыту тәсілі деректердің әртүрлі сипаттамаларын байланыстыратын логикалық ережелерді құрастырады. Мысалы, ережелер жиынтығы маманның белгілі бір салада магистр дәрежесі болса және ірі компанияда тағылымдамадан өткен болса, оның мансаптық өсуінің белгілі бір лауазымға ықтималдығы айтарлықтай артады деген қорытынды жасауға болады. Бұл әдіс бірнеше факторларды ескере алатын күрделі және бейімделгіш модельдерді жасауға көмектеседі, дегенмен оларды оқыту өте көп еңбекті қажет етуі мүмкін (Mohrman and Lawler, 2012).

Деректерді талдау әдістерінің алуан түрлілігіне қарамастан, кластерлік талдау мансаптық траекторияларды зерттеуге ең тиімді әдістердің бірі болып табылады, әсіресе мансабын жаңадан бастаған жас мамандарға қатысты. Кластерлік талдау - бұл бақылаусыз оқытуға негізделген әдіс, мұнда мақсат нақты құндылықтарды болжау емес, ұқсас сипаттамалары бар объектілер тобын анықтау (Banerjee and Sharma, 2024). Мансаптық траекториялар контексінде кластерлік талдау мамандарды олардың білім беру, кәсіби және жеке қасиеттеріне қарай топтастыруға қолданылады. Бұл әдіс жас мамандарды бірнеше кластерлерге бөлуге мүмкіндік береді, олардың әрқайсысы мансаптық жолдары ұқсас адамдар тобын білдіреді. Кластерлік талдаудың негізгі артықшылықтарының бірі - бұл тәуелділіктердің сипаты туралы алдын ала болжам жасамай-ақ, мансаптық өсуге әсер ететін жасырын заңдылықтар мен факторларды ашу мүмкіндігі (Banerjee and Sharma, 2024). Мысалы, кластерлік талдау белгілі бір дағдылар жиынтығы және гуманитарлық ғылымдар бойынша жоғары білімі бар мамандар тобының маркетинг пен жарнама саласында жұмыс істеуі мүмкін екенін, ал техникалық білімі мен бағдарламалау дағдылары бар басқа топ ақпараттық технологиялар саласында мансап құруға бейім екенін анықтай алады. Осылайша, кластерлік талдау мансаптық траекторияларды түрлерін анықтауға ғана емес, сонымен қатар белгілі бір мамандар тобы үшін қандай мансаптық жолдардың ең сәтті болуы мүмкін екенін болжауға да көмектеседі (Venkatesh et al., 2012). Кластерлік талдауды жас мамандарға да, жұмыс берушілерге де ұсыныстар жасауға пайдалануға болады. Түлектерге ол өткен тәжірибе мен қалауларға негізделген оңтайлы мансаптық жолдарды ұсына алады, бұл оларға еңбек нарығында тезірек және дәл жүруге көмектеседі. Жұмыс берушілерге кластерлік талдау белгілі бір лауазымдарда қандай кандидаттардың ең табысты болуы мүмкін екенін дәл анықтау арқылы жұмысқа қабылдау процестерін жақсартады, осылайша жалдау шығындарын азайтады және тиімділікті арттырады. Осылайша, кластерлік талдау мансаптық траекторияларды болжауда қуатты құрал болып табылады, жұмыс іздеушілер мен жұмыс берушілерге дәлірек

және жекелендірілген ұсыныстар ұсынады (Zhang and Gu, 2025). Бұл әдіс, әсіресе, мансаптық жолдары әлі қалыптасып келе жатқан жас мамандарды талдауға өте қолайлы және білімі, жеке қызығушылықтары мен дағдылары сияқты көптеген факторларға байланысты айтарлықтай өзгеруі мүмкін.

Кластерлік талдау: әдістері мен мансаптық траекторияны болжаудағы қолданылуы

Кластерлік талдаудың негізгі принципі – объектілерді топтарға бөлу, яғни топ ішіндегі ұқсастықты барынша арттырумен қатар топтар арасындағы ұқсастықты барынша азайту. Алдын ала анықталған санаттар қолданылатын классификация әдістерінен айырмашылығы, кластерлік талдау бастапқы деректермен жұмыс істейді және жасырын құрылымдарды анықтауға бағытталған (Maciariello, 2018). Бұл әдіс маркетинг, биология және әлеуметтік ғылымдар сияқты салаларда, сондай-ақ мансаптық траекторияларды зерттеу кезінде өте пайдалы, өйткені бұл салаларда айқын санаттар немесе алдын ала анықталған даму жолдары жиі болмайды.

Кластерлік талдаудағы негізгі ұғымдардың бірі – қашықтық метрикасы, ол объектілер арасындағы ұқсастықты бағалауға қолданылады. Ең жиі қолданылатын метрикалар – Евклидтік қашықтық және Манхэттен қашықтығы. Евклидтік қашықтық (немесе жай ғана қашықтық) объектілердің сәйкес белгілері арасындағы айырмашылықтардың квадраттарының қосындысының квадрат түбірі ретінде есептеледі. Мансаптық траекторияларға бұл белгілерге білім, кәсіби дағдылар, тәжірибе жылдары және қалаулар кіруі мүмкін. Егер екі маманның дағдылары мен тәжірибесі ұқсас болса, олардың арасындағы қашықтық аз болады және олар бір кластерге жатады (Cappelli and Keller, 2012).

Нәтижелер мен талқылау. K-Means алгоритмі – статистика мен машиналық оқытуда кеңінен қолданылатын, ең тиімді кластерлеу әдістерінің бірі. Бұл алгоритм деректер жиынтығын олардың ұқсастығына қарай K кластерге бөлуге мүмкіндік береді және кластер ішіндегі қашықтықты (variance) барынша азайтуға бағытталған (Ibrahim and Iqbal, 2025). Әдістің мәні — әрбір кластердің орталығын (центроидын) табу және осы орталықтан кластердегі барлық нүктелерге дейінгі орташа қашықтықты минимизациялау болып табылады.

Математикалық негіз. K-Means алгоритмі кластер ішіндегі дисперсияны (within-cluster variance) азайту қағидасына негізделген. Ол деректер нүктелерінің өз кластерінің центроидынан қашықтықтарының квадраттарының қосындысы ретінде анықталады. Минимизация процесі келесі формуламен өрнектеледі:

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left|\left|x_j - \mu_i\right|\right|^2, \qquad (1)$$

Мұндағы:

J — мақсатты функция (кластерлеу сапасын сипаттайды);

K — кластерлер саны;

C_i — i-кластерге тиесілі деректер нүктелерінің жиыны.

x_j — Сі кластеріндегі j-деректер нүктесі.

μ_i — i-кластердің центрі (ор

тасы немесе орташа мәні).

$\|x_j - \mu_i\|^2$ — деректер нүктесі xj мен кластер центрі μi арасындағы Евклидтік қашықтықтың квадраты.

K-Means алгоритмі кластердегі барлық нүктелер мен центроидтар арасындағы квадраттық қашықтықтардың қосындысын білдіретін JJJ функциясын азайтуға бағытталған. JJJ мәні төмен болса, центроидтар тиісті кластерлердегі барлық нүктелерді дәлірек көрсетеді дегенді білдіреді (Ibrahim and Iqbal, 2025).

Алгоритм қадамдары

1. Бастапқыда, K центроидтары (кластер орталықтары) кездейсоқ таңдалады.

2. Кластерлерге нүктелерді тағайындау: Әрбір деректер нүктесі кластерге центроидқа дейінгі минималды қашықтыққа негізделіп тағайындалады. Нақтырақ айтқанда, әрбір xj деректер нүктесі центроиді μi ең кіші Евклидтік қашықтыққа ие кластерге тағайындалады:

$$C = \left\{ x_j \middle| \left|\left| x_j - \mu_i \right|\right|^2 \leq \left|\left| x_j - \mu_k \right|\right|^2 \forall k \neq i \right\}, \qquad (2)$$

Мұндағы $\|x_j - \mu_i\|^2$ x_j нүктесі мен μ_i центрі арасындағы қашықтық.

3. Центроидтарды жаңарту: Барлық деректер нүктелері кластерлерге тағайындалғаннан кейін, центроидтар жаңартылады. Әрбір кластерге жаңа центроид Сі сол кластерге тағайындалған барлық нүктелердің орташа мәні ретінде есептеледі:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \qquad (3)$$

4. 2 және 3 қадамдарды қайталау: 2 және 3 қадамдар орталық нүктелер өзгеруді тоқтатқанша немесе алдын ала белгіленген ең көп итерация санына жеткенше қайталанады.

Практикалық мысал. Бізде түлектердің дағдылары мен мансаптық қызығушылықтары туралы деректер жиынтығы бар делік. Біз оларды әртүрлі мансаптық жолдарға сәйкес келетін үш топқа бөлгіміз келеді: аналитика, даму және шығармашылық мамандықтар.

Кесте 1 – Студенттерге жасалған деректер

Студент	**Дағдылар**	Мансаптық қызығушылық
A	8, 5	Деректерді талдау
B	7, 6	Деректерді талдау
C	1, 8	Бағдарламалық қамтамасыз етуді әзірлеу
D	2, 9	Бағдарламалық қамтамасыз етуді әзірлеу
E	9, 1	Шығармашылық дизайн
F	8, 2	Шығармашылық дизайн

Мұнда «Дағдылар» екі өлшемді, аналитика (бірінші параметр) және бағдарламалау (екінші параметр) саласындағы дағдыларды білдіреді, ал «Мансапқа қызығушылық» осы студенттердің назар аударатын мамандықтарын білдіреді.

Біз кездейсоқ түрде үш бастапқы центроидты таңдаймыз деп есептейік (үш кластер үшін):

$\mu1 = (8,5)$ — аналитиктерге,

$\mu2 = (1,8)$ — әзірлеушілерге,

$\mu3 = (9,1)$ — шығармашылық мамандарына.

Әрбір деректер нүктесі үшін біз әрбір центроидқа дейінгі қашықтықты есептеп, нүктені ең жақын кластерге тағайындаймыз. Мысалы:

A оқушысы: $\mu1 = (8,5)$ дейінгі қашықтық $\mu2$ және $\mu3$ дейінгі қашықтықтан кіші, сондықтан A оқушысы бірінші кластерге жатады.

B оқушысы: $\mu1$ дейінгі қашықтық басқа центроидтарға қарағанда кіші, сондықтан B оқушысы да бірінші кластерге жатады.

C оқушысы: $\mu3$ дейінгі қашықтық минималды, сондықтан C оқушысы екінші кластерге жатады.

Барлық оқушыларды кластерлерге тағайындағаннан кейін, біз әрбір кластерге жаңа центроидтарды қайта есептейміз. Мысалы, егер A және B оқушылары бірінші кластерге жататын болса, жаңа центроид олардың дағдыларының орташа мәні ретінде есептеледі:

$$\mu_1 = \frac{(8,5) + (7,6)}{2} = (7.5, 5.5)$$

Екінші кластер үшін, егер C және D оқушылары сол топқа жататын болса:

$$\mu_2 = \frac{(1,8) + (2,9)}{2} = (1.5, 8.5)$$

Үшінші топқа E және F студенттері жатады:

$$\mu_3 = \frac{(9,1) + (8,2)}{2} = (8.5, 1.5)$$

Орталық центроидтарды жаңартқаннан кейін, біз студенттерді жаңа орталық центроидтарға негізделген кластерлерге қайтадан тағайындаймыз. Орталық центроидтар өзгермейінше қадамдар қайталанады. K-Means алгоритмін пайдалану арқылы біз студенттердің мансаптық қызығушылықтары мен дағдылары туралы деректерді әртүрлі кластерлерге тиімді түрде топтастыра аламыз, бұл бізге олардың болашақ мансаптық траекторияларын болжауға мүмкіндік береді. Бұл процесс мансаптық ұсыныстар жасау, еңбек нарығындағы танымал мансаптық жолдарды анықтау және жас мамандардың кәсіби өсуін оңтайландыруға пайдалы болуы мүмкін.

Мансап жолын талдауға K-орташа әдісін қолдану. K-орташа алгоритмі жас мамандарды білім деңгейі, тағылымдамадан өту тәжірибесі, техникалық және жұмсақ дағдылар, сондай-ақ кәсіби қызығушылықтары мен қалаулары сияқты әртүрлі сипаттамаларға негізделген топтастыруға қолданылады. Бұл әдісті қолдану арқылы ұқсас мансаптық күтулер мен дағдыларды бөлісетін бірнеше түлек тобын анықтауға болады. Мысалы, бір кластерге ақпараттық технологиялар саласында тәжірибесі бар мамандар кіруі мүмкін, ал екіншісіне маркетинг немесе PR саласындағы мансапқа бағытталған гуманитарлық ғылымдар дәрежесі бар түлектерден тұруы мүмкін (Levenson, 2017).

Бұл тәсіл әртүрлі мамандар топтарына ең ықтимал мансап жолдарын айқын визуализациялауға мүмкіндік береді. Кластерлік талдауды қолдану арқылы кәсіби тәжірибесі мен қалауларына байланысты негізгі мансаптық өсу бағыттарын бөлуге болады. Мысалы, техникалық білімі бар, бағдарламалаудың негізгі дағдыларына ие жас мамандар тобы IT саласына қатысты лауазымдарға назар аударуы мүмкін, ал келіссөздер жүргізу және командалық жұмыста білікті гуманитарлық ғылымдар дәрежесі бар түлектері жобаларды басқару немесе маркетинг саласындағы мансапқа бейімделуі мүмкін.

Мансаптық траекторияларға негізделген дағдыларды дамыту бойынша ұсыныстар. Жас мамандар туралы деректер кластерлерге бөлінгеннен кейін, әрбір топтың мансаптық траекториясына негізделген дағдыларды дамыту бойынша ұсыныстарды тұжырымдауға болады. Мысалы, егер кластер бағдарламалау саласындағы мансапқа бағытталған мамандар тобын анықтаса, олардың алгоритмдер, дерекқорлар мен бағдарламалық жасақтаманы әзірлеу сияқты салалардағы білімдерін нығайту ұсынылады. Екінші жағынан, маркетингтік мансапқа бағытталған түлектерге цифрлық маркетинг, брендті басқару және деректерді талдау саласындағы қосымша білім пайдалы болар еді (Ibrahim and Iqbal, 2025).

Сонымен қатар, кластерлік талдау әрбір топ үшін қандай нақты дағдыларға сұраныс жоғары екенін анықтауға көмектеседі. Мысалы, бағдарламалық жасақтаманы әзірлеу саласындағы мамандар тобына машиналық оқыту мен жасанды интеллект саласындағы білімдерін тереңдетуді ұсынуға болады, бұл оларға мансаптарында тезірек алға жылжуға мүмкіндік береді. Жобаларды басқару саласында мансаптық өсуді қалайтындарға Agile және Scrum

әдіснамалары саласындағы дағдылар, сондай-ақ көшбасшылық қасиеттер өте құнды болар еді (Cappelli and Keller, 2012).

K-орташа әдісінің мансаптық траекторияларды талдауға тартымды ететін бірнеше артықшылықтары бар. Біріншіден, бұл ең қарапайым әрі түсінікті алгоритмдердің бірі, бұл оны жұмыспен қамтуды қоса алғанда, әртүрлі салаларда қолдануға қолжетімді етеді. Екіншіден, ол үлкен деректер жиынтығымен жұмыс істеуге мүмкіндік береді, бұл мансап жолын жаңа бастаған көптеген жас мамандарды талдау кезінде өте маңызды. Үшіншіден, алгоритм мамандарды дәл топтастыруға мүмкіндік береді, бұл мансаптық даму бойынша жеке ұсыныстар жасауға және кәсіби траекторияларды құруға мүмкіндік береді.

Сонымен қатар, икемділігінің арқасында K-орташа әдісі мансаптың әртүрлі аспектілерін, соның ішінде табыс деңгейін, мансаптық өсуді және тіпті дамудың әртүрлі кезеңдеріндегі кәсіби күйіп кетуді болжау үшін пайдаланылуы мүмкін. Бұл әдіс мансаптық қалаулардың егжей-тегжейлі көрінісін береді және жас мамандарға жеке ұсыныстар жасауға мүмкіндік береді.

Сценарийлік зерттеу: Нақты деректермен тәжірибе. Зерттеуіміз үшін біз жас мамандардың топтарын олардың білім деңгейіне, негізгі дағдыларына, мақсатты салаларына және мансаптық қызығушылықтарына қарай анықтауға кластерлік талдауды қолдандық. Кластерлеу бізге мамандарды ұқсас мансаптық траекторияларымен топтарға бөлуге және олардың кәсіби дағдыларын жақсарту бойынша ұсыныстар жасауға мүмкіндік береді. Біздің экспериментімізге әрқайсысының әртүрлі салалардағы білім деңгейі, дағдылары мен қызығушылықтары әртүрлі 30 түлек қатысты. Талдауға біз деректерді әртүрлі топтарға анық бөлуге мүмкіндік беретін кластерлеуге өте қолайлы K-орташа алгоритмін қолдандық. Кластерлеу нәтижелері нақты деректерге емес, иллюстрациялық мақсат үшін модельденген деректерге негізделгенін атап өту маңызды. Бұл бізге кластерлік талдауды нақты әлемдегі жағдайда қалай қолдануға болатынын анық көрсетуге мүмкіндік береді.

1-қадам: Деректерді жасау

Зерттеуіміздің әрбір қатысушысына келесі параметрлер тағайындалды:

1. Білім – Қатысушының білім беру профилін анықтайды (мысалы, экономика, IT, дизайн).

2. Негізгі дағдылар – Ең өзекті дағдылар жиынтығы (мысалы, деректерді талдау, бағдарламалау, графикалық дизайн).

3. Мақсатты сала – Түлек жұмыс істегісі келетін сала (мысалы, қаржы, IT, медиа).

4. Мансаптық қызығушылықтар – Түлектің кәсіби тұрғыдан дамытқысы келетін бағыттары (мысалы, қаржылық талдау, бағдарламалық жасақтаманы әзірлеу, шығармашылық дизайн).

Кесте 2 – Қатысушыларға жасалған деректер

Білім саласы	Дағдылар	Мақсатты сала	Мансаптық қызығушылықтар	Кластер
Дизайн	Бағдарламалау	Қаржы	Деректерді талдау	0
Экономика	Деректер ғылымы	Медиа	Бағдарламалық қамтамасыз етуді әзірлеу	0
Дизайн	Деректер ғылымы	Медиа	Қаржылық талдау	2
Дизайн	Веб-дизайн	Ақпараттық технологиялар (IT)	Жобаларды басқару	2
Экономика	Веб-дизайн	Қаржы	Жобаларды басқару	0
Экономика	Веб-дизайн	Ақпараттық технологиялар (IT)	Жобаларды басқару	0
Дизайн	Графикалық дизайн	Ақпараттық технологиялар (IT)	Деректерді талдау	2
Ақпараттық технологиялар (IT)	Деректер ғылымы	Ақпараттық технологиялар (IT)	Қаржылық талдау	1
Дизайн	Деректер ғылымы	Ақпараттық технологиялар (IT)	Деректерді талдау	2
Дизайн	Аналитика	Ақпараттық технологиялар (IT)	Деректерді талдау	1

2-қадам: K-Means әдісін қолдану

K-means алгоритмін қолдана отырып, біз деректерді түлектердің мансаптық жолдарын көрсететін үш кластерге бөлдік. Кластерлеу нәтижелері үш негізгі топты анықтады:

1-кластер: Экономика және IT саласында тәжірибесі бар, қаржы және жобаларды басқаруға қызығушылық танытқан мамандар.

2-кластер: Дизайн және IT саласында арнайы білімі бар, шығармашылық және аналитикалық рөлдерге бағытталған түлектер.

3-кластер: Деректерді талдау және графикалық дизайн саласындағы дағдылары бар медиа және дизайн саласындағы шығармашылық лауазымдарға ұмтылатын мамандар.

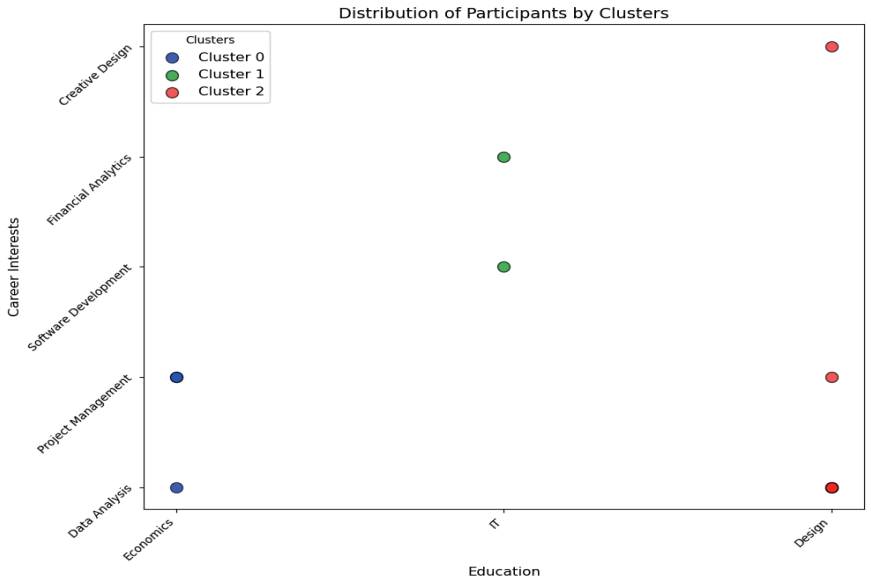
3-қадам: Нәтижелерді түсіндіру

1-кластер: Қаржы және менеджментке бағытталған экономистер мен IT мамандары. Бұл топ экономика және IT саласында білімі бар, аналитика және жобаларды басқаруға қызығушылық танытқан түлектерден тұрады. Бағдарламалау, аналитика мен веб-дизайн саласындағы дағдылары оларды IT және қаржы секторларындағы компаниялардағы қаржылық аналитиктер немесе жоба менеджерлері сияқты рөлдерге өте қолайлы етеді.

2-кластер: Шығармашылық қызығушылықтары бар IT және дизайн мамандары. Бұл кластердегі түлектердің веб-дизайн, графикалық дизайн және бағдарламалау саласында күшті дағдылары бар, сонымен қатар өз жұмыстарының шығармашылық аспектілеріне қызығушылық танытады. Олар медиа, қолданбаларды әзірлеу немесе UI/UX дизайны сияқты техникалық және шығармашылық элементтерді біріктіретін лауазымдарды табуы мүмкін.

3-кластер: Шығармашылық мансаптық мақсаттары бар дизайнерлер мен аналитиктер. Бұл кластерге графикалық дизайн саласындағы тәжірибесі бар және деректерді талдау мен дизайнға қызығушылық танытқан адамдар кіреді. Бұл түлектер медиа немесе аналитика немесе маркетинг сияқты деректерге қатысты рөлдер сияқты шығармашылық салаларда мансап құра алады. Бұл кластерлеу арқылы біз әр топтың бірегей дағдылары мен қызығушылықтарына бейімделген әлеуетті мансаптық жолдарды анықтай аламыз, бұл олардың мансаптық дамуына бағытталған ұсыныстар жасауға мүмкіндік береді.

4-қадам: Кластерлеу нәтижелерін визуализациялау. Кластерлеу нәтижелерін көрсетуге қатысушылардың білімі мен мансаптық қызығушылықтарына негізделген кластерлер бойынша таралуын көрсететін диаграмма құрайық.



Сурет 1 – Қатысушылардың кластерлер бойынша бөлінуі

Диаграммада қатысушылардың білімі мен мансаптық қызығушылықтарына негізделген үш кластер бойынша таралуы көрсетілген. Әрбір кластер әртүрлі түспен белгіленген, бұл ұқсас сипаттамалары бар топтарды ажыратуды жеңілдетеді. Бұл білім мен мансаптық қызығушылықтардың қалай байланысты екенін елестетуге көмектеседі.

5-қадам: Ұсыныстар мен қорытындылар

Кластерлік талдау нәтижелеріне сүйене отырып, біз әрбір кластер үшін мақсатты ұсыныстар жасай аламыз. 1-кластердегі жас мамандарға еңбек нарығында бәсекеге қабілетті болуға қаржылық талдау, жобаларды басқару және IT біліктілігін арттыру дағдыларын дамыту өте маңызды. 2-кластердегі мамандар дизайн және өнімді әзірлеу саласындағы білімдерін тереңдетуден, БАҚ немесе шығармашылық салалардағы мансаптық мүмкіндіктерге

портфолиосын кеңейтуден пайда көруі мүмкін. Сонымен қатар, 3-кластердегі қатысушылар аналитикалық дағдыларын жетілдіруге және UX/UI дизайны немесе маркетинг сияқты шығармашылықты технологиямен біріктіретін жобалармен жұмыс істеуге назар аудара алады.

Қорытындылай келе, кластерлік талдауды пайдалану тек қазіргі мансаптық жолдарды ашып қана қоймай, сонымен қатар кәсіби дағдыларды жақсарту және жас мамандарға перспективалы бағыттарды анықтау бойынша ұсыныстар береді. Кластерлік талдау мансаптық траекторияларды зерттеудің қуатты құралы болып табылады, зерттеушілерге жасырын үлгілерді және топтық деректерді ашуға мүмкіндік береді, бұл оны жас мамандардың мансаптық жолдарын талдауда құнды әдіске айналдырады. Бұл әдіс кең ауқымды, әртүрлі ақпаратты ұқсас сипаттамалары бар бірнеше кластерлерге немесе топтарға сегменттеуге мүмкіндік береді. Мансаптық траекториялар тұрғысынан кластерлік талдауды пайдалану зерттеушілерге де, мансаптық даму мамандарға да айтарлықтай артықшылықтар береді. Негізгі артықшылықтарының бірі - мансаптық қызығушылықтар, білім деңгейлері, дағдылары мен тәжірибесі туралы деректерді талдау мен жүйелеу мүмкіндігі, бұл мүмкін мансаптық жолдарды болжауға, қолайлы мамандықтарды ұсынуға және тіпті жұмыспен қамту процестерін оңтайландыруға көмектеседі.

Кластерлік талдаудың бірінші үлкен артықшылығы - деректердегі жасырын үлгілерді анықтау мүмкіндігі, олар бір қарағанда айқын көрінбеуі мүмкін. Жас мамандардың мансаптық траекторияларына келгенде, қарапайым статистикалық есептеулер немесе сызықтық модельдер сияқты дәстүрлі талдау әдістері мансаптық жолдардың әртүрлілігін толық қамтымауы мүмкін. Кластерлік талдау көптеген факторларды - білім, кәсіби қызығушылықтар, дағдылар, тәжірибе және тіпті әлеуметтік факторларды - ұқсас сипаттамалары бар топтарға «сұрыптай» алады. Бұл зерттеушіге әртүрлі параметрлердің бір-бірімен қалай әрекеттесетінін көруге мүмкіндік береді, бұл әлеуетті мансаптық жолдарды жан-жақты түсінуді ұсынады. Мысалы, ол кластерлеу әдісінсіз айқын болмас еді, белгілі бір жұмыс түрлеріне немесе салаларға бейім нақты дағдылары бар жас мамандар тобын ашуы мүмкін.

Екінші артықшылығы - үлкен көлемдегі деректерді өңдеу мүмкіндігі. Бүгінгі таңда жас мамандардың білімі, кәсіби дағдылары және мансаптық қалаулары әртүрлі. Кластерлік талдау түлектердің сауалнамалары, түйіндемелер немесе жұмыс дерекқорлары сияқты үлкен деректер жиынтығын өңдеу және әртүрлі айнымалылар арасындағы байланыстарды анықтауға өте қолайлы. Кластерлеу процесінде K-means сияқты алгоритмдерді білім деңгейі, дағдылары, тәжірибесі, мансаптық қызығушылықтары және т.б. сияқты сипаттамаларға негізделген деректерді топтастыруға пайдалануға болады. Бұл үлкен деректер жиынтығын талдауды айтарлықтай жеңілдетеді әрі жылдамдатады, әсіресе тек жеке тұлғалар туралы деректерді ғана емес, сонымен қатар оны динамикалық түрде өзгеретін еңбек нарығымен үйлестіру қажет болған кезде.

Кластерлік талдаудың негізгі аспектілерінің бірі - мақсатты топтарды түсінуді жақсарту мүмкіндігі. Жас мамандардың мансаптық траекторияларын зерттеу кезінде көбінесе жеке деректерді талдау ғана емес, сонымен қатар жұмыс берушілерге қызықты болуы мүмкін топтардың профильдерін жасау қажет. Мысалы, кластерлік талдау арқылы нақты салаларда қандай дағдыларға сұраныс көп екенін және әртүрлі салаларда қандай кәсіби профильдер ізделетінін анықтауға болады. Бұл ұқсас мансаптық қызығушылықтары бар мамандар тобын анықтауға көмектеседі, жас мамандардың әртүрлі санаттарыға жұмысқа орналасуды дәлірек анықтауға көмектеседі. Нәтижесінде жұмыс берушілер өз лауазымдарына лайықты кандидаттарды тезірек әрі дәлірек таба алады, ал мамандар мансаптық бағыттар мен жұмысқа орналасудың мүмкін нұсқаларын анық түсінеді. Кластерлік талдаудың тағы бір маңызды артықшылығы - оны мансаптық траекторияны болжауға қолдану (Zhang and Gu, 2025). Кластерлеу ұқсас сипаттамалары бар адамдар тобын анықтай алады және осы топтарға сүйене отырып, болашақта оларға қандай мамандықтар немесе салалар ең перспективалы болатынын болжайды. Мысалы, IT мамандарына еңбек нарығының қазіргі үрдістеріне сүйене отырып, алдағы жылдары қандай нақты мамандықтарға сұраныс көп болатынын болжауға болады. Бұл жас мамандарға болашақта қажет болатын дағдыларды дамыту арқылы мансап жолдары туралы неғұрлым хабардар шешім қабылдауға мүмкіндік береді, осылайша олардың табысты жұмысқа орналасу мүмкіндігін арттырады. Мансаптық траекторияларды болжау білім беру мекемелеріне де құнды құрал бола алады, олар өз бағдарламаларын еңбек нарығының қажеттіліктеріне жақсырақ сәйкестендіру және түлектердің бәсекеге қабілеттілігін арттыруға бейімдей алады.

Қорытынды. Кластерлік талдаудың икемділігі мен әмбебаптығы да артықшылықтар ретінде ерекшеленеді. Бұл әдісті тек мансаптық жолдарды талдауға ғана емес, сонымен қатар деректерді талдаудың кең ауқымды міндеттерін шешуге де қолдануға болады. Мысалы, оны еңбек нарығын сегменттеуге, HR саласындағы тұтынушылардың профильдерін жасауға, бос жұмыс орындарын талдауға және жұмыс берушілерге ұсыныстар әзірлеуге қолдануға болады. Кластерлік талдауды әртүрлі деректер түрлері мен сценарийлеріне бейімдеуге болады, бұл оны HR саясаты, маркетинг, еңбек нарығын талдау және басқа да салаларды қоса алғанда, әртүрлі салаларда қолданылатын әмбебап құралға айналдырады. Географиялық орналасуы, тілдік дағдылары немесе әртүрлі компаниялардағы тәжірибесі сияқты көптеген факторлар мен айнымалыларды ескере отырып, егжей-тегжейлі мамандардың профильдерін жасауға болады, бұл дәлірек болжамдарға әкеледі. Сондай-ақ, кластерлік талдау жалдау процесін айтарлықтай оңтайландыра алатынын атап өткен жөн. Қатысушыларды топтарға бөлу арқылы зерттеушілер мен жұмыс берушілер мамандардың дағдыларын жұмыс талаптарымен тиімдірек сәйкестендіре алады. Кластерлеу жұмыс берушілерге негізгі дағдылар мен үміткерлердің қалауын жақсы анықтауға

көмектеседі, ал жұмыс іздеушілер қандай кәсіби топтар мен салалар оларға ең тартымды болуы мүмкін екендігі туралы түсінік ала алады. Бұл тәсіл тек жұмысқа қабылдау процесін жақсартып қана қоймай, сонымен қатар жұмыс берушілер мен қызметкерлерге қанағаттанушылықтың жоғары деңгейін қамтамасыз етеді. Жұмыс берушілер қажетті біліктілікке ие үміткерлерді дәл әрі тез анықтай алудан пайда көреді, ал жас мамандар өздерінің дағдылары нарықтық сұраныстарға сәйкес келетінін біле отырып, жұмыста өздерін сенімді сезінеді (Levenson, 2017).

Кластерлік талдаудың тағы бір маңызды артықшылығы - басқа әдістерді қолдана отырып, байқалмауы мүмкін екіұшты немесе онша айқын емес топтарды анықтау мүмкіндігі. Мысалы, жас мамандардың мансаптық траекторияларын талдау кезінде кластерлік талдау тек аналитика немесе бағдарламалық жасақтаманы әзірлеу сияқты дәстүрлі мансаптық жолдарды ғана емес, сонымен қатар жаңа салалардағы жаңа технологиялық немесе тар лауазымдардағы сұранысқа ие рөлдер сияқты жаңа үрдістерді де көрсете алады. Бұл талдау ауқымын кеңейтеді және алдағы жылдары қандай мамандықтардың танымал болатынын болжауға ғана емес, сонымен қатар жас мамандардың жаңа дағдыларды игеру немесе қызығушылықтарын өзгерту арқылы мансаптық жолдарын қалай қайта бағдарлай алатынын көрсетуге көмектеседі.

Қорытындылай келе, кластерлік талдау - білім беру, мансаптық қызығушылықтар мен жұмысқа орналасу мүмкіндіктері арасындағы байланысты түсінуді айтарлықтай жақсарта алатын мансаптық траекторияларды талдаудың тиімді құралы. Бұл әдістің артықшылықтары оның үлкен көлемдегі деректерді өңдеу, жасырын үлгілерді ашу, мансаптық жолдарды болжау, жас мамандарға дағдыларды дамыту бойынша ұсыныстар беру қабілетінде. Бұл әдіс жұмыспен қамту процестерін оңтайландыруға ғана емес, сонымен қатар мансаптық траекторияны тиімдірек жоспарлауға мүмкіндік береді, жас мамандардың үнемі өзгеріп отыратын еңбек нарығына бейімделуін арттырады.

References

Banerjee P., & Sharma N. (2024) Digital transformation and talent management in industry 4.0: A systematic literature review and future directions. The Learning Organization, 32(4). — P. 620–640. DOI: 10.1108/tlo-10-2023-0183 (in Eng.)

Bishop C.M. (2006) Pattern recognition and machine learning. Springer. (in Eng.)

Breiman L. (2001) Random forests. Machine Learning, 45. — P. 5–32. DOI: 10.1023/A:1010933404324 (in Eng.)

Cappelli P., & Keller J. (2012) Classifying work in the new economy. Academy of Management Review, 38(4). — P. 575–596. DOI: 10.5465/amr.2011.0302 (in Eng.)

Friedman T.L. (2016) Thank you for being late: An optimist's guide to thriving in the age of accelerations. Farrar, Straus and Giroux. (in Eng.)

Ibrahim I., & Iqbal N.M. (2025) Career pattern analysis of SMKN 1 Stabat graduates using K-means clustering algorithm on tracer study dataset. Journal of Data Science, 3(1). — P. 43–57. DOI: 10.58471/jds.v3i01.6543 (in Eng.)

Levenson A. (2017) Using workforce analytics to improve strategy execution. Human Resource Management, 57(3). — P. 685–700. DOI: 10.1002/hrm.21850 (in Eng.)

Li Y., & Zhao L. (2025) Application of support vector machine algorithm in predicting the career development path of college students. International Journal of High Speed Electronics and Systems. DOI: 10.1142/S012915642540230X (in Eng.)

Maciariello J.A. (2018) Drucker, Peter: The Drucker strategic management system. In Palgrave Macmillan UK eBooks. — P. 431–440. DOI: 10.1057/978-1-137-00772-8_629 (in Eng.)

McKinsey Global Institute. (2020). The future of work after COVID-19. McKinsey & Company. https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19 (in Eng.)

Mohrman S.A., & Lawler E.E. (2012) Generating knowledge that drives change. Journal of Applied Behavioral Science, 48(2). — P. 197–222. DOI: 10.5465/amp.2011.0141 (in Eng.)

Noack B. (2019) Big data analytics in human resource management: Automated decision-making processes, predictive hiring algorithms, and cutting-edge workplace surveillance technologies. Psychosociological Issues in Human Resource Management, 7(2), 37. DOI: 10.22381/PIHRM7220196 (in Eng.)

Pereira O., & Raposo M. J. (2019) Soft skills in knowledge-based economics. Marketing and Management of Innovations. — P. 182–195. DOI: 10.21272/mmi.2019.1-15 (in Eng.)

Rigas B. (2021) Trends shaping education and innovative learning environments: A discourse analysis of OECD CERI projects. http://hdl.handle.net/10464/15063 (in Eng.)

Venkatesh V., Thong J.Y.L., & Xu X. (2012) Consumer acceptance and use of information technology. MIS Quarterly, 36(1). — P. 157–178. DOI: 10.2307/41410412 (in Eng.)

Zhang L., & Gu X. (2025) A novel career prediction method based on fuzzy model: Fuzzy clustering approach. Acta Psychologica, 259. — 105475 p. DOI: 10.1016/j.actpsy.2025.105475 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 167–181

https://doi.org/10.32014/2026.2518-1726.407

IRSTI: 28.23.01
UDC 004.855

© Dzhsupbekova G.[1], Othman M.[2], Ordabayeva G. [3,4*], 2026.

[1]M. Auezov South Kazakhstan State University, Shymkent, Kazakhstan;
[2]Universiti Putra Malaysia, Selangor D.E., Malaysia;
[3]Bolashak Program, Scientific internship, UC Davis, USA;
[4]Kazakh National University named after Al-Farabi, Almaty, Kazakhstan.
E-mail: gordabayeva@ucdavis.edu

COMPARATIVE ANALYSIS OF ARTIFICIAL INTELLIGENCE ALGORITHMS TO DETECT NETWORK ATTACKS

Dzhsupbekova Gulzat — Candidate of Pedagogical Sciences, Head of the Department of Information Technology, M. Auezov South Kazakhstan State University, Shymkent, Kazakhstan,
E-mail: gulzat20.10@mail.ru, ORCID ID: https://orcid.org/0000-0003-1727-0966;
Mohamed Othman — Professor, Department of Communication Technology and Networks, Universiti Putra Malaysia, Serdang, Malaysia,
E-mail: mothman@upm.edu.my, ORCID ID: https://orcid.org/0000-0002-5124-5759;
Ordabayeva Gulzinat — Bolashak Program, Scientific internship, UC Davis, USA; Senior lecturer at the Department of Cybersecurity and Cryptology, Kazakh National University named after Al-Farabi, Almaty, Kazakhstan,
E-mail: gordabayeva@ucdavis.edu, ORCID ID: https://orcid.org/0000-0001-9952-1620;

Abstarct. The rapid development of information technologies and the expansion of digitalization trends have a huge impact on all spheres of life of society and the economy. The complication of network infrastructures and the sharp increase in the number of devices connected to the internet have increased the volume of information flow, as well as created new security challenges. The types and methods of cyberattacks are changing rapidly, and traditional methods of protection, as before, are often insufficient. Large enterprises are using artificial intelligence (AI) systems that can independently learn patterns and analyze complex data streams instead of traditional rule-based systems. These methods allow you to detect attacks faster and more accurately because they can quickly detect patterns and anomalies. The purpose of the article is to conduct a comparative analysis based on deep learning methods to accurately and quickly identify threats in network traffic and prove its effectiveness in practice. To analyze the potential of machine learning (ML) and deep learning (DL) methods, traditional algorithms such as Random

Forest, XGBoost, and SVM were employed together with a convolutional neural network (CNN) architecture. Model training was conducted in Python using the CIC-IDS-2017 dataset. The data were stored in * parquet format and underwent preprocessing, including the removal of missing values, scaling of numerical features, and encoding of categorical variables. Attack types were encoded using predefined labels, and scaling was performed using StandardScaler. The dataset was split into 70% for training and 30% for testing. Artificial intelligence algorithms have been relatively analyzed in terms of classifying attack types, distinguishing between normal and abnormal traffic, as well as detecting new or rare attacks. As a result of the study, indicators of accuracy, sensitivity, F1-score, ROC-AUC curve and error matrix were evaluated. In addition, the real-time application of the proposed methods and their practical significance were analyzed.

Keywords: network attack, artificial intelligence, deep learning, machine learning, neural network, types of attacks, comparative analysis

For citations: Dzhsupbekova G., Othman M., Ordabayeva G. Comparative analysis of artificial intelligence algorithms to detect network attacks. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 167–181. DOI: https://doi.org/10.32014/2026.2518-1726.407

© Джусупбекова Г.[1], Othman M.[2], Ордабаева Г.[3,4*], 2026.
[1]М. Ауэзов атындағы Оңтүстік Қазақстан университеті,
Шымкент, Қазақстан;
[2]Putra университеті, Selangor D.E., Малайзия;
[3]«Болашақ» бағдарламасы, ғылыми тағылымдама, UC Davis, USA;
[4]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.
E-mail: gordabayeva@ucdavis.edu

ЖАСАНДЫ ИНТЕЛЛЕКТ АЛГОРИТМДЕРІН ЖЕЛІЛІК ШАБУЫЛДАРДЫ АНЫҚТАУ ҮШІН САЛЫСТЫРМАЛЫ ТАЛДАУ

Джусупбекова Гулзат — педагогика ғылымдарының кандидаты, «Ақпараттық-коммуникациялық технологиялар» кафедрасының меңгерушісі, М. Ауэзов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан,
E-mail: gulzat20.10@mail.ru, ORCID ID: https://orcid.org/0000-0003-1727-0966;
Mohamed Othman — Communication Technology and Networks кафедрасының профессоры, Putra университеті, Серданг, Малайзия,
E-mail: mothman@upm.edu.my, ORCID ID: https://orcid.org/0000-0002-5124-5759;
Ordabayeva Gulzinat — "Болашак" бағдарламасы, ғылыми тағылымдама, UC Davis, USA; «Киберқауіпсіздік және криптология» кафедрасының аға оқытушысы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: gordabayeva@ucdavis.edu, ORCID ID: https://orcid.org/0000-0001-9952-1620.

Аннотация. Қазіргі кезеңде ақпараттық технологиялардың жедел дамуы және цифрландыру үрдістерісінің кең таралуы қоғамның, экономиканың және адам өмірінің көптеген салаларына елеулі ықпал етуде. Желілік инфрақұрылымдардың күрделене түсуі және интернетке қосылған құрылғылар санының қарқынды өсуі ақпараттық ағын көлемінің ұлғаюына әкеліп қана қоймай, сонымен қатар ақпараттық қауіпсіздік саласында жаңа қауіп-қатерлердің пайда болуына себеп болды. Кибершабуылдардың түрлері мен жүзеге асыру әдістері үнемі өзгеріп, жетілдіріліп отырғандықтан, дәстүрлі қорғаныс тәсілдері қазіргі сәтте көптеген жағдайларда жеткіліксіз болып отыр. Бүгінде ірі ұйымдарда дәстүрлі ережелерге негізделген жүйелердің орнына деректерден заңдылықтар мен үлгілерді өздігінен үйрене алатын және күрделі ақпараттық ағындарды талдайтын жасанды интеллект (AI) технологиялары кеңінен қолданылуда. Мұндай тәсілдер кибершабуылдарды неғұрлым жылдам әрі дәл анықтауға мүмкіндік береді, өйткені олар деректердегі жасырын үлгілер мен аномалияларды тиімді түрде айқындай алады. Бұл мақаланың мақсаты – желілік трафиктегі қауіп-қатерлерді дәл әрі жедел анықтауға бағытталған терең оқыту әдістеріне негізделген модельдеоге салыстырмалы талдау жасап, олардың тиімділігін тәжірибелік зерттеу нәтижелері арқылы көрсету. Зерттеу барысында жасанды интеллект негізіндегі бірнеше әдіс сыналды. Машиналық оқыту (МО) және терең оқыту (Deep Learning) алгоритмдерінің мүмкіндіктері зерттеле отырып, Random Forest, XGBoost, SVM сияқты дәстүрлі әдістермен қатар нейрондық желілердің заманауи архитектурасы – CNN қолданылды. Барлық модельдер CIC-IDS-2017 деректер қорын қолдана отырып Python негізінде үйретілді. Ақпараттар *.parquet форматында сақталынды және алдын ала өңделді, яғни бос мәндер жойылды, сандық шкалаға келтірілді, категориялар кодталды. Шабуыл түрлері сандық кодтау арқылы белгіленіп, деректер StandardScaler көмегімен масштабталды. Деректер жиыны 70% оқу және 30% тестілеу бөліктеріне бөлінді. Жасанды интеллект алгоритмдері шабуылдарды жіктеу, қалыпты және аномальды трафикті анықтау, сондай-ақ жаңа немесе сирек шабуылдарды табу қабілеті бойынша салыстырмалы түрде бағаланды. Зерттеу нәтижесінде дәлдік, сезімталдық, F1-score, ROC-AUC қисығы және қателік матрицасы көрсеткіштері бағаланды. Бұдан бөлек, ұсынылған әдістердің нақты уақыт режимінде қолданылу мүмкіндіктері мен олардың практикалық тиімділігіне талдау жүргізілді.

Түйін сөздер: желі шабуылы, жасанды интеллект, терең оқыту, машиналық оқыту, нейрондық желі, шабуыл түрі, салыстырмалы талдаулар

©Джусупбекова Г.[1], Othman M.[2], Ордабаева Г.[3,4*], 2026.
[1]Южно-Казахстанский университет имени М. Ауэзова, Шымкент, Казахстан;
[2]Университет Putra, Selangor D.E., Малайзия;
[3]Программа «Болашак», научная стажировка, UC Davis, USA;
[4]Казахский Национальный университет имени ал-Фараби,
Алматы, Казахстан.
E-mail: gordabayeva@ucdavis.edu

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ОБНАРУЖЕНИЯ СЕТЕВЫХ АТАК

Джусупбекова Гулзат – кандидат педагогических наук, зав. кафедрой «Информационно-коммуникационные технологии», Южно-Казахстанский университет имени М. Ауэзова, Шымкент, Казахстан,
E-mail: gulzat20.10@mail.ru, ORCID ID: https://orcid.org/0000-0003-1727-0966;
Mohamed Othman – профессор кафедры Communication Technology and Networks, Университет Putra, Серданг, Малайзия,
E-mail: mothman@upm.edu.my, ORCID ID: https://orcid.org/0000-0002-5124-5759;
O**rdabayeva Gulzina**t – Программа “Болашак”, научная стажировка, UC Davis, USA; ст.преп. кафедры «Кибербезопасность и криптология», Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: gordabayeva@ucdavis.edu, ORCID ID: https://orcid.org/0000-0001-9952-1620.

Аннотация. Стремительное развитие информационных технологий и расширение процессов цифровизации оказывают значительное влияние на все сферы жизни общества и экономики. Усложнение сетевой инфраструктуры и рост числа устройств, подключённых к сети Интернет, приводят к увеличению объёма трафика и появлению новых угроз информационной безопасности. В современных условиях в крупных организациях всё шире применяются системы искусственного интеллекта (ИИ), способные автоматически выявлять закономерности и анализировать сложные потоки данных. Такие методы обеспечивают более быстрое и точное обнаружение атак за счёт выявления скрытых зависимостей и аномалий в сетевом трафике. Целью исследования является сравнительный анализ методов машинного и глубокого обучения, направленных на эффективное выявление сетевых атак, а также экспериментальная оценка их результативности. В работе рассмотрены как классические алгоритмы машинного обучения (Random Forest, XGBoost, SVM), так и современные нейросетевые архитектуры, включая сверточные нейронные сети (CNN). Модели обучались с использованием языка программирования Python на основе набора данных CIC-IDS-2017. Предварительная обработка данных включала удаление пропущенных значений, масштабирование числовых признаков с использованием StandardScaler, кодирование категориальных переменных, а также преобразование данных в формат *.parquet. Данные

были разделены на обучающую и тестовую выборки в соотношении 70% и 30% соответственно. Сравнительный анализ алгоритмов проводился по ряду критериев, включая точность классификации, способность различать нормальный и аномальный трафик, а также выявление редких и ранее неизвестных атак. В качестве метрик оценки использовались accuracy, recall (чувствительность), F1-score, ROC-AUC и матрица ошибок. Результаты исследования показали, что методы глубокого обучения демонстрируют более высокую эффективность при выявлении сложных паттернов сетевых атак, в то время как классические алгоритмы обеспечивают устойчивые результаты при меньших вычислительных затратах.

Ключевые слова: сетевая атака, искусственный интеллект, глубокое обучение, машинное обучение, нейронная сеть, типы атак, сравнительный анализ

Кіріспе. Желілік шабуыл – ұйымның ақпараттық желісіне рұқсатсыз кіру арқылы деректерді заңсыз алу немесе зиян келтіру мақсатындағы әрекет. Желілік шабуылдардың ең жиі кездесетін түрлері – пассивті және белсенді шабуылдар. Пассивті шабуыл кезінде шабуылдаушы желі трафигін тыңдап немесе ақпаратты ұрлап алады, бірақ деректерді өзгертпейді, яғни олардың бастапқы күйін сақтайды. Белсенді шабуылдарда рұқсатсыз қатысушылар жүйеге заңсыз кіріп, деректерді жоюға, шифрлауға немесе бүлдіруге мүмкіндік алады.

Кең таралған желілік шабуылдар сипаттамалары:

- Таратылған қызмет көрсетуден бас тарту (Distributed Denial of Service, DDoS) шабуылдары. Зиянкестер ботнеттерді – басқараылатын зиянды құрылдғылар желісін қолданып серверлерге немесе желілерге жалған трафик жібереді. DDoS-шабуылдар желілік деңгейде (серверді шамадан тыс жүктейтін SYN/ACK пакеттерінің легі) немесе қолданбалы деңгейде (дерекқорларды бұзуға бағытталған күрделі SQL-сұраныстар арқылы) жүзеге асыруы мүмкін (Aljabri et al., 2021; Singh et al., 2022).

- MITM (Man-In-The-Middle) шабуылдарды зерттеулерде (Salem et al., 2022) IoMT жүйелерінде қашықтан денсаулық мониторингін бұзудан қорғау және шабуылдардың ықпалын төмендету жолдару ретінде талданған. Деректердің құпиялылығын қамтамасыз ету үшін физиологиялық көрсеткіштер орнына LSH-қолтаңбасы енгізілді, ал олардың тұтастығын сақтау мақсатында RSSI мәніне негізделген HMAC кілті қолданылды.

- Кодтық және SQL инъекция шабуылдары. Көптеген веб-сайттар пайдаланушы енгізген деректерді өңдейді, алайда оларды дұрыс тексеріп, зарарсыздандырмауы мүмкін. Осы жағдайларда зиянкестер веб-формалар мен API сұраныстарын пайдалана отырып, жүйеге зиянды код енгізуге қабілетті. Осы код серверде іске қосылған жағдайда жүйенің осалдықтарын пайдаланып, дерекқорға кіруге немесе оны басқаруға мүмкіндік береді (Tadhani et al., 2024.

- Артықшылықтарды кеңейту (Privilege Escalation) кезінде шабуылдаушы-

лар желіге кіргеннен соң өз мүмкіндіктерін арттыру мақсатында сәйкес әдістерді қолданады. (Muslam, 2023) зерттеуінде инсайдерлік шабуылдарды анықтау және жіктеу үшін машиналық оқыту алгоритмдерін ұсынады және талдауда бірнеше Cert деректер жиынтығы қолданылды.

- Ішкі қауіптер (Insider Threats) – ұйымдағы артықшылықты рұқсаттарға ие қызметкерлер тарапынан туындайтын қауіптер, олар желілік инфрақұрылымның осалдығын арттыра алады. (Ye et al., 2025) өз зерттеулерінде Cert деректер жиынтығын конвульсиялық нейрондық желі (CNN) негізінде инновациялық архитектура ұсынды. Зерттеу нәтижелері көрсеткендей, бұл тәсіл корпоративтік ортада инсайдерлік қауіптерді анықтаудың дәлдігі мен сенімділігін елеулі түрде жақсартады.

Материалдар мен әдістер. Қазіргі уақытта МО алгоритмдері үлкен деректер жиынтығын өңдеп, ықтимал қауіптерді көрсететін үлгілер мен аномалияларды анықтауға, сондай-ақ күрделі кибершабуылдарды тануға қабілетті.

МО алгоритмдеріне негізделген жүйелер деректерді нақты уақыт режимінде өңдеуге мүмкіндік беріп, киберқауіптерді жылдам анықтау және оларға оперативті жауап беру арқылы кибершабуылдардың ықтимал зиянын төмендетуге ықпал етеді.

Машиналық оқытуға негізделген мінез-құлықты терең талдау пайдаланушының қалыпты белсенділік үлгілерін тиімді анықтауға мүмкіндік береді. Қалыпты әрекет үлгілерінен кез келген ауытқулар қауіпсіздікке қатер төндіруінің ықтимал белгісі ретінде қарастырылады, бұл зиянды әрекеттерді ерте анықтап, алдын алу мүмкіндігін қамтамасыз етеді. Болжамдық талдау әдістері өткен деректерді талдай отырып, МО арқылы ықтимал болашақ қауіптер мен жүйелік осалдықтарды алдын ала болжауға мүмкіндік береді. Бұл белсенді қорғаныс әдістері ұйымдарға шабуылдардың алдын алу мақсатында қауіпсіздік шараларын алдын ала күшейтуге мүмкіндік береді (Zou, 2025).

Әдеби шолу

ЖИ даму кезеңдері

ЖИ әдістері желілік шабуылдарды алдын ала анықтау және олардың нақты классификациясын жүргізуге мүмкіндік береді. Бұл тәсілдер шабуылдардың индикаторларын ерте тануға және оларға жылдам әрі тиімді түрде әрекет етуге мүмкіндік береді.

ЖИ алгоритмдері МО әдістерін пайдаланып желідегі әрекет үлгілерін меңгеріп, жаңа қауіп-қатерлерге тиімді бейімделуге қабілетті. ЖИ жүйелері инциденттерге автоматты түрде әрекет ете отырып, киберқауіптерді басқаруда мамандардың уақытын тиімді пайдалануға жағдай эжасайды.

ЖИ эволюциясы және киберқауіпсіздікке ықпалы:

1. Алғашқы кезең (1900-1950 жж.) – ережеге негізделген жүйелердің дамуы.

2. Өтпелі кезең (1950–1980 жж.). Статистикалық әдістер және алғашқы МО модельдері..

3. Қалыптасу кезеңі (1980-2020 жж.) – үлкен деректер және МО модельдерінің жетілу кезеңі.

4. Қазіргі кезең (2021 – қазіргі уақытқа дейін) – терең оқыту және өзіндік үйрену қабілеті бар жүйелердің қалыптасуы мен жетілдірілуі.

Киберқауіпсіздік саласында терең оқыту (көпқабатты нейрондық желілер) кеңінен қолданылады. Өздігінен үйренетін ЖИ жүйелері, мысалы, Darktrace және Vectra AI, нақты уақытта желідегі әрекеттерді талдап, шабуылдарды автоматты түрде анықтайды. Сонымен қатар, бұл технологиялар бұлттық қауіпсіздік, IoT құрылғыларды қорғау, блокчейн қауіпсіздігі және API трафигін бақылау сияқты жаңа бағыттарды дамытуға мүмкіндік береді (Asambaev, 2011; Mucci, 2024; Muthukrishnan, 2020).

Классикалық IDS/IPS жүйелері сигнатураларға негізделгендіктен, жаңа немесе белгісіз шабуылдарды тиімді анықтай алмайды. Бұл кемшілікті ЖИ негізіндегі жүйелер жоя алады. ЖИ технологиялары шабуылдарда тануда келесі әдістерді қолдануға мүмкіндік береді:

1. Зиянды трафикті анықтау – ЖИ алгоритмдері желілік трафикті автоматты түрде талдай отырып, қалыпты және зиянды мінез-құлық үлгілерін тиімді ажырата алады. Терең нейрондық желілер (DNN) және рекуренттік нейрондық желілер (RNN) желілік сессиялардағы уақыттық заңдылықтарды идентификациялауға арналған. NSL-KDD немесе CICIDS2017 секілді датасеттерді пайдалана отырып, ЖИ жүйесі шабуылдарды (DoS, Probe, R2L, U2R) нақты ажырата алады. Мұнда әрбір пакет немесе байланыс сеансы белгіленген сипаттамалар (features) бойынша өңделеді.

2. Аномалияларды анықтау – ЖИ көмегімен "қалыптыдан ауытқыған" әрекеттерді табу. Мұнда бақылаусыз оқыту (unsupervised learning) әдістері, K-means, DBSCAN, Autoencoder кеңінен қолданылады. Бұл әдіс жаңа шабуыл түрлерін анықтауға мүмкіндік береді, себебі олар әлі белгілі шаблондарға сәйкес келмеуі мүмкін.

3. Кластерлеу және жіктеу - ЖИ шабуылдарды түрлеріне қарай жіктеу үшін SVM, CNN, Random Forest сияқты алгоритмдерді қолданады. Алдын ала оқытылған үлгілер негізінде жүйелер трафиктің зиянды немесе қалыпты екенін анықтайды.

4. DDoS шабуылдарын ерте анықтауда ЖИ пакеттер ағынын нақты уақыт режимінде талдай отырып, шабуылдың бастапқы белгілерін анықтай алады. LSTMсияқты терең оқыту модельдері уақытша тәуелділіктерді зерттеу арқылы шабуыл басталмай тұрып есекерту сигналдарын береді (Zhunisov, 2024).

ЖИ модельдерін оқытуда қолданылатын дереккөздер

Желілік шабуылдарды анықтауға арналған ЖИ жүйелерін тиімді үйрету үшін нақты және репрезентативті деректер жиынтықтары қажет. Қазіргі зерттеулерде кеңінен қолданылатын деректер жиынтықтарына мыналар жатады:

- KDD Cup 99 – классикалық кибершабуылдарды қамтитын деректер;

- NSL-KDD – KDD Cup 99 деректер жиынтығының жетілдірілген және жаңартылған нұсқасы;

- CSE-CIC-IDS2017 – заманауи шабуыл түрлерін және толық метаақпаратты қамтитын деректер жиынтығы;

- UNSW-NB15 – қалыпты және зиянды трафиктің теңгерімді үлгілерін қамтитын деректер жиынтығы.

Зерттеуде негізгі дереккөз ретінде желілік шабуылдарды анықтауда жиі қолданылатын CIC-IDS-2017 таңдалды. Бұл дерекқор құрамында зертханалық ортада құрастырылған нақты желі трафигі және қалыпты белсенділігі бар түрлі шабуылдар бар. Мәліметтер flow-based features типтес желі легімен берілген. Мұнда қосылу ұзақтығы, пакеттер саны, байт сипаттамалары және уақыт параметрлері сияқты желілік өзара әрекеттесудің статистикалық сипаттамасы бар. Деректер қоры құрамында бірнеше кластар бар және деректер көп кластық бойынша жіктеледі: Benign (қалыпты трафик), DDoS, DoS (Hulk, GoldenEye, Sloworis, Slowhttptest), PortScan, Bot, FTP-Patator, SSH-Patator, Web Attack (Brute Forse, XSS), Infiltration.

Аталған деректер жиынтықтары негізінде ЖИ модельдерін оқытып, оларды нақты желілік шабуылдарды тануға үйретуге болады. ЖИ жүйелерінің жұмыс тиімділігін бағалау үшін келесі негізгі метрикалар қолданылады:

- Accuracy – дұрыс классификацияланған жағдайлардың жалпы үлесі;
- Precision – шабуыл ретінде анықталған әрекеттердің ішінде нақты шабуыл болғандарының үлесі;
- Recall (Sensitivity) – барлық нақты шабуылдардың ішінде жүйе дұрыс анықталғандарының үлесі;
- F1-score – precision және recall көрсеткіштерінің гармониялық орташа мәні (Khan et al., 2021).

ЖИ әлемдік компанияларда қолданылуы

1. Google компаниясының Gmail қызметіндегі фишингтік шабуылдарды анықтау мақсатында ЖИ және МО алгоритмдері кеңінен қолданылады. Мысалы, Gmail жүйесі әрбір кіріс хатты жүздеген параметр бойынша талдайды, оның ішінде мәтін құрылымы, тақырып мазмұны, сілтемелердің сипаты, IP мекенжай мәліметтері және басқа сипаттамалар бар. Осындай талдау нәтижесінде жүйе күн сайын 100 миллионнан астам фишингтік хабарламаны автоматты түрде бұғаттай алады (Google AI Blog, 2023).

2. IBM компаниясы дамытқан QRadar жүйесі ақпараттық қауіпсіздік орталықтарында (Security Operation Center, SOC) кеңінен қолданылады. Аталған жүйе ЖИ технологияларын пайдалана отырып, желідегі аномалияларды, рұқсатмыз әрекеттерді және ықтимал шабуылдарды нақты уақыт режимінде анықтайды (IBM Security, 2022).

3. Darktrace - Ұлыбританияда құрылған киберқауіпсіздік саласындағы компания. Ол желілік шабуылдарды анықтау үшін ЖИ негізделген өзіндік оқытылатын жүйені пайдаланады. Аталған жүйе ұйымның ішкі желісіндегі қалыпты әрекеттерді талдап, олардан ауытқыған кез келген аномалды белсенділікті ықтимал қауіп ретінде анықтап, жедел хабарлама жібереді. Darktrace технологиялары әлемнің 100-ден астам елінде қолданылып, қаржы, денсаулық сақтау және өндіріс секторларында кеңінен енгізілген (Darktrace, 2025).

4. Kaspersky компаниясы зиянды бағдарламалардың жаңа түрлерін анықтау мақсатында DL және үлгіге негізделген (pattern-based) талдау әдістерін пайдаланады. Бұл тәсіл компанияға жыл сайын миллиондаған жаңа кибершабуыл түрлерін автоматты түрде анықтап, оларды жоюға мүмкіндік береді (Kaspersky, 2024).

5. Қазақстан Республикасында 2017 жылдан бастап жүзеге асырылып келе жатқан «Киберқалқан» бағдарламасы аясында ақпараттық қауіпсіздік жүйелеріне жасанды интеллект енгізілуде. Аталған жүйелер мемлекеттік органдар мен ұлттық инфрақұрылымды кибершабуылдардан қорғау мақсатында желілік трафикті үздіксіз мониторингтен өткізіп, шабуылға тән аномалияларды автоматты түрде анықтайды. Нәтижесінде жүйе мемлекеттік деректер базаларына бағытталған шабуыл әрекеттерінің алдын алуға мүмкіндік береді (Киберқалқан, 2025).

Нәтижелер

Модельді бағалаудағы классификация метрикалары

ЖИ модельдерін желілік шабуылдарды анықтау және жіктеу үшін қолданғанда, ең басты міндет – модельдің тиімділігін дәл бағалау. Нақты уақытта қолданылатын модель тек болжам жасап қана қоймай, сол болжамның сенімді және қауіпсіз екенін де қамтамасыз етуі қажет.

ЖИ мен МО әдістері негізінде құрылған желілік шабуылдарды анықтау модельдерінің сапасын дұрыс бағалау – жүйенің сенімділігі мен тиімділігін қамтамасыз етудің негізгі құрамдас бөлігі. Бағалау критерийлері мен метрикалар шабуылдарды нақты және толық анықтауға мүмкіндік беріп қана қоймай, сонымен қатар жалған позитивтер (false positives) мен жалған негативтерді (false negatives) де өлшеуге мүмкіндік береді.

Көбінесе модельді бағалауда бинарлық классификация метрикалары қолданылады, себебі көптеген жүйелер қалыпты және шабуыл трафигін ажыратуға бағытталған. Егер шабуылдардың түрлерін бөлу қажет болса, көпкласты классификация метрикалары да қолданылады, атап айтқанда:

1. CIC-IDS-2017 деректер қорындағы кластарды талдау (1-кесте).

1 кесте – Қалыпты трафиктің және шабуыл түрлерінің саны (CIC-IDS-2017 дерек қоры)

Р/н	Класс атауы	Саны
1	Benign	2273097
2	DoS Hulk	231073
3	PortScan	158930
4	DDoS	128027

5	DoS GoldenEye	10293
6	FTP-Patator	7938
7	SSH-Patator	5897
8	DoS Sloworis	5796
9	DoS Slowhttptest	5499
10	Bot	1966
11	Web Attack Brute Forse	1507
12	Web Attack XSS	652
13	Infiltration	36

Қарастырылған деректер қорында басым көпшілікті Benigh класы құрайды. Сонымен қатар, ең төменгі мәлімет Infiltration класына тиесілі. Бұл теңгерімсіздік МО модельдерін құру кезінде ескерілген.

2. Дәлдік (Accuracy) – модельдің барлық болжамдар ішінде дұрыс болжам жасаған жағдайларының үлесін сипаттайтын метрика (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Дәлдік (Accuracy) тек теңгерілген (balanced) деректерде ғана толық ақпарат береді. Егер шабуылдардың саны аз болса, модель барлық трафикті «қалыпты» деп бағалап, жоғары дәлдік көрсеткіші алуы мүмкін. Алайда бұл жағдай модельдің шынайы тиімділігін көрсетпейді, себебі нақты шабуылдарды анықтау қабілеті төмен болады.

3. Нақтылық (Precision) – модельдің шабуыл деп белгіленген трафигінің ішінде шынайы шабуылдардың үлесін көрсететін метрика (2):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Егер модель көп мөлшерде FP (жалған оң) жағдайлар тудырса, жүйеге сенім төмендейді. Сол себепті нақтылық (Precision) маңызды метрика болып табылады, себебі ол модельдің қате ескертулерін бағалауға мүмкіндік береді.

4. Recall немесе Detection Rate - модель нақты шабуылдардың қаншасын дұрыс анықтағанын көрсетеді (3).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

Бұл көрсеткіш әсіресе қауіпсіздік саласында аса маңызды,өйткені FN (жіберіп алған шабуылдар) жүйеге үлкен зиян келтіруі мүмкін.

5. F1-score - нақтылық пен сезімталдықтың үйлесімі, екі метриканың арасындағы тепе-теңдікті көрсетеді (4).

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

F1-score теңгерімсіз деректерде (шабуыл үлесі 10%-дан аз) өте маңызды. Ол модельдің жалпы тиімділігін жақсы сипаттайды.

Берілген формулалар негізінде тұрғызылған МО модельдерін талдау 2 кестеде берілген.

2 кесте – МО модельдерін салыстыру

Модель атауы	Accuracy	Precision	Recall	F1-Score
CNN	0.953111	0.950174	0.953111	0.949316
Random Forest	0.995111	0.994960	0.995111	0.994887
SVM	0.945778	0.943968	0.945778	0.942091

Жалпы алғанда, Random Forest алгоритмі салыстырмалы талдауда жоғары көрсеткішке ие болды.

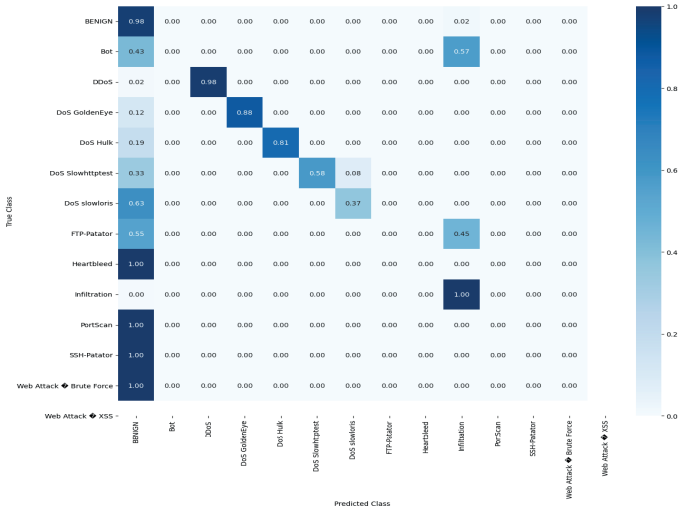
Модельдерді оқыту алгоритмдерін салыстырмалы талдау

Келесі МО және DL модельдеріне салыстырмалы талдау жүргізілді:

- Random Forest – ансамльдік ағаштар әдісі;
- SVM – қолдау векторлар машиналары;
- CNN – конволюциялық нейрондық желі.

Барлық модельдер CIC-IDS-2017 датасеті негізінде үйретілді. Мәліметтер MachineLearningCSV.csv форматында сақталып, алдын ала өңделді (бос мәндер жойылды, сандық шкалаға келтірілді, категориялар кодталды). Шабуыл түрлері сандық кодтау арқылы көрсетілді, ал ерекшеліктер StandardScaler негізінде масштабталды. Деректер оқыту және тест жиынтығына бөлінген, мұнда оқыту жиынтығы деректердің 70% , ал тест жиынтығы 30% қамтиды.

Кластар бойынша қателік матрицасы CNN алгоритмінің тест деректерінде әр шабуыл класын қаншалықты дұрыс немесе бұрыс анықтағанын көрсетеді (2 - сурет).



2 сурет – Кластар бойынша қателік матрицасы

Қателік матрицасының нәтижелеріне сүйенсек, CNN моделі желілік шабуылдарды жіктеуде жоғары тиімділік көрсетті. Талдау көрсеткендей Benign, DoS Hulk, DDoS Portscan кластары жоғары дәлдікті көрсететті, мұнда диагональдағы мәндер 0,85-0,98 деңгейінде болады. Сонымен қатар, кейбір кластар үшін жіктеу қателері бар. Графикте Bot класы Benign және басқа шабуылдаушы сыныптармен шатастырылады. Сол сияқты, DoS Slowloris және DoS Slowhttptest кластары өзара қателіктерді көрсетеді. Жалпы модель қалыпты желілік трафикті сенімді түрде ажырата алатыны көрсетілді.

Талқылау

(AL-Jaberi et al., 2025) зерттеулерінде МО алгоритмдеріне арналған жаңа мұздату әдісі ұсынылды және пирамида сұлбасын пайдалана отырып, фишингтік веб-сайттарды анықтаудың жетілдірілген тәсілі жасалды. Жаңа модель ең маңызды белгілерге назар аударып, фишинг сайттарды дәл анықтауды қамтамасыз етеді және қайта оқытуды қажет етпейді. Ұсынылған модельде 30 атрибуты бар 2456 ақпараттан тұратын деректер қоры негізінде Random Forest алгоритмі фишингтік сайттарды анықтауды 97,4% дәлдікте көрсетті.

(Cárdenas-Haro et. al., 2025) өз зерттеулерін Sybil шабуылына қатысты әдеби шолудан бастаған. Мақалада Sybil шабуылына осал желілер қауіпсіздігін қамтамасыз етудегі қорғаныс шаралары және шабуылдың алдын-алу алгоритмдері қарастырылған. Онлайн әлеуметтік желілерде, соның ішінде Twitter-де Sybil байланысын анықтау алгоритмі зерттеудің негізі болып табылады. Sybil шабуылын KNN алгоритмі 96% - 97%, *SVM* алгоритмінің көрсеткіші 97% - 100% дәлдікпен анықтады. Зерттеуде қорғаныс шараларын қолданудың қадамдары, нәтижелерді талдау және болашақтағы зерттеу бағыттары баяндалған.

(Kuraś et.al., 2026) зерттеуінде IoT-құрылғыларындағы шектеулі ресурстар жағдайында желі трафигіндегі аномальды әсерді бақылау мақсатында МО модельдерін оңтайландыруды қарастыра отырып RandomForest, ExtraTrees, AdaBoost, XGBoost, CatBoost алгоритмдеріне салыстырмалы талдау жасалған. Қорытынды да XGBoost алгоритмі бинарлы классификация негізінде 92% дәлдікпен 8 секунд арасында МО оңтайлы моделі екенін көрсетті.

(Altwaijry et al., 2024) зерттеулерінде фишингтік электрондық пошта хабарларын анықтау үшін бірөлшемді конволюциялық нейрондық желілерге (1D-CNNPD) негізделген модельдер қолданылды. Базалық 1D-CNNPD модель рекуренттік қабаттармен – LSTM, Bi-LSTM, GRU және Bi-GRU – күшейтілгенде оның анықтау дәлдігі артатыны дәлелденген. Зерттеу барысында фишингті анықтауға арналған конволюциялық нейрондық желі (CNN) негізіндегі 12 DL моделі ұсынылған. Талдау нәтижесінде Leaky ReLU және Bi- GRU моделінің негізінде өнімділік көрсеткіштері Acc.99.68%, F1 99.66% тең болды. Бұл көрсеткіш 1D-CNNPD моделінің электрондық пошта арқылы фишингтік шабуылдармен күресу үшін киберқауіпсіздік шешімдерін іске асырудағы әлеуетін көрсетеді.

(Bolatbek et. al., 2024) зерттеулерінде деструктивті хабарламалардың таралуын YouTube, Vkontakte және Telegram сияқты әлеуметтік желілерден ақпарат жинау арқылы қарастырған. Мақалада e uni-bi-gram(1,2) көмегімен Logistic Regression, SVM, Naive Bayes әдістерінен алынған нәтижелердің жиынтығы көрсетілген. Сонымен қатар, фишингтік веб-сайттарды анықтау, желі қауіпсіздігін қамтамасыз ету және шабуылдарды алдын алу, сондай-ақ әлеуметтік желілердегі деструктивті хабарламалардың таралуын болжау үшін МО модельдері мен DL алгоритмдері қарастырылған.

Біздің зерттеуімізде желілік шабуылдарды анықтауда CNN моделі негізіндегі көрсеткіш 98,13% дәлдікпен дұрыс анықтады.

(Altwaijry et. al., 2024) зерттеуінде LSTM, Bi-LSTM, GRU және Bi-GRU қабаттарымен кеңейтілген кездегі өнімділік көрсеткіштері 99.68%, ал біз қарастырған Random Forest, XGBoost, SVM орташа нәтижелері 98%, яғни бұл тұрғызылған модель сенімділігінің жоғарылығын көрсетеді. Желілік талдауда Benign (қауіпсіз трафик) класы, DoS Hulk шабуылы, DDoS шабуылы, DoS GoldenEye шабуылдары толықтай ажыратылды және модельдің тиімділігін көрсетті.

Қорытынды. Мақалада желілік шабуылдарды анықтау және оларды жіктеу мәселесі зерттелді. Осы мақсатта ЖИ негізделген бірнеше МО және DL алгоритмдері – Random Forest, SVM және CNN қолданылып, олардың тиімділігі салыстырмалы түрде талданды.

Зерттеу нәтижелері келесі негізгі **қорытындыларды** көрсетті:

- Деректерді алды ала өңдеу кезінде маңызды белгілерді іріктеу модельдердің жалпы өнімділігіне айтарлықтай әсер ететіні анықталды. Масштабтау және деректерді сандық форматқа келтіру сияқты процедуралар, әсіресе SVM және NN модельдерінің жұмысында маңызды рөл атқарды;

- SVM алгоритмі кейбір шабуыл түрлерін анықтауда жоғары нәтижелер көрсеткенімен, деректер көлемі ұлғайған сайын оның есептеу жылдамдығының төмендейтіні байқалды. Бұл фактор оны нақты уақыт режимінде жұмыс жасайтын жүйелерде қолдануды белгілі бір деңгейде шектейді;

- CNN моделі шабуылдардың күрделі құрылымдық ерекшеліктерін тиімді меңгеріп, жақсы нәтижелер көрсетті. Дегенмен, мұндай модельдерді оқыту үшін жоғары есептеу ресурстары мен қосымша уақыт керек.

Зерттеу нәтижелері көрсеткендей, желілік шабуылдарды анықтау барысында бір ғана алгоритмді қолдану жеткіліксіз болуы мүмкін. Себебі шабуылдардың сипаты, деректердің көлемі мен құрылымы өзгерген жағдайда әртүрлі модельдер әртүрлі тиімділік көрсетеді. Сондықтан болашақта нақты міндеттерге байланысты бірнеше модельді біріктіру немесе ансамбльдік тәсілдерді қолдану арқылы жүйенің сенімділігін арттыру мүмкіндігі бар.

ЖИ-ке енгізделген шешімдерді желілік қауіпсіздік саласына енгізу цифрлық инфрақұрылымды қорғаудың маңызды бағыттарының бірі болып табылады. Мұндай тәсілдер желілік аномалияларды ерте кезеңде автоматты түрде анықтауға және кибершабуылдардың алдын алуға мүмкіндік береді.

Осы бағыттағы зерттеулер болашақ интеллектуальды қауіпсіздік жүйелерінің қалыптасуына негіз болады және қосымша ғылыми зерттеулер жүргізуді талап етеді.

Сонымен қатар, зерттеу барысында ЖИ-ке негізделген МО және DL алгоритмдерінің мүмкіндіктері қарастырылды. Атап айтқанда, Random Forest, SVM сияқты дәстүрлі әдістермен қатар, NN-дің заманауи архитектурасының бірі – CNN қолданылды. Бұл алгортимдер шабуыл түрлерін жіктеу, қалыпты және аномальды желілік трафикті ажырату, сондай-ақ жаңа немесе сирек кездесетін шабуылдарды анықтау мүмкіндіктері тұрғысынан салыстармалы түрде талданды.

References

Al-Jaberi A.J.K., Kurnaz S., Naser R.A.S., et al. (2025) A novel architecture based on weight freezing and random forest for website phishing detection. *International Journal of Computational Intelligence Systems*, 18. — 284 p. https://doi.org/10.1007/s44196-025-00975-5 (in Eng.)

Aljabri M., Aljamee S.S., Mohammad R.M.A., Almotiri S.H., Mirza S., Anis F.M., Aboulnour M., Alomari D.M., Alhamed D.H., & Altamimi H.S. (2021) Intelligent techniques for detecting network attacks: Review and research directions. *Sensors*, 21. — 7070 p. https://doi.org/10.3390/s21217070 (in Eng.)

Altwaijry N., Al-Turaiki I., Alotaibi R., & Alakeel F. (2024) Advancing phishing email detection: A comparative study of deep learning models. *Sensors*, 24. — 2077 p. https://doi.org/10.3390/s24072077 (in Eng.)

Asambaev A.Zh. (2011) *Zhasandy intellekt negizderi: Okulyk* [Fundamentals of artificial intelligence: Textbook]. Almaty: Dauir. (in Kaz.)

Bolatbek M., Saginai M., & Musiralieva Sh. (2024) Using machine learning methods to detect destructive web content in the Kazakh language. *Academic Scientific Journal of Computer Science*, (4). — P. 99–111. https://doi.org/10.32014/2024.2518-1726.310 (in Eng.)

Cárdenas-Haro J.A., Salem M., Aldaco-Gastélum A.N., López-Avitia R., & Dawson M. (2024) Enhancing security in social networks through machine learning: Detecting and mitigating Sybil attacks with SybilSocNet. *Algorithms*, 17. — 442 p. https://doi.org/10.3390/a17100442 (in Eng.)

Cyber Shield (2025) Access: https://www.gov.kz/memleket/entities/knb/activities/250?lang=kk (in Eng.)

Darktrace (2025) *AI and cybersecurity: Predictions for 2025*. https://www.darktrace.com/blog/ai-and-cybersecurity-predictions-for-2025 (in Eng.)

IBM Security (2022) *Cost of a data breach report*. https://www.ibm.com/security/data-breach (in Eng.)

Issa M.M., Aljanabi M., & Muhialdeen H.M. (2024) Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. *Journal of Intelligent Systems*, 33(1). — 20230248 p. https://doi.org/10.1515/jisys-2023-0248 (in Eng.)

Kaspersky (2024) *Kaspersky predicts AI and privacy to shape consumer cybersecurity landscape in 2025*. https://www.kaspersky.com/about/press-releases/kaspersky-predicts-ai-and-privacy-to-shape-consumer-cybersecurity-landscape-in-2025 (in Eng.)

Khan M.A., & Usman M. (2021) A deep learning approach for intrusion detection using LSTM and GRU. *Sensors*, 21(15). — 5112 p. https://doi.org/10.3390/s21155112 (in Eng.)

Kuraś P., Bolanowski M., & Łoza M. (2026) Optimization of machine learning models for effective anomaly detection in industrial IoT systems. *Advances in Science and Technology Research Journal*, 20(1). — P. 203–221. https://doi.org/10.12913/22998624/210686 (in Eng.)

Matias Y. (2025) *Google Research 2025: Bolder breakthroughs, bigger impact*. https://research.google/blog/google-research-2025-bolder-breakthroughs-bigger-impact/ (in Eng.)

Muslam M. (2023) Privilege escalation attack detection and mitigation in cloud using machine learning. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3273895 (in Eng.)

Muthukrishnan N., Maleki F., Ovens K., Reinhold C., Forghani B., & Forghani R. (2020) Brief history of artificial intelligence. *Neuroimaging Clinics of North America*, 30(4). — P. 393–399. https://doi.org/10.1016/j.nic.2020.07.004 (in Eng.)

Salem O., Alsubhi K., Shaafi A., Gheryani M., Mehaoua A., & Boutaba R. (2022) Man-in-the-middle attack mitigation in Internet of Medical Things. *IEEE Transactions on Industrial Informatics*, 18(3). — P. 2053–2062. https://doi.org/10.1109/TII.2021.3089462 (in Eng.)

Singh A., & Gupta B.B. (2022) Distributed denial-of-service (DDoS) attacks and defense mechanisms in various web-enabled computing platforms: Issues, challenges, and future research directions. *International Journal on Semantic Web and Information Systems*, 18(1). — P. 1–43. https://doi.org/10.4018/IJSWIS.297143 (in Eng.)

Tadhani J.R., Vekariya V., Sorathiya V., et al. (2024) Securing web applications against XSS and SQLi attacks using a novel deep learning approach. *Scientific Reports*, 14. — 1803 p. https://doi.org/10.1038/s41598-023-48845-4 (in Eng.)

Ye X., Luo F., Cui H., et al. (2025) Research on insider threat detection based on personalized federated learning and behavior log analysis. *Scientific Reports*, 15. — 19214 p. https://doi.org/10.1038/s41598-025-04029-w (in Eng.)

Zhou Y. (2025) Design and implementation of computer network security monitoring system. *Journal of Computer, Signal, and System Research*, 2(4). — P. 63–68. https://doi.org/10.71222/qkantx75 (in Eng.)

Zhunisov N.M., Aben A.B., & Isakov D. (2024) Detection of network attacks using a comparative analysis of machine learning algorithms. *Mechanics and Technologies*, 4(86). — P. 430–439. https://doi.org/10.55956/ZCJD4515 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 182–201

https://doi.org/10.32014/2026.2518-1726.408

IRSTI 27.35.21; 27.35.47; 27.41.19
UDC 533.9.01; 519.63; 519.684

© Issakhov A.[1]*, Orazmoldayev N.[1], Zharkynbek Y.[2], Abylkassymova A.[1], 2026.
[1] Kazakh-British Technical University, Almaty, Kazakhstan;
[2] King Abdullah University of Science & Technology, Thuwal, Saudi Arabia.
E-mail: alibek.issakhov@gmail.com

NUMERICAL MODELING OF THE SPREAD OF VIRAL INFECTION BY AIRBORNE DROPLETS IN CONFINED SPACES

Issakhov Alibek — PhD, Professor, Kazakh-British Techinical University, Almaty, Kazakhstan,
E-mail: alibek.issakhov@gmail.com, https://orcid.org/0000-0002-1937-8615;
Orazmoldayev Nurbol — Master's student, Kazakh-British Techinical University, Almaty, Kazakhstan,
E-mail: nurbol.orazmoldayev@gmail.com, https://orcid.org/0009-0004-2444-7621;
Zharkynbek Yeskendir — Master's student, King Abdullah University of Science & Technology, Thuwal, Saudi Arabia,
E-mail: yeskendir.zharkynbek@gmail.com, https://orcid.org/0009-0001-0616-567X;
Abylkassymova Aizhan — PhD, Associated Professor, Kazakh-British Techinical University, Almaty, Kazakhstan,
E-mail: Abylkassymova.Aizhan@gmail.com, https://orcid.org/0000-0002-5967-6959.

Abstract. Airborne virus transmission is a key mechanism for the spread of infectious diseases in enclosed spaces, where asymptomatic carriers are especially dangerous, which can increase the rate of infection foci formation. The behavior of aerosol particles in the air is determined by the complex dynamics of air flows, thermal effects, and characteristics of ventilation systems, which makes the problem relevant for developing effective strategies to prevent indoor contamination. In this paper, a numerical study of the propagation of droplets and virus particles in ventilated rooms is performed based on the solution of three–dimensional Navier-Stokes equations for incompressible flows in combination with the discrete phase (DPM) model and the turbulent SST k–ω model. The developed mathematical model takes into account the influence of human body temperature, gravitational deposition, room geometry, location of ventilation openings, as well as the behavior of particles during breathing, coughing and sneezing. Two room configurations and two particle ejection rates are considered (6, 20 m/s), which makes it possible to

cover a wide range of real-world scenarios. The simulation results demonstrate that particles can spread over a distance of more than 5-6 meters, and improper or uneven ventilation leads to the formation of zones of increased aerosol concentration. The nonlinear effect of the human heat plume and ventilation design on the dynamics of air flow, precipitation and particle transport is shown. The data obtained can be used to optimize ventilation systems, design safe indoor spaces, and develop protocols to reduce the risk of indoor transmission of infections. Computational Fluid Dynamics (CFD) provides valuable information for substantiating safety protocols and reducing the risk of airborne transmission, which makes the proposed approach particularly relevant for practical applications.

Key words: particles, enclosed space, SST k-omega, viral diseases

For citations: Issakhov A., Orazmoldayev N., Zharkynbek Y., Abylkassymova A. Numerical modeling of the spread of viral infection by airborne droplets in confined spaces. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 182–201. DOI: https://doi.org/10.32014/2026.2518-1726.408

© Исахов А.А.[1]*, Оразмолдаев Н.[1], Жаркынбек Е.[2], Абылкасымова А.[1], 2026.

[1]Қазақ-Британ техникалық университеті, Алматы, Қазақстан;

[2] Сауд Арабиясының Король Абдулла атындағы ғылым және технология университеті, Тувал, Сауд Арабия.

E-mail: Alibek.issakhov@gmail.com

АУА ТАМШЫЛАРЫ АРҚЫЛЫ ВИРУСТЫҚ ИНФЕКЦИЯНЫҢ ШЕКТЕУЛІ КЕҢІСТІКТЕ ТАРАЛУЫН САНДЫҚ МОДЕЛЬДЕУ

Исахов Алибек — PhD, Профессор, Қазақ-Британ техникалық университеті, Алматы, Қазақстан,

E-mail: alibek.issakhov@gmail.com, https://orcid.org/0000-0002-1937-8615;

Оразмолдаев Нұрбол — магистрант, Қазақ-Британ техникалық университеті, Алматы, Қазақстан,

E-mail: nurbol.orazmoldayev@gmail.com, https://orcid.org/0009-0004-2444-7621;

Жаркынбек Ескендір — магистрант, Сауд Арабиясының Король Абдулла атындағы ғылым және технология университеті, Тувал, Сауд Арабия,

E-mail: yeskendir.zharkynbek@gmail.com, https://orcid.org/0009-0001-0616-567X;

Абылкасымова Айжан — PhD, қауымдастырылған профессор, Қазақ-Британ техникалық университеті, Алматы, Қазақстан,

E-mail: Abylkassymova.Aizhan@gmail.com, https://orcid.org/0000-0002-5967-6959.

Аннотация. Вирустардың ауа-тамшылары арқылы таралуы инфекция ошақтарының қалыптасу жылдамдығын арттыруға қабілетті, асимптоматикалық тасымалдаушылар аса қауіпті жабық үй-жайларда жұқпалы аурулардың таралуының негізгі механизмі болып табылады. Ауадағы аэрозоль бөлшектерінің әрекеті ауа ағындарының күрделі динамикасымен,

жылу әсерлерімен және желдету жүйелерінің сипаттамаларымен анықталады, бұл мәселені ішкі инфекцияның алдын алудың тиімді стратегияларын әзірлеу үшін өзекті етеді. Бұл жұмыста дискретті фазалық модельмен (DPM) және SST K–ω турбулентті моделімен біріктірілген сығылмайтын ағындар үшін үш өлшемді Навье–Стокс теңдеулерін шешу негізінде желдетілетін бөлмелерде тамшылар мен вирустық бөлшектердің таралуын сандық зерттеу жүргізілді. Әзірленген математикалық модель адамның дене температурасының, гравитациялық тұндырудың, бөлме геометриясының, желдеткіш саңылаулардың орналасуының және тыныс алу, жөтелу және түшкіру кезіндегі бөлшектердің әрекетінің әсерін ескереді. Бөлменің екі конфигурациясы және екі бөлшектердің шығарылу жылдамдығы қарастырылады (6, 20 м/с). Бұл нақты сценарийлердің кең ауқымын қамтуға мүмкіндік береді. Модельдеу нәтижелері бөлшектердің 5-6 метрден астам қашықтыққа таралуы мүмкін екенін көрсетеді, ал дұрыс емес немесе біркелкі емес желдету аэрозоль концентрациясының жоғарылау аймақтарының пайда болуына әкеледі. Адамның жылу шлейфі мен желдету құрылымының ауа ағынының динамикасына, тұндыруға және бөлшектердің тасымалдануына сызықтық емес әсері көрсетілген. Нәтижелер желдету жүйелерін оңтайландыру, қауіпсіз ішкі кеңістіктерді жобалау және үй ішінде инфекциялардың таралу қаупін азайту хаттамаларын әзірлеу үшін пайдаланылуы мүмкін. Есептеу гидродинамикасы (CFD) қауіпсіздік хаттамаларын негіздеу және ауа арқылы жұқтыру қаупін азайту үшін құнды ақпарат береді, бұл ұсынылған тәсілді практикалық қолдануға ерекше маңызды етеді.

Түйін сөздер: бөлшектер, жабық бөлме, SST k-omega, вирустық аурулар

© Исахов А.А.[1]*, Оразмолдаев Н.[1], Жаркынбек Е.[2], Абылкасымова А.[1], 2026.

[1]Казахстанско-Британский технический университет, Алматы, Казахстан;
[2] Университет науки и технологий имени короля Абдаллы, Тувал, Саудовская Аравия
E-mail: Alibek.issakhov@gmail.com

ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ РАСПРОСТРАНЕНИЕ ВИРУСНОЙ ИНФЕКЦИИ ВОЗДУШНО-КАПЕЛЬНЫМ ПУТЕМ В ЗАМКНУТЫХ ПОМЕЩЕНИЯХ

Исахов Алибек — PhD, Профессор, Казахстанско-Британский технический университет, Алматы, Казахстан,
E-mail: alibek.issakhov@gmail.com, https://orcid.org/0000-0002-1937-8615;
Оразмолдаев Нұрбол — магистрант, Казахстанско-Британский технический университет, Алматы, Казахстан,
E-mail: nurbol.orazmoldayev @gmail.com, https://orcid.org/0009-0004-2444-7621;
Жаркынбек Ескендир — магистрант, Университет науки и технологий имени короля Абдаллы, Тувал, Саудовская Аравия,
E-mail: yeskendir.zharkynbek @gmail.com, https://orcid.org/0009-0001-0616-567X;

Абылкасымова Айжан — PhD, ассоциированный профессор, Казахстанско-Британский технический университет, Алматы, Казахстан,
E-mail: Abylkassymova.Aizhan@gmail.com, https://orcid.org/0000-0002-5967-6959.

Аннотация. Воздушно-капельная передача вирусов является одним из ключевых механизмов распространения инфекционных заболеваний в закрытых помещениях, где особую опасность представляют бессимптомные носители, способные ускорять формирование очагов заражения.Поведение аэрозольных частиц определяется сложной динамикой воздушных потоков, тепловыми эффектами и характеристиками вентиляционных систем, что обуславливает актуальность разработки эффективных стратегий предотвращения внутрипомещенного заражения. В работе выполнено численное моделирование распространения капель и вирусных частиц в вентилируемых помещениях на основе решения трёхмерных уравнений Навье–Стокса для несжимаемых течений в сочетании с моделью дискретной фазы (DPM) и турбулентной моделью SST k–ω. Разработанная математическая модель учитывает влияние температуры тела человека, гравитационного осаждения частиц, геометрии помещения, расположения вентиляционных отверстий, а также режимов дыхания, кашля и чихания. Рассмотрены две конфигурации помещения и два режима скорости выброса частиц (6 и 20 м/с), что позволяет охватить широкий диапазон реальных сценариев. Результаты моделирования показывают, что частицы способны распространяться на расстояние более 5–6 метров, а неравномерная или неэффективная вентиляция приводит к формированию зон повышенной концентрации аэрозолей. Выявлено нелинейное влияние теплового шлейфа человека и параметров вентиляции на динамику воздушных потоков, перенос и осаждение частиц. Полученные результаты могут быть использованы для оптимизации систем вентиляции, проектирования безопасных внутренних пространств и разработки мер по снижению риска передачи инфекций в помещениях. Методы вычислительной гидродинамики (CFD) предоставляют важную информацию для обоснования санитарно-гигиенических рекомендаций и повышения эффективности систем вентиляции.

Ключевые слова: частицы, замкнутое помещение, SST k-omega, вирусные заболевания.

Introduction. In the last decades, it became evident that understanding and modelling the spreading of airborne droplet particles, such as viruses, is of great importance to public health and safety. The COVID-19 pandemic clearly demonstrated the significance of studying the aerosol ways of infection transmission to develop efficient prevention strategies. The global statistics of COVID-19 are as follows: the number of infected exceeds 775 million people, and more than 7 million people have passed away (World Health Organization WHO, 2023). Among the symptoms of virus, the most common are: an increase in

body temperature to 38° C, significant fatigue and loss of sense of smell and taste. One of the foremost aspects of such research is the modelling of the behavior and dynamics of particles in air flows, which requires accounting for multiple factors, such as turbulence, lift forces, and inter-particle interaction. The primary mode of disease transmission is recognized as droplets containing viral particles expelled by an infected individual through breathing, coughing, and sneezing (Bipasha et al., 2015). Data from the study indicate that a healthy individual typically coughs twice and sneezes four times daily (Hansen et al., 2002; Surinder et al., 2006). SARS-CoV-2, the coronavirus SARS-CoV-2, as many other viruses, are mainly transmitted through the airborne droplet route (Busco et al., 2020). The primary transmission mechanisms include:

The SARS-CoV-2 coronavirus, like many other viruses, is transmitted mainly through airborne droplets. The main transmission mechanisms include:

Drip transmission: Large droplets released when coughing, sneezing or talking may contain viral particles. These droplets settle on the surface within a few meters of the source.

Aerosol transmission: Fine particles (aerosols), also released when breathing, talking or coughing, can remain suspended in the air for a long time and travel long distances, especially in conditions of insufficient ventilation. In fact, the speed of reflex processes can vary significantly, averaging between 1 - 22 m/s (Gao et al., 2007; Gupta et al., 2009).

Contact transmission: The virus can be transmitted by touching contaminated surfaces and then to mucous membranes (eyes, nose, mouth) (Wolfer et al., 2020; Pan et al., 2019).

The main focus of this work is on the study of the behavior of aerosol particles containing SARS-CoV-2 virus particles. In particular, three basic physical models were used: the Saffman lifting force, stochastic collision, and the particle rupture model. Each of these models plays an important role in accurately describing the dynamics of particles in turbulent flows.

The Saffman lifting force takes into account the effect of the velocity gradient on particles, which is especially important for small particles in streams with high shear rates. Stochastic collision describes the probabilistic interactions between particles, and the rupture model takes into account the process of breaking large particles or droplets into smaller ones under the action of turbulence and shear forces.

To accomplish this task, modern numerical modeling methods were used, including discrete phase models (DPM) (Pirker et al., 2011) and turbulence models. The use of such models makes it possible to obtain a detailed picture of particle propagation under various conditions, as well as to assess the influence of various physical factors on this propagation.

The thesis also considers the initial and boundary conditions used in modeling, including temperature and particle composition, which allows us to reproduce the real conditions of the spread of viral particles in the air as accurately as possible.

The ventilation rate of the room, along with both the individual's body temperature and the ambient room temperature, exert significant influence on the characteristics of infectious aerosols (Mutlu et al., 2020). The human body temperature was set at 37°C (310.15 K), which was taken into account in the simulation to accurately reproduce the thermal effect on air flows and, consequently, on the propagation of particles.

To assess the nature of the flow, the Reynolds number was calculated, which made it possible to determine the flow regime and take into account its influence on the behavior of particles. The turbulent nature of the air flow plays a key role in the spread of particles, especially in confined spaces where viral particles can remain suspended in the air for a long time. Therefore, proper modeling of turbulence and its effect on the trajectory and distribution of particles is a critically important aspect of this study.

Various numerical methods, including the finite element method and the finite volume method, have also been used to better understand and predict the behavior of particles in air flows. These methods make it possible to solve complex systems of equations describing the motion and interaction of particles in a multicomponent medium.

Literature Review. Computational fluid dynamics (CFD) was used to study the formation and spread of viruses. Experimental investigation of such processes presents significant difficulties, therefore CFD offers a valuable tool for modeling the distribution and spread of viral droplets that occur during reflex processes such as coughing, sneezing and breathing. Numerical modeling in the study (Hamid et al., 2021), various ventilation options were considered in order to influence ventilation on the transfer of pathogen droplets. The authors of the work (Mengfan et al., 2022) numerically analyzed the spread of droplets during coughing in different seasons of time indoors. It was concluded that the rate of droplet transmission accelerates at low temperature and the exhalation height of droplets plays an important role in droplet transmission. The study (Aliyu et al., 2021) considered airborne transport with and without masks in ventilated rooms. As a result, in cases of wearing masks, more than 96% of all discarded droplets were trapped within the recommended social distancing radius of 2 m. Without wearing masks, 80% of the drops were trapped, and 20% of the drops got into the air and spread. CFD was used in the above-mentioned works and the authors of these works suggest using CFD for such tasks, namely particle propagation.

The purpose of this thesis is to develop and apply an integrated approach to modeling the aerosol spread of viruses, which can make a significant contribution to understanding the mechanisms of transmission of infections and developing measures to prevent them. The results of the study can be used to improve ventilation systems, develop more effective disinfection methods and reduce the risk of transmission of infections in public places.

Materials and methods

Mathematical model. A mathematical model has been developed to study

the spread of viral particles in airborne droplets in a confined spaces. To simulate the airflow and particle dynamics within a confined indoor environment, the incompressible Navier–Stokes equations were employed as the governing equations for fluid motion:

$$\nabla u = 0, \tag{1}$$

$$\rho \left(\frac{\partial u}{\partial t} + (\nabla u)u\right) = -\nabla p + \mu \nabla^2 u - \rho \beta g (T - T_0), \tag{2}$$

$$\frac{\partial \mathrm{T}}{\partial \mathrm{t}} + u_j \frac{\partial \mathrm{T}}{\partial x_j} = D_a \left(\frac{\partial^2 T}{\partial x_j^2}\right) + \frac{S_T}{\rho c_p}, \tag{3}$$

Where u – velocity vector, $(\nabla u)u$ – the convective term describing the change in velocity due to fluid motion, ρ – density, p – pressure, μ – effective dynamic viscosity, β – thermal expansion coefficient, g – acceleration due to gravity, T – temperature, T_0 – initial temperature, D_a - thermal diffusivity, S_T- heat source, c_p - specific heat capacity.

To describe the conservation of mass in a fluid flow, the continuity equation shown in equation (1) is used. To describe the dynamics of air in a room, the equations of motion presented in equation (2) is used and equation of temperature (3).

To model the thermal effects due to human body temperature, the Boussinesq approximation was applied, assuming that density variations are negligible except in the buoyancy term. The human body acts as a localized heat source, generating thermal plumes that influence particle movement. "β" is an indicator of the relative change in the density of a liquid depending on temperature. It is defined as:

$$\beta = \frac{2}{T_1 + T_2}. \tag{4}$$

The particle motion was modeled using a dispersed phase (DPM), which is solved by integrating the force balance. In a set of differential equations that calculate the positions and velocities of particles. In the CFD code, all particles are considered as point masses, the kinematic relationship between the position of the particles (respiratory droplet) and the velocity of the particles (respiratory droplet) is equal to

$$\frac{dx_p}{dt} = u_p, \tag{5}$$

$$m_p \frac{du_p}{dt} = F_D + F_G. \quad (6)$$

The trajectory of particles is determined by these forces, since they include the main effects on their movement. F_D - It is calculated as follows:

$$F_D = \frac{1}{2}\rho_f \frac{\pi d_p^2}{4} C_D(u_f - u_p)|u_f - u_p|, \quad (7)$$

$$C_D = \begin{cases} \frac{24}{Re}; & (Re < 1) \\ \frac{24}{Re}(1 + 0.15Re^{0.687}); & (1 \le Re \le 1000) \end{cases} \quad (8)$$

Re is the Reynolds number, calculated as follows:

$$Re = \frac{\rho d_p |u - u_p|}{\mu_d}. \quad (9)$$

Where x_p - the location of the particles, u_p - particle velocity, m_p - particle mass, u_f - the speed of movement of liquids, F_D - drag force, F_G – gravity, C_D - drag coefficient, d_p - particle diameter, Re - Reynolds number, μ_d - dynamic viscosity, ρ_p - particle density, ρ_f - the density of the liquid.

There are two fundamental concepts that play a crucial role in the accurate modeling of particle propagation in airflow: drag force and drag coefficient.

The drag force represents the resistance exerted by the surrounding medium on a moving particle. It arises due to viscous friction and inertial effects as the particle travels through a fluid (gas or liquid).

The drag coefficient is a dimensionless parameter that characterizes the magnitude of this resistance. It depends on the Reynolds number, particle shape, and flow conditions.

These two concepts are essential for describing the interaction between particles and the ambient air, especially when modeling particle transport and deposition in aerosols, liquid droplets, or solid particles. In the current study, the drag force influences both the trajectory and the eventual settling of particles, while the drag coefficient helps determine how particles of different sizes behave under airflow conditions in confined spaces.

Turbulent model. To calculate turbulent flows, a two-level k- SST turbulence model was used. At the same time, this method predicts the behavior of the boundary layer, especially in areas with a high velocity gradient. The main advantage of this method is that it demonstrates stable and accurate results for problems with complex geometry. It should be noted that this method has also proven its reliability and stability when modeling the air flow inside the room (Issakhov et al., 2021).

These equations are discretized using the finite volume method. The equations for the k-ω model can be written as follows:

SST k-ω

$$\frac{\partial k}{\partial t} + u\frac{\partial u_j}{\partial x} = \frac{1}{\rho}(P_k - \eta^* k\omega + \frac{\partial}{\partial x_j}((\mu + \delta^*\mu_t)\frac{\partial k}{\partial x_j})), \tag{10}$$

$$\frac{\partial \omega}{\partial t} + u\frac{\partial u_j \omega}{\partial x_j} = \frac{1}{\rho}\left(\alpha S^2 - \eta\omega^2 + \frac{\partial}{\partial x_j}\left((\mu + \delta_\omega \mu_t)\frac{\partial \omega}{\partial x_j}\right) + \right.$$

$$\left. + 2(1 - f_1)\delta_{\omega_2}\frac{1}{\omega}\frac{\partial k}{\partial x_i}\frac{\partial \omega}{\partial x_i}\right), \tag{11}$$

where:

$$\mu_t = \frac{\alpha_1 k}{max(\alpha_1\omega_1 f_2 S)}, \tag{12}$$

$$P_{k_1} = \mu_t(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})\frac{\partial u_i}{\partial x_j} - \frac{2}{3}\frac{\partial u_k}{\partial x_k}(3\mu_t\frac{\partial u_k}{\partial x_k} + \rho k), \tag{13}$$

$$P_k = min(P_{k_1}, 10\eta^* k\omega), \tag{14}$$

$$f_1 = tanh((min(max(\frac{2\sqrt{k}}{\eta^*\omega y}, \frac{500\mu}{y^2\omega}), \frac{4\delta\omega_2 k}{CD_{k\omega}y^2}))^4), \tag{15}$$

$$f_2 = tanh((max(\frac{2\sqrt{k}}{\eta^*\omega y}, \frac{500\mu}{y^2\omega}))^2), \tag{16}$$

$$CD_{k\omega} = max(2\rho\delta_{\omega_2}\frac{1}{\omega}\frac{\partial k}{\partial x_i}\frac{\partial \omega_i}{\partial x_i}, 10^{-10}), \tag{17}$$

$$\phi = \phi_1 f_1 + \phi_2(1 - f_1). \tag{18}$$

k - kinetic energy of turbulence, u_j – components of the velocity vector, ω - specific dissipation rate, P_k - turbulent kinetic energy production, μ_t - turbulence viscosity, η^* - the constant of the model, δ^* - the constant of the model, S – rate of

deformation, f_1 – function for switching between different model modes, $CD_{k\omega}$ – the correction term for ω, φ – is a unified function, φ_2 – function values in the second zone.

Constants of the model:

$$\alpha_1 = 1, \quad \alpha_2 = 0.52, \quad \eta_1 = \frac{3}{40}, \quad \eta_2 = 0.0828,$$
$$\eta^* = 0.09, \quad \delta_{k_1} = 1.178, \quad \delta_{k_2} = 1, \quad \delta_{\omega_1} = 2, \quad \delta_{\omega_2} = 1.168$$

Boundary conditions. Boundary conditions are conditions that are set at the boundaries of the area necessary for accurate and complete determination of the behavior of the simulated system. The following boundary conditions were set for my thesis:

Inlet (ventilation):

- The air velocity at the inlet is set constant and is 1 m/s.
- The inlet pressure is set according to the Neumann condition and is equal to 0 Pa.
- The temperature of the incoming air is set to 291° K, which corresponds to standard ventilation conditions.
- The conditions of the discrete phase model were like an escape.

Outlet (exhaust):

- The outlet pressure is set to 0 Pa.
- By default, the temperature on the outlet is set at 300 K.
- The air velocity through the outlet is determined by the mass balance in the design area.
- The conditions of the discrete phase model were like an escape.

Inlet (mouth):

- The particles are ejected in the time interval from 0.1 to 0.3 seconds.
- The rate of particle release varies depending on the reflex process (for example 6, 20 m/s).
- The particle temperature is set as 310.15 K.
- The particle composition was set to 0.59% for water, and the remaining 0.41% for air.
- The conditions of the discrete phase model were like a reflect.

Walls and desks:

- The walls and desks of the room are set to be impenetrable and smooth. (No slip)
- The heat flow on the walls and desks are set, which means that the walls and desks temperature will be calculated in the process of solving the problem.
- The conditions of the discrete phase model were like a trap.

People

- The surface of the body (teacher and students) is set at a constant temperature equal to 309.7 K.

• The condition of zero normal velocity is set on the surface of the body. (No slip)

• The conditions of the discrete phase model were like a reflect.

Numerical simulation. Numerical modeling allows to describe the modeling process in detail and solve the problem using computational algorithms. If we cannot solve the equation analytically and make an experiment, then a numerical solution is the best and less expensive way to study the problem. At the same time, the error that may be in the solution is not quite large, but exactly what error we cannot say for sure. By solving the problem numerically, we can take the results more natural and more accurate than nothing. For my research on solving equations, numerical methods implemented in the ANSYS Fluent package were used. The relationship between pressure and velocity was achieved using the SIMPLE numerical algorithm (Semi-Implicit Method for Pressure-Linked Equations). The SIMPLE method is an effective and reliable method for solving the Navier-Stokes equations in problems with incompressible flows. The SIMPLE method has a number of advantages, namely stability and convergence, ease of implementation and can be applied to various sampling schemes and used to solve complex geometries and boundary conditions (Han et al., 2019). Its use in this study allows us to obtain accurate and stable results of modeling the spread of viral particles in confined spaces. Complete the SIMPLE algorithm includes following steps:

1. Guess the pressure field and solve the momentum equations to obtain .

$$\frac{u^{**}-u^{n}}{\Delta t}=-\frac{1}{\rho}\frac{\partial P^{*}}{\partial x}+L_{1}\rightarrow u^{**}=u^{n}+\Delta t(-\frac{1}{\rho}\frac{\partial P^{*}}{\partial x}+L_{1}), \quad (19)$$

2. Solve the equation by substitute into the continuity equation and we get the Poisson equation.

$$\frac{\imath^{n+1}-u^{**}}{\Delta t}=-\frac{1}{\rho}\frac{\partial P'}{\partial x}+L_{1}\rightarrow u^{n+1}=u^{**}-\frac{\Delta t}{\rho}\frac{\partial P'}{\partial x} \quad (20)$$

$$\frac{\partial^{2}P'}{\partial x^{2}}+\frac{\partial^{2}P'}{\partial y^{2}}+\frac{\partial^{2}P'}{\partial z^{2}}=\frac{\rho}{\Delta t}\left(\frac{\partial u^{**}}{\partial x}+\frac{\partial v^{**}}{\partial y}+\frac{\partial w^{**}}{\partial z}\right), \quad (21)$$

3. Correct pressure field $P=P^{*}-P'$, (22)

4. Correct velocity field $u=u^{*}-u'$, (23)

For simplicity, the derivation above considers only the x-component of the velocity field. The corresponding expressions for the y- and z-components can be obtained analogously by applying the same discretization procedure to the respective momentum equations

Table 1— Simulation scenario.

Variant	Scenario	Particle velocity (m/s)	Supply air velocity (m/s)	Particle diameter (м)
1	1	V = 6	1	$1 \times 10^{-6} to\ 1 \times 10^{-4}$
	2	V = 20	1	$1 \times 10^{-6} to\ 1 \times 10^{-4}$
2	3	V = 6	1	$1 \times 10^{-6} to\ 1 \times 10^{-4}$
	4	V = 20	1	$1 \times 10^{-6} to\ 1 \times 10^{-4}$

All equations in the discrete phase model and turbulent models are solved by default using the explicit method.

The momentum was discretized utilizing the second-order discretization

scheme. The energy which means temperature was discretized utilizing the first-order discretization scheme. The solution is considered converged when all the scaled residuals stabilize and reach a minimum of for the k, ω, continuity, and x, y, and z momentum equations, as well as for the energy equation.

Geometry and mesh. The COVID-19 pandemic has highlighted the importance of maintaining social distance, which has led to increased interest in analyzing the geometry of classrooms with the presence of teachers and students. The creation of a geometric model of the classroom was carried out taking into account the average parameters of the room. I conducted a detailed study of the geometry of the classroom, including its dimensions and the location of the entrances and exits. Thus, for the study, I took the average room size of X*Y*Z = 7*3*4 m. This room served as the main model for analyzing the spread of viral particles in a confined space

We examined two types of geometry of the classroom where the infected teacher is, and also took into account the presence of ventilation and entrance to the room. The height of the teacher inside the room was 1.94 m, and the teacher stood 0.2 m taller, which could affect the propagation of particles, as shown in Figure 2. The height of the infected person's mouth was about 1.76 m.

Ventilation in the room was carried out using a single grate on the side wall with a size Y*Z = 0.125*0.5 m, while an extractor hood of the same size was located 0.275 m below the ventilation on the opposite wall. All geometries consider made a standing teacher and four sitting students. The teacher looked at the students or at the blackboard, depending on the geometry option. For the realism of the model, desks in front of the students with dimensions X*Y*Z = 0.5*0.9*0.8 m were added.

The following distances were taken into account when describing the geometry in detail:

- The distance from the teacher to the students was 2.5 m.
- The distance between the students is 1.5 m and between the desks is 1 m.
- The distance from the wall to the desk is 0.7 m.
- The distance between the desk and the front wall is 2.5 m.

• The distance to the wall at the back is 2 m.

In this study, two types of geometry were considered:

1)The teacher looks at the students taking into account ventilation and exhaust. (which is shown in the Figure 1).

2)The teacher looks at the board, taking into account ventilation and exhaust. (which is shown in the Figure 1).

The simulation also took into account various velocities (1 m/s, 6 m/s, 20 m/s) corresponding to reflex processes (breathing, sneezing, coughing) in order to fully study the spread of viral particles.

Figure 3 shows a three-dimensional (3D) computational grid of the study area, in which the total number of elements is 2898449. The value of the grid size plays a key role in ensuring the accuracy of the results. The unstructured geometry caused by the presence of people in the room was taken into account when choosing a hexahedral grid. This choice of technology is due to its efficiency when working with solids and simplification of the grid generation process. Given this unstructured nature, creating a grid requires additional efforts to ensure its quality and accuracy. Nevertheless, the hexahedral grid remains the preferred choice, as it provides sufficient accuracy of the results and successfully solves the tasks set. It is important to note that the resulting hexahedral elements are used in the Finite Volume Method (FVM) for subsequent numerical analysis.

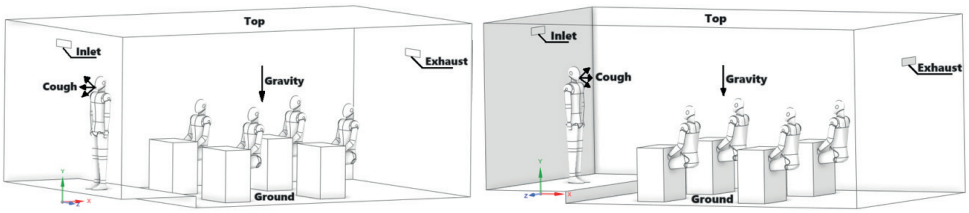


Figure 1— The geometry of the area under study.

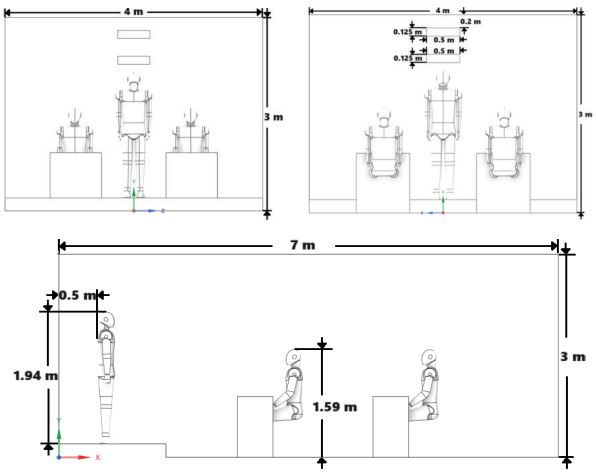


Figure 2— The geometry of the area under study.

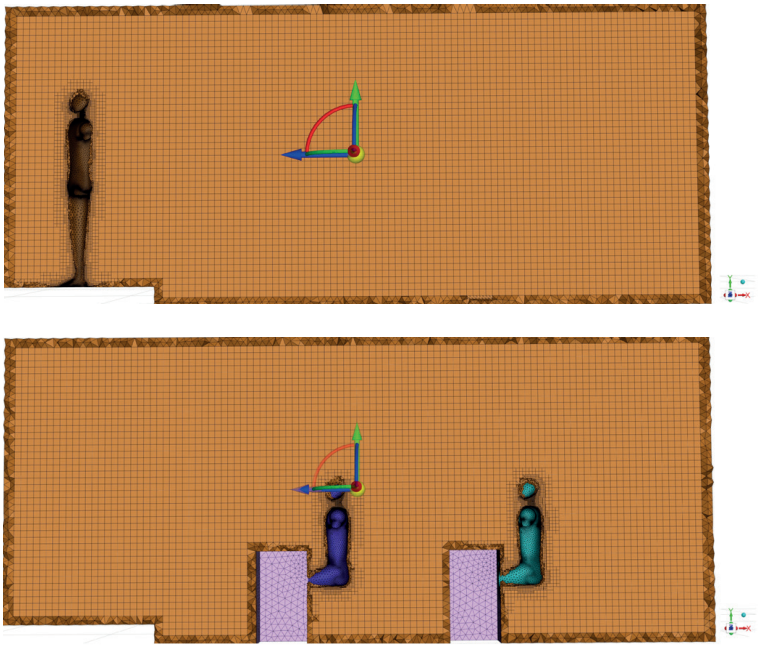


Figure 3 — Computational grid of the studied area.

Properties of materials. Air was modeled as an incompressible fluid with constant properties which shown in **Table 2.** The coefficient of thermal expansion was calculated using the formula given in 1.1.6, which indicates temperatures corresponding to the air through ventilation (291 K) and human body temperature (309.7 K), respectively. These values were chosen to most accurately reflect the conditions in the real system.

Table 2 — The property of air.

Property	Value	Unit
Density	1.225	kg/m^3
Specific Heat	1006.43	$J/(kg\ K)$
Thermal Conductivity	0.0242	$W/(m\ K)$
Viscosity	1.7894e-05	$kg/(m\ s)$
Molecular Weight	28.966	$\underline{kg/kmol}$
Thermal Expansion Coefficient	0.00332696	K^{-1}

The particles were modeled as a mixture of water and air, simulating the composition of real respiratory droplets. The average particle density was set to 600 kg/m³, based on an assumed composition of 59% water and 41% air.

$$\rho_p = 0.41\rho_{air} \cdot 0.59\rho_{water}, \quad (24)$$

$$\rho_p = 0.41 \cdot 1.225 + 0.59 * 1000 = 590.5\ kg/m^3. \quad (25)$$

This allowed us to take into account the peculiarities of the composition of viral particles and their impact on the environment. These steps provide a more accurate and realistic simulation of the spread of viral particles in the air, which is important for understanding and preventing transmission of infection.

Results

Particle distribution. The results of a numerical study of scenarios 1-2 are presented in Figures 4, 5. Figures 4, 5 show the results of particle propagation in a room with ventilation of 1 m/s at different points in time. Figure 4 shows the results of particle propagation at different points in time when coughing or sneezing 6 m/s. These results show that after 30 seconds the particles are transported by 3.36 m in length, which exceeds the social distance. The results presented in Figure 5 show that after 30 seconds the particles travel a distance of 4.94 m in length, which exceeds the social distance recommended by WHO. This means that option 3 is also dangerous for people in this range. As can be seen from this scenario, the spread has reached the opposite wall in the room. As we have seen, options 1 and 2 differ only with ventilation and exhaust. It should be noted that when using ventilation, particles spread a much longer distance than without ventilation, since particles with a smaller diameter under the ventilation flow can spread longer and further around the room.

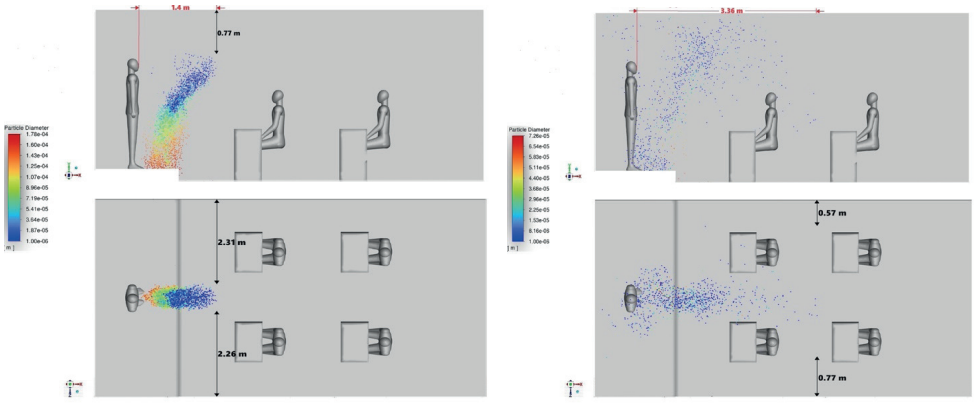


Figure 4 — Particle distribution during coughing = 6 m/s and ventilation = 1 m/s (5 s, 30 s)

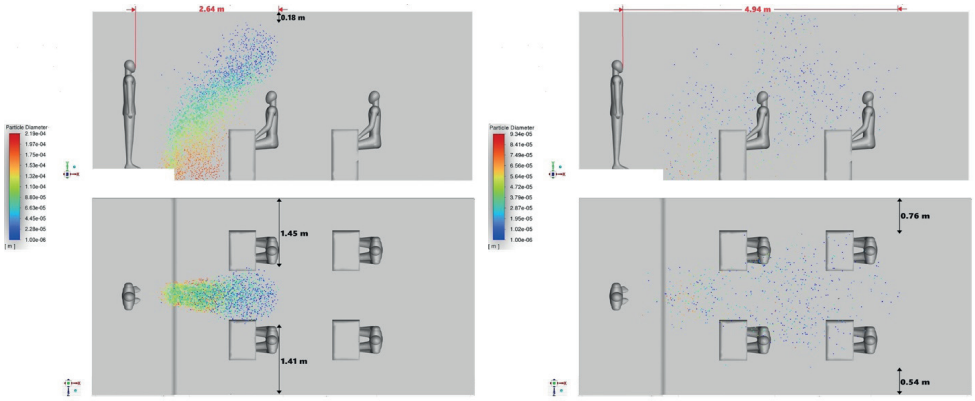


Figure 5 — Particle distribution during coughing = 20 m/s and ventilation = 1 m/s (5 s, 30 s)

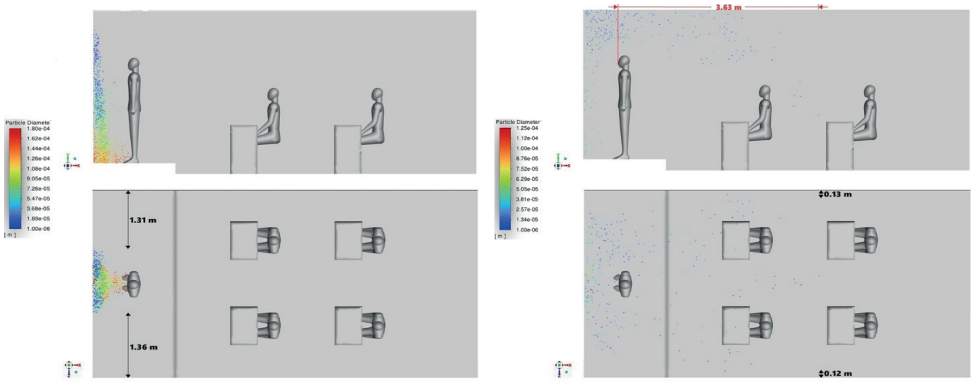


Figure 6 — Particle distribution during coughing = 6 m/s and ventilation = 1 m/s (5 s, 30 s)

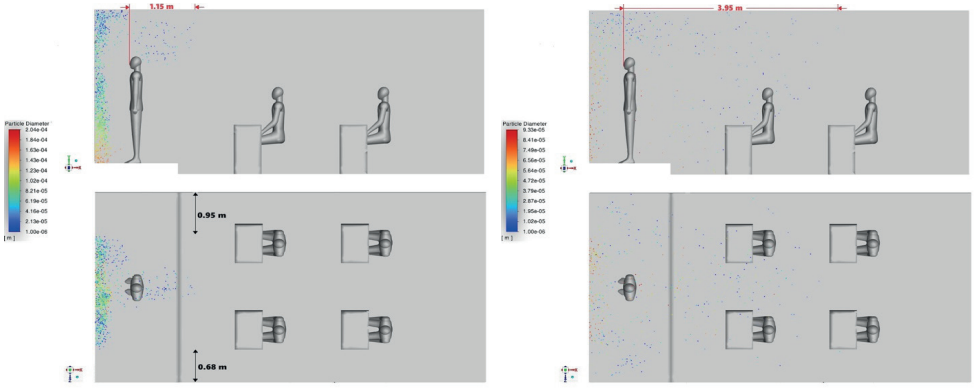


Figure 7 — Particle distribution during coughing = 20 m/s and ventilation = 1 m/s (5 s, 30 s)

Temperature of human body. Figures 8-11 show the results of human body temperature for all variants. It should be noted that in the ventilation and exhaust versions, the air flow significantly affects the temperature distribution, which leads to a higher temperature at the end of the room. The ventilation options have shown that the air temperature is more evenly distributed throughout the room, creating comfortable conditions.

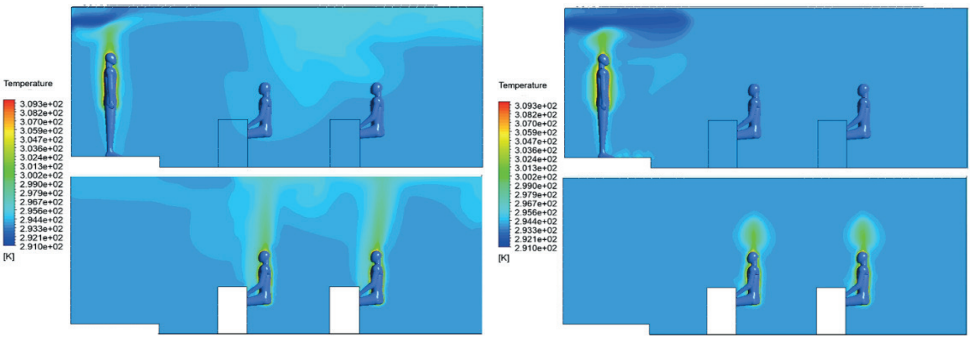


Figure 8 — The effect of body temperature on particle distribution during coughing = 6 m/s and ventilation = 1 m/s (5 s, 30 s)

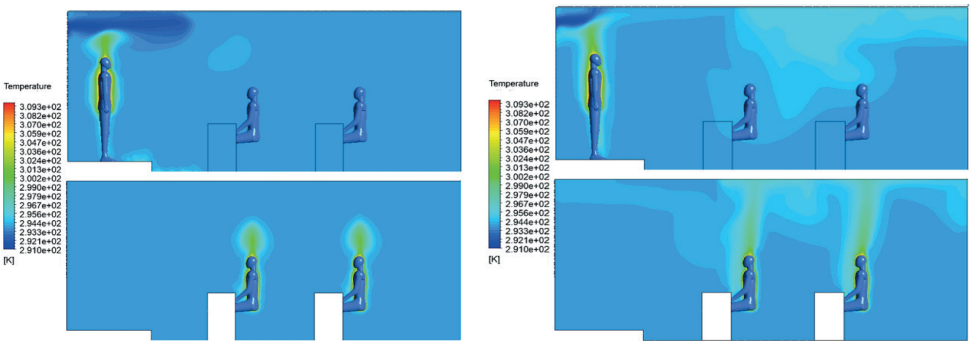


Figure 9 —The effect of body temperature on particle distribution during coughing = 20 m/s and ventilation = 1 m/s (5 s, 30 s)

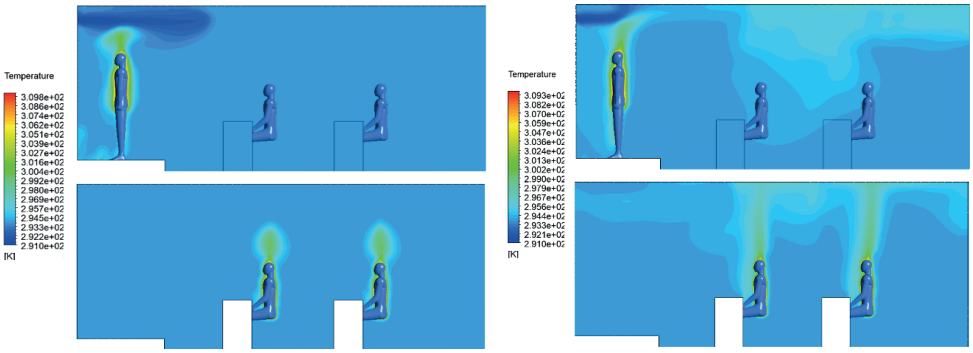


Figure 10 —The effect of body temperature on particle distribution during coughing = 6 m/s and ventilation = 1 m/s (5 s, 30 s)

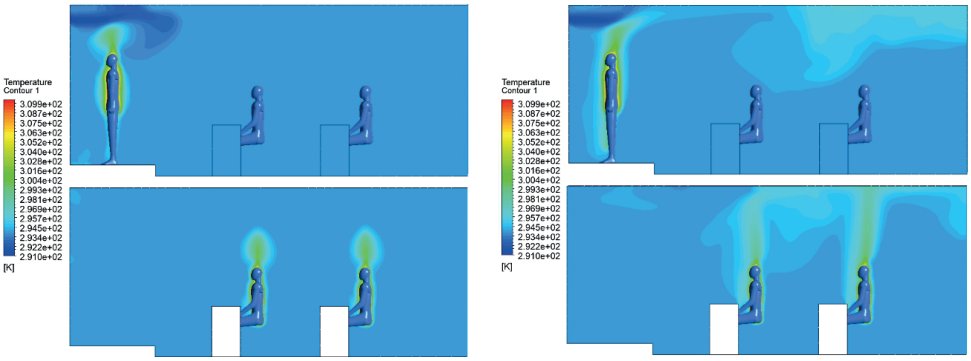


Figure 11— The effect of body temperature on particle distribution during coughing = 20 m/s and ventilation = 1 m/s (5 s, 30 s)

Discussion. The results of a numerical study of scenarios 3-4 are presented in Figures 6, 7 Figures 6, 7 show the results of particle propagation in a room with ventilation of 1 m/s at different points in time. Figure 6 shows the results of particle propagation at different points in time when coughing or sneezing 6 m/s. These results show that after 30 seconds the particles are transported by 3.63 m in length, which exceeds the social distance. The results presented in Figure 7 show that after 30 seconds the particles travel a distance of 3.95 m in length, which exceeds the social distance recommended by WHO. This means that option 2 is also dangerous for people in this range. As can be seen from this scenario, the spread has reached the opposite wall in the room. As we have seen, options 2 differ only with ventilation and exhaust. It should be noted that when using ventilation, since particles with a smaller diameter under the ventilation flow can spread longer and further around the room. Thus, this means that rooms with ventilation and exhaust are more dangerous than rooms. Because particles travel social distances and even spread further.

The processes of evaporation and condensation of particles depend on temperature conditions, and ventilation options contribute to faster evaporation of particles, which can reduce their concentration in the air. Efficient ventilation systems have shown their ability to quickly remove viral particles from the room, which underscores their importance for hygiene and safety.

Active ventilation creates turbulent flows that contribute to the chaotic spread of particles, increasing the likelihood of their deposition on surfaces or removal from the breathing zone.

However, the analysis of particle propagation showed that options with ventilation can be more dangerous for humans, since particles spread over a longer distance. This indicates the need for careful design of ventilation systems to minimize the risk of spreading viral particles. A possible solution may be to optimize the direction of air flows and use additional methods of air filtration and purification.

Conclusion. In this thesis, numerical modeling of the spread of viral particles

in various scenarios is carried out, taking into account the presence or absence of ventilation and exhaust. The study covered 4 scenarios, the results of which are shown in Figures 4-7 and 8-11. Data analysis allowed us to draw several important conclusions about the influence of various conditions on the distribution and deposition of particles.

Main conclusions:

- Particles with large diameters are deposited by gravity, which somewhat limits their propagation.
- Ventilation promotes the chaotic movement of particles due to turbulent flows, which increases the likelihood of their deposition on surfaces or their removal from the breathing zone.
- Poorly designed ventilation systems may unintentionally direct contaminated air toward occupants, increasing the risk of infection.
- Thermal effects from the human body induce upward convective currents that prolong particle suspension and affect their vertical distribution.

These findings emphasize the importance of optimizing HVAC systems in public indoor environments, especially classrooms, to minimize the transmission of airborne pathogens. The computational approach applied in this study can serve as a practical tool for assessing infection risks and guiding preventive strategies.

References

Aliyu A., Uzoka C., Singh D., & Mishra R. (2021) Dispersion of virus-laden droplets in ventilated rooms: Effect of homemade facemasks. Journal of Building Engineering, 44. — 102933 p. https://doi.org/10.1016/j.jobe.2021.102933 (in Eng.)

Benjamin P., David H.E., Ian D.P., Ronnak B.P., Surinder S.B., & Sergio M. (2006) Cough frequency, cough sensitivity and health status in patients with chronic cough. Respiratory Medicine, 100 (6). — P. 1105–1109. https://doi.org/10.1016/j.rmed.2005.09.023 (in Eng.)

Busco G., Hassan Y.A., Seo J., & Yang S.R. (2020) Sneezing and asymptomatic virus transmission. Physics of Fluids, 32 (7), 073309. https://doi.org/10.1063/5.0019090 (in Eng.)

Chen Q., Lin C.H., & Gupta J.K. (2009) Flow dynamics and characterization of a cough. Indoor Air, 19 (6). — P. 517–525. https://doi.org/10.1111/j.1600-0668.2009.00619.x (in Eng.)

Gao N. P., & Niu J.L. (2007) Modeling particle dispersion and deposition in indoor environments. Atmospheric Environment, 41 (18). — P. 3862–3876. https://doi.org/10.1016/j.atmosenv.2007.01.016 (in Eng.)

Goniva C., Kahrimanovic D., & Pirker S. (2011) Improving the applicability of discrete phase simulations by smoothening their exchange fields. Applied Mathematical Modelling, 35 (12). — P. 5896–5909. https://doi.org/10.1016/j.apm.2010.11.066 (in Eng.)

Han M., Kikumoto H., & Ooka R. (2019) Lattice Boltzmann method-based large-eddy simulation of indoor isothermal airflow. International Journal of Heat and Mass Transfer, 130. — P. 700–709. https://doi.org/10.1016/j.ijheatmasstransfer.2018.10.137 (in Eng.)

Hansen B., & Mygind H. (2002) How often do normal persons sneeze and blow the nose? Rhinology, 40 (1). — P. 10–12. (in Eng.)

Issakhov A., Alimbek A., & Zhandaulet Y. (2021) The assessment of water pollution by chemical reaction products from the activities of industrial facilities: Numerical study. Journal of Cleaner Production, 282. — 125239 p. https://doi.org/10.1016/j.jclepro.2020.125239 (in Eng.)

Lednicky J.A., Pan M., & Wu C. (2019) Collection, particle sizing and detection of airborne viruses. *Journal of Applied Microbiology, 127 (6). — P. 1596–1611. https://doi.org/10.1111/jam.14278 (in Eng.)

Mei D., Li J., Jia M., Hou S., Duan W., & Liu Z. (2022) Analysis of the spread of cough droplets and body deposition fraction in the smart classroom in different seasons. Health & Environmental Research Online, 5. — 100015 p. https://doi.org/10.1016/j.heha.2022.100015 (in Eng.)

Motamedi H., Shirzadi M., Mirzaei P., & Tominaga Y. (2021) CFD modeling of airborne pathogen transmission of COVID-19 in confined spaces under different ventilation strategies. Sustainable Cities and Society, 72. — 103397p. https://doi.org/10.1016/j.scs.2021.103397 (in Eng.)

Mutlu M. (2020) Numerical investigation of indoor air quality in a floor heated room with different air change rates. Building Simulation, 13. — P. 1299–1312. https://doi.org/10.1007/s12273-020-0683-5 (in Eng.)

Srivastava A., Ghosh B., & Lal H. (2015) Review of bioaerosols in indoor environment with special reference to sampling, analysis and control mechanisms. Environment International, 85. — P. 254–272. https://doi.org/10.1016/j.envint.2015.09.018 (in Eng.)

Wölfel R., et al. (2020) Virological assessment of hospitalized patients with COVID-2019. Nature, 581 (7809). — P. 465–469. https://doi.org/10.1038/s41586-020-2196-x (in Eng.)

World Health Organization (2023) WHO coronavirus (COVID-19) dashboard with vaccination data. https://covid19.who.int/ (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 202–217

https://doi.org/10.32014/2026.2518-1726.409

IRSTI: 20.23.17
UDC 004.8:614.8

© Kantureeva M.[1], Omarova G.S.[1], Duisen Z.D.[2], Shekerbek A.A.[1], Tulebayev Y.B.[2], 2026.
[1] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Astana IT University, Astana, Kazakhstan.
E-mail: duisenzhanerke50@gmail.com

APPLICATION OF MACHINE LEARNING METHODS IN FORECASTING AND OPTIMIZING THE PROCESSES OF EVACUATION OF PEOPLE IN HIGH-RISE BUILDINGS

Kantureeva Mansiya — PhD, senior lecturer at the Department of Information Systems at the L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: kantureyeva_ma@enu.kz, ORCID ID: https://orcid.org/0000-0001-5904-820X;
Omarova Gulmira — PhD, senior lecturer at the Department of Information Systems at the L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: ogs12@mail.ru, ORCID ID: 0000-0003-2053-0255;
Duisen Zhanerke — Master's student at the Faculty of Information Technologies, Astana IT University, Astana, Kazakhstan,
E-mail: duisenzhanerke50@gmail.com, ORCID ID: https://orcid.org/0009-0009-8746-5946;
Shekerbek Ainur — PhD of the Department of Information Systems of the Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan,
E-mail: shekerbek80@mail.ru, 800520401702@enu.kz, ORCID ID: https://orcid.org/0000-0002-1088-4239;
Tulebayev Yersultan — Master of Technical Sciences, School of Software Engineering, Astana IT University, Astana, Kazakhstan,
E-mail: yersultan.tulebayev@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-7965-1476.

Abstract. High-rise buildings in major cities pose significant safety challenges during emergencies, including fires, explosions, and technological accidents. Factors such as dense population, limited evacuation routes, and complex multi-level structures complicate rapid and secure evacuation. This paper explores the use of machine learning methods to analyze factors affecting the effectiveness of human evacuation from high-rise buildings and to predict the risk of complicated evacuation based on real-world statistical data. The analysis used a publicly available data set on the Toronto fires, which included more than 12,000 individual fire records. This dataset contained information about the response time of the fire department, the operational status of the fire alarm and sprinkler systems, the technical characteristics

of the building, the source of the fire, and the final results. Before the analysis, the data went through several stages of preprocessing. These included deleting incomplete records, encoding categorical variables, creating new variables, and visualizing key metrics to identify important patterns. To solve the forecasting problem, models of Logistic regression, Random Forest and XGBoost were used, and their effectiveness was evaluated and compared using established classification indicators. The results of the study showed that the response time of the fire service, the availability and effective operation of alarm and fire extinguishing systems, the height of the building and the location of the fire significantly affected the success of the evacuation. The Random Forest model has achieved an optimal balance between accuracy and sensitivity, thereby demonstrating the ability of machine learning to improve fire safety.

Key words: High-rise buildings, evacuation, fire safety, emergency situations, machine learning

For citations: Kantureeva M., Omarova G.S., Duisen Z.D., Shekerbek A.A., Tulebayev Y.B. Application of machine learning methods in forecasting and optimizing the processes of evacuation of people in high-rise buildings. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 202–217. DOI: https://doi.org/10.32014/2026.2518-1726.409

© Кантуреева М.А.[1], Омарова Г.С.[1], Дүйсен Ж.Д.[2], Шекербек А.Ә.[1], Тулебаев Е.Б.[2], 2026.

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Astana IT University, Астана, Қазақстан.
E-mail: duisenzhanerke50@gmail.com

БИІК ҒИМАРАТТАРДАҒЫ АДАМДАРДЫ ЭВАКУАЦИЯЛАУ ПРОЦЕСТЕРІН БОЛЖАУ ЖӘНЕ ОҢТАЙЛАНДЫРУДА МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ

Кантуреева Мансия — PhD, «Ақпараттық жүйелер» кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: kantureyeva_ma@enu.kz, ORCID ID: https://orcid.org/0000-0001-5904-820X;
Омарова Гульмира — PhD, «Ақпараттық жүйелер» кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: ogs12@mail.ru, ORCID ID: 0000-0003-2053-0255;
Дүйсен Жанерке — «Ақпараттық технологиялар» факультетінің магистранты, Astana IT University, Астана, Қазақстан,
E-mail: duisenzhanerke50@gmail.com, ORCID ID: https://orcid.org/0009-0009-8746-5946;
Шекербек Айнұр — PhD, Ақпараттық жүйелер кафедрасы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: shekerbek80@mail.ru, 800520401702@enu.kz, ORCID ID: https://orcid.org/0000-0002-1088-4239;
Тулебаев Ерсұлтан — Техникалық ғылымдар магистрі, Бағдарламалық инженерия мектебі, Astana IT University, Астана, Қазақстан,
E-mail: yersultan.tulebayev@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-7965-1476.

Аннотация. Ірі қалалардағы биік ғимараттар өрт, жарылыс және техногендік апаттарды қоса алғанда, төтенше жағдайлар кезінде қауіпсіздікке үлкен қиындықтар туғызады. Халық саны, шектеулі эвакуация жолдары және күрделі көп деңгейлі құрылымдар сияқты факторлар тез және қауіпсіз эвакуациялауды қиындатады. Бұл жұмыста биік ғимараттардан адамдарды эвакуациялау тиімділігіне әсер ететін факторларды талдау және эвакуацияның күрделену қаупін болжау үшін машиналық оқыту әдістерін қолдану қарастырылады. Талдауда Торонтодағы өрттер туралы ашық қол жетімді мәліметтер жиынтығы қолданылды, оған 12000-нан астам жеке өрт жазбалары кіреді. Бұл деректер жиынтығында өрт сөндіру қызметінің жауап беру уақыты, өрт дабылы мен спринклер жүйелерінің жұмыс жағдайы, ғимараттың техникалық сипаттамалары, өрт көзі және соңғы нәтижелер туралы ақпарат болды. Талдау алдында деректер алдын ала өңдеудің бірнеше кезеңдерінен өтті. Оларға толық емес жазбаларды жою, категориялық айнымалыларды кодтау, жаңа айнымалылар құру және маңызды заңдылықтарды анықтау үшін негізгі көрсеткіштерді визуализациялау кірді. Болжау мәселесін шешу үшін логистикалық регрессия, Random Forest және XGBoost модельдері қолданылды және олардың тиімділігі белгіленген жіктеу көрсеткіштерін қолдана отырып бағаланды және салыстырылды. Зерттеу нәтижелері өрт сөндіру қызметінің жауап беру уақыты, дабыл және өрт сөндіру жүйелерінің болуы және тиімді жұмыс істеуі, ғимараттың биіктігі мен өрттің орналасуы эвакуацияның сәттілігіне айтарлықтай әсер еткенін көрсетті. Random Forest моделі дәлдік пен сезімталдық арасындағы оңтайлы тепе-теңдікке қол жеткізді, осылайша машиналық оқытудың өрт қауіпсіздігін арттыру қабілетін көрсетті.

Түйін сөздер: Биік ғимараттар, эвакуация, өрт қауіпсіздігі, төтенше жағдайлар, машиналық оқыту

© Кантуреева М.А.[1], Омарова Г.С.[1], Дүйсен Ж.Д. [2], Шекербек А.Ә.[1], Тулебаев Е.Б.[2], 2026.

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

[2]Astana IT University, Астана, Казахстан.

E-mail: duisenzhanerke50@gmail.com

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ И ОПТИМИЗАЦИИ ПРОЦЕССОВ ЭВАКУАЦИИ ЛЮДЕЙ В ВЫСОТНЫХ ЗДАНИЯХ

Кантуреева Мансия — PhD, старший преподаватель кафедры «Информационные системы», Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан, E-mail: kantureyeva_ma@enu.kz, ORCID ID: https://orcid.org/0000-0001-5904-820X;

Омарова Гульмира — PhD, старший преподаватель кафедры «Информационные системы», Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан, E-mail: ogs12@mail.ru, ORCID ID: 0000-0003-2053-0255;

Дүйсен Жанерке — магистрант факультета «Информационные технологии», Astana IT University, Астана, Казахстан,
E-mail: duisenzhanerke50@gmail.com, ORCID ID: https://orcid.org/0009-0009-8746-5946;
Шекербек Айнур — PhD, старший преподаватель кафедры «Информационные системы», Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: shekerbek80@mail.ru, 800520401702@enu.kz, ORCID ID: https://orcid.org/0000-0002-1088-4239;
Тулебаев Ерсултан — магистр технических наук, Школа программной инженерии, Astana IT University, Астана, Казахстан,
E-mail: yersultan.tulebayev@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-7965-1476.

Аннотация. Высотные здания в крупных городах представляют серьёзные вызовы для обеспечения безопасности в условиях чрезвычайных ситуаций, включая пожары, взрывы и техногенные аварии. Высокая плотность населения, ограниченное количество путей эвакуации и сложная многоуровневая структура зданий существенно затрудняют оперативную и безопасную эвакуацию людей. В работе рассматривается применение методов машинного обучения для анализа факторов, влияющих на успешность эвакуации, а также для прогнозирования риска осложнённой эвакуации на основе реальных статистических данных. В качестве эмпирической базы использован открытый набор данных о пожарах в городе Торонто, включающий более 12 000 записей. Данные содержат информацию о времени реагирования пожарных служб, состоянии систем сигнализации и пожаротушения, характеристиках зданий, источниках возгорания и исходах инцидентов. Предварительная обработка данных включала удаление неполных наблюдений, кодирование категориальных переменных, создание новых признаков и визуализацию ключевых показателей для выявления закономерностей. Для построения прогностических моделей применялись логистическая регрессия, Random Forest и XGBoost. Эффективность моделей оценивалась с использованием стандартных классификационных метрик. Результаты исследования показали, что время реагирования пожарной службы, наличие и работоспособность систем сигнализации и пожаротушения, высота здания и локализация пожара оказывают существенное влияние на успешность эвакуации. Модель Random Forest продемонстрировала наилучший баланс между точностью и чувствительностью, что подтверждает высокий потенциал методов машинного обучения для повышения уровня пожарной безопасности и оптимизации процессов эвакуации.

Ключевые слова: Высотные здания, эвакуация, пожарная безопасность, чрезвычайные ситуации, машинное обучение

Кіріспе. Қазіргі ірі қалалардағы биік ғимараттардың салынуы қарқынды дамып келе жатыр. Бұл өрт, жарылыс және техногендік апаттар сияқты төтенше жағдайлар кезінде адамдардың қауіпсіздігін арттыруды қажет етеді. Биік ғимараттар күрделі инженерлік жүйелер болып табылады және эвакуацияны ұйымдастыру әсіресе халықтың тығыздығының жоғары болуына, эвакуация-

лық шығулардың шектеулі санына және көп деңгейлі құрылымға байланысты қиынға соғады. Жылдам және қауіпсіз эвакуацияны қамтамасыз ету өрт қауіпсіздігі жүйелерінің ғана емес, сонымен қатар қалалық жұмысшылардың жалпы тәуекелдерін басқарудың негізгі міндеті болып табылады.

Адам ағынының инженерлік үлгілері, мінез-құлық модельдеулері, агентке бағытталған жүйелер және маршрутты оңтайландыру әдістері эвакуацияны талдаудың заманауи әдістері болып табылады. Алайда, шын жағдайларда эвакуацияның тиімділігі жоспарлау шешімдеріне ғана емес, сонымен қатар адам факторына, яғни қауіпті қабылдау, дүрбелең, адамдар арасындағы өзара әрекеттесуге және қызметтердің уақытылы әрекет етуіне байланысты. Осыған байланысты, нақты оқиғалар негізінде заңдылықтарды анықтауға және экстремалды жағдайларда адамдардың мінез-құлқын болжаудың дәлдігін арттыруға мүмкіндік беретін деректерді талдау және машиналық оқыту әдістері белсенді дамып келеді

Ашық өрт деректерін пайдалану эвакуацияның сәттілігіне әсер ететін факторларды эмпирикалық түрде зерттеуге жаңа мүмкіндіктер береді. Мұндай дереккөздердің бірі – Торонтодағы өрт оқиғалары туралы мәліметтер жиынтығы (Toronto Fire Incidents Dataset). Бұл деректер жиынтығында жауап беру уақыты, нысан түрлері, өрт дабылы жүйелері және басқа да оқиғалар параметрлері туралы ақпарат бар (Kaggle, 2023). Осындай секілді деректерді талдау үшін қолданылатын машиналық оқыту алгоритмдері оқиғаларды ұзақтығы бойынша жіктеуге, әртүрлі факторлардың әсерін бағалауға және көп қабатты үйлерді эвакуациялау кезіндегі қауіптер туралы түсінікті арттыруға мүмкіндік береді. Сол себепті эвакуация тиімділігін болжай алатын және адамдардың қауіпсіздігіне әсер ететін негізгі факторларды анықтай алатын интеллектуалды модельдерді әзірлеу қажеттілігі туындайды.

Зерттеудің мақсаты – ашық өрт деректерін пайдалана отырып, адамдарды биік ғимараттардан эвакуациялаудың тиімділігін талдау және болжау үшін машиналық оқыту үлгілерін жасау және бағалау болып табылады. Жұмыс барысында эвакуацияның сәттілігіне және жауап беру уақытына әсер ететін негізгі элементтерді зерделеу, сондай-ақ эвакуация процесінде адам шығыны мен ықтимал кідірістерді бағалау үшін болжамды модель жасау жоспарлануда.

Әдеби шолу. Соңғы жылдары биік ғимараттардан адамдарды эвакуациялау саласындағы зерттеулер қарапайым инженерлік модельдерден адамдардың мінез-құлқын, өрт динамикасын және ортаның белгісіздігін ескеретін кешенді интеллектуалды жүйелерге көшуін көрсетеді. Қазіргі тәсілдер оңтайландыру әдістерін, агентке бағытталған модельдеуді, машиналық оқытуды және сандық егіздерді біріктіріп, болжам дәлдігін арттыруға және эвакуацияны басқарудың тиімділігін жақсартуға мүмкіндік береді.

1. Инженерлік және оңтайландыру модельдері

Классикалық инженерлік модельдер эвакуацияны талдаудың негізі болып қала береді, себебі олар ғимараттың геометриясын, жолдардың өткізу қабілетін және адамдар ағынының тығыздығын ескеруге мүмкіндік береді. Ronchi &

Nilsson (2013) шолуы көрсеткендей, мұндай модельдер биік ғимараттарды жобалау кезінде кеңінен қолданылады, алайда олар адамдардың мінез-құлқындағы жеке ерекшеліктер мен стресс факторларының әсерін әлсіз көрсетеді. Есептеу әдістерінің дамуы эвакуациялық ағындарды динамикалық түрде тарату үшін оңтайландыру алгоритмдерін енгізуге әкелді. Gao et al. (2022) Шығу жолдарын таңдауда жеке қалауларды ескеретін модель ұсынды, бұл кептелістердің пайда болу ықтималдығын азайтып, жалпы эвакуация уақытын қысқартуға мүмкіндік берді. Сол сияқты тәсіл Li et al. (2024) еңбегінде жүзеге асырылған, онда жетілдірілген құмырсқа алгоритмі интеллектуалды өрт сөндіру жүйесінде адаптивті маршруттарды құру үшін қолданылады Gravit et al. (2018) ұсынған фазалық эвакуацияның инженерлік алгоритмдері баспалдақ торларының шамадан тыс жүктелуін азайту мақсатында биік ғимараттардың қабаттарын кезең-кезеңімен босатуға бағытталған. Carev et al. (2024) терезелер, балкондар және шатырлар арқылы эвакуация жүргізуге мүмкіндік беретін қашықтықтан басқарылатын механикаландырылған платформаға негізделген инженерлік құтқару жүйесін ұсынды. Осылайша, оңтайландыру әдістері дәстүрлі модельдердің тиімділігін арттырады, алайда әлі де нақты кіріс деректерін қажет етеді және адамдардың когнитивті ерекшеліктерін толық ескермейді.

2. Адамның мінез-құлқы мен когнитивтік факторлар

Адам факторын ескеру эвакуациялық модельдердің шынайылығын арттырудың негізгі шарты ретінде қарастырылады. Zhang et al. (2024) эксперименттік зерттеуі, тік эвакуация кезінде адамдардың тығыздығына және визуалды стимулдарға байланысты топтың қозғалысын синхрондау эффектісі пайда болатынын көрсетті, бұл баспалдақтармен жылжу жылдамдығына тікелей әсер етеді. Teng et al. (2024) қоршаған орта контексті мен гендерлік айырмашылықтар эвакуациялық шығуды таңдауға статикалық маңызды әсер ететінін анықтады, бұл модельдің психологиялық параметрлерін қосу қажеттілігін растайды. Gerges et al. (2021) агенттік модельдеу негізінде жасалған жекелендірілген эвакуация нұсқаулары адамдардың ғимараттан шығу уақытын 50% дейін қысқартуға мүмкіндік беретінін көрсетті. Li et al. (2022) алдын ала эвакуация кезеңіндегі шешім қабылдау процесін қарастырып, қозғалысты кеш бастау адам шығынының жалпы қауіпін айтарлықтай арттыратынын көрсетті. Сондай-ақ Richardson et al. (2017) ғимараттың жоспарын білу өрт және түтін жағдайында сәтті эвакуациялау ықтималдығын айтарлықтай арттыратынын анықтады. Wang et al. (2023) байестік тепе-теңдікке негізделген агенттікке бағытталған модельді ұсынды, онда адамдар маршрутты таңдағанда болашақ топтың тығыздығын болжайды. Zablotsky et al. (2024) кооперативті мінез-құлықты модельдеу жергілікті тығыздықты және жалпы эвакуация уақытын азайтатынын дәлелдеді.

3. Интеллектуалды және машиналық оқыту жүйелері

Машиналық оқыту әдістері эвакуация саласындағы зерттеулердің ең перспективті бағыттарының біріне айналды. Testa et al. (2020) егжей-тегжейлі

3D модельдеуді қажет етпейтін CrowdEst эвакуация параметрлерін бағалау әдісін ұсынды. Әдіс симуляциялық деректерде оқытылған нейрондық желілер мен эвристикалық біріктіру алгоритмдерін қолданады. Saadatmand et al. (2025) Random Forest және Gradient Boosting алгоритмдері дағдарысқа дейінгі кезеңде ғимараттардың эвакуациялық тәуекелдерге осалдығын тиімді бағалауға мүмкіндік беретінін көрсетті. Қазіргі зерттеулер Random Forest негізіндегі модельдердің деректердегі шуға жоғары төзімділігін және тәуекел факторларын анықтауға мүмкіндік беретінін растайды, ал градиентті бустинг күрделі сызықтық емес тәуелділіктер кезінде болжаудың жоғары дәлдігін қамтамасыз етеді. Бұл әдістер өрттер мен эвакуация туралы үлкен деректер жиынтықтарын талдау үшін перспективті етеді.

4. Геоақпараттық технологиялар және цифрлық егіздер

Эвакуацияны модельдеуге цифрлық технологияларды интеграциялау аясында, оңтайлы эвакуация маршруттарын анықтау үшін BIM (Building Information Modeling) және GIS (Geographic Information Systems) жүйелерін бірлесіп пайдалану әдістері ұсынылды. Atyabi et al. (2019) BIM-нің ғимараттың ішкі кеңістігінің геометриясы туралы деректерін GIS-талдауымен біріктіріп, құрылымдық сипаттамалар мен өрт ошағының орналасуын ескере отырып, ең қауіпсіз эвакуация жолдарын есептейтін әдістеме әзірледі. Сол сияқты Bayat et al. (2020) осындай интеграцияланған әдістер 3D-граф ішкі құрылымын және маршруттарды талдау үшін GIS алгоритмдерін қолдана отырып эвакуацияны модельдеуге мүмкіндік беретінін көрсетті, бұл төтенше жағдайларға жедел әрекет ету қызметтерінің шешім қабылдауын жақсартады. Senanayake et al. (2024) жүйелі шолуы, эвакуацияны модельдеу саласындағы қазіргі зерттеулердің нақты орталардың цифрлық егіздерін жасауға бағытталғанын көрсетеді, мұнда агенттік модельдердің (ABM) және машиналық оқыту (ML) әдістерінің үйлесімі статикалық модельдердің шектеулерін жеңіп, эвакуация сценарийлерін динамикалық, бейімделгіш және шынайы болжауға көшуге мүмкіндік береді. Бұл симуляциялардың шынайылығы мен пайдалылығын арттырудың негізгі перспективалық бағыты ретінде қарастырылады.

Кесте 1 — Төтенше жағдайлар кезіндегі адамдарды эвакуациялау модельдерінің салыстырмалы сипаттамасы

Модель типі	**Негізгі идея**	**Артықшылықтар**	**Шектеулер**
Инженерлік-физикалық	Геометрияны, өткізу қабілетін және тығыздықты ескере отырып, ағындардың қозғалысын талдау	Есептеудің қарапайымдылығы; көрнекілігі; жобалық және нормативтік есептеулерге ыңғайлы болуы	Индивидуалды және психологиялық әрекеттерді ескермейді
Агенттік мінез-құлық модельдері (ABM)	Агенттердің жеке мінез-құлқын, қауіп-қатерді қабылдау, дүрбелең және эвакуация қатысушылары арасындағы өзара әрекеттестікті ескере отырып модельдеу	Көпшілікті шынайы модельдеу; жеке реакциялар мен мінез-құлық динамикасын есепке алу	Көптеген параметрлерді қажет етеді; жоғары есептеу күрделілігі

Оптимизациялық және гибридті алгоритмдер	Эвристикалық әдістер, мысалы, құмырсқа алгоритмі, генетикалық алгоритмдер және бөлшектердің рой алгоритмі, эвакуацияның оңтайлы маршруттарын іздеуде қолданылады	Адаптивтілік; өзгеріп отыратын жағдайларда эвакуациялық маршруттарды тиімді жоспарлау	Кіріс деректерінің сапасына тәуелділік; нақты жүйелерге интеграциялаудың күрделілігі
Машиналық оқыту және деректерді талдау (ML/AI)	Нақты деректер негізінде эвакуациялық шешімдер мен тәуекелдерді болжау	Жоғары дәлдік; жасырын заңдылықтарды анықтау; масштабталу қабілеті	Үлкен көлемдегі деректер жиынтықтарының қажеттілігі; модельдердің шектеулі түсіндіру мүмкіндігі
BIM–GIS және цифрлық егіздер	Эвакуация жолдарын есептеу үшін ғимараттың 3D моделін (BIM) геоақпараттық жүйелермен (GIS) интеграциялау	Эвакуацияны жобалау сатысында визуализациялау мүмкіндігі; ғимараттың геометриясын дәл есепке алу.	Толық цифрлық ғимарат модельдері қажет; енгізу үдерісі жоғары еңбек шығынын талап етеді

1-кестеде адамдардың төтенше жағдайлар кезіндегі қозғалысын бағалаудың әртүрлі тәсілдерін ұсынады. Инженерлік-физикалық модельдер ғимараттардың көлемін және шығу жолдарының өткізу қабілеттілігін ескере отырып, адамдар ағынын есептеуге негізделген. Олар жобалық есептеулер үшін ыңғайлы және оларды пайдалану салыстырмалы түрде оңай, бірақ психологиялық ерекшеліктерді және адамдардың мінез-құлқының ерекшеліктерін ескермейді. Агенттік мінез-құлық модельдері жеке адамдарды тәуелсіз агенттер ретінде имитациялау арқылы масса динамикасын және қауіп-қатерге реакцияны шынайы түрде көрсетуге мүмкіндік береді. Бұл әдістің қолданылуы модельдеудің дәлдігін жақсартады, бірақ үлкен есептеу ресурстарын және көптеген мінез-құлық деректерін талап етеді. Өзгермелі жағдайларда эвакуациялық маршруттарды табу үшін оңтайландыру алгоритмдері қолданылады. Олар жоспарлаудың бейімделгіштігін қамтамасыз етеді, алайда кіріс деректер сапасына және нақты жүйелермен интеграцияның күрделілігіне сезімтал. Ең заманауи бағыт – машиналық оқыту модельдері, олар шынайы деректер негізінде эвакуациялық шешімдерді болжауға және жасырын заңдылықтарды анықтауға мүмкіндік береді. Олардың басты шектеулері – үлкен деректер жиынтықтарына қажеттілік пен нәтижелерді түсіндірудің күрделілігі болып табылады. Жалпы алғанда, заманауи зерттеулер физикалық, мінез-құлықтық және интеллектуалды модельдерді біріктіретін гибридті шешімдерге ұмтылады, бұл эвакуациялық жүйелердің тиімділігі мен сенімділігін арттыруға бағытталған.

Жүргізілген әдебиеттер талдауы заманауи зерттеулер инженерлік модельдерді, мінез-құлық талдауын және машиналық оқыту әдістерін

біріктіретін гибридті интеллектуалды жүйелерге қарай ығысып бара жатқанын көрсетеді. Инженерлік және оңтайландыру алгоритмдері эвакуация ағындарын формализациялауды қамтамасыз етеді, агенттік мінез-құлық модельдері когнитивті және әлеуметтік факторларды ескеруге мүмкіндік береді, ал Random Forest және Gradient Boosting әдістері эмпирикалық деректерді жоғары дәлдікпен өңдеуді қамтамасыз етеді. Цифрлық егіздерді және BIM-GIS интеграциясын пайдалану нақты уақыттағы эвакуацияны басқарудың адаптивті жүйелерін құруға негіз қалайды. Алайда, нақты деректер жиынтығының шектеулігі және ML модельдерін түсіндірудің күрделілігі мәселесі өзекті болып қала береді, бұл түсіндірілетін жасанды интеллект бағытында және ғимараттардың жоспарлауы мен халық тығыздығы туралы деректерді интеграциялауда одан әрі зерттеулер жүргізу қажеттілігін анықтайды.

Материалдар мен әдістер. Зерттеу Toronto Fire Incidents Dataset (Kaggle, 2023) ашық деректер жинағын талдау негізінде орындалды. Бұл деректер жиынтығы 12 000-нан астам оқиға туралы мәліметтерді, соның ішінде ғимараттардың сипаттамаларын, өрт сөндіру қызметтерінің әрекет ету уақытын, өрт дабылы жүйесінің жұмысын, зардап шеккендердің санын және өртті сөндіру параметрлерін қамтиды.

Зерттеу мақсатында эвакуация динамикасымен және әрекет ету тиімділігімен тікелей байланысты келесі сипаттамалар таңдап алынды:

- Өртке жауап беру және оны жою уақыты(TFS_Alarm_Time,TFS_Arrival_Time, Fire_ Under_Control_Time);
- Өрт және түтін сигнализациясының болуы мен жағдайы (Fire_Alarm_System_Presence, Smoke_Alarm_at_Fire_Origin,Sprinkler_System_Presence);
- Ғимараттар мен өрт ошақтарының сипаттамалары (Property_Use, Level_Of_Origin, Area_of_Origin);
- Оқиға салдарының көрсеткіштері (Civilian_Casualties, Count_of_Persons_Rescued, Estimated_Dollar_Loss).

Бастапқы деректер негізінде келесі туынды параметрлер құрылды:

- Response_Time_min – өртке жауап беру уақыты (шақыру уақыты мен келген уақыт арасындағы айырмашылық);
- Fire_Control_Time_min – өртті сөндіру ұзақтығы;
- Has_Alarm_System – белсенді сигнализация жүйесінің болуы туралы бинарлық көрсеткіш;
- Has_Sprinkler – өрт сөндіру жүйесінің болуы және жұмыс жасауы туралы индикатор;
- High_Casualty_Risk – зардап шеккендердің қаупін немесе эвакуацияның кешігуін көрсететін мақсатты айнымалы.

Деректерді алдын ала өңдеу датасетті жүктеу және тазалауды, бос орындарды жоюды, уақыт белгілерін datetime форматына келтіруді, туынды белгілерді генерациялауды және категориялық деректерді кодтауды қамтыды. Одан әрі факторлардың корреляциялық талдауы, жауап беру уақыты мен зардап шеккендердің санының таралуын визуализациялау, сондай-ақ ғимараттардың

инженерлік сипаттамаларының оқиғалардың нәтижелеріне әсерін зерттеу жүргізілді.

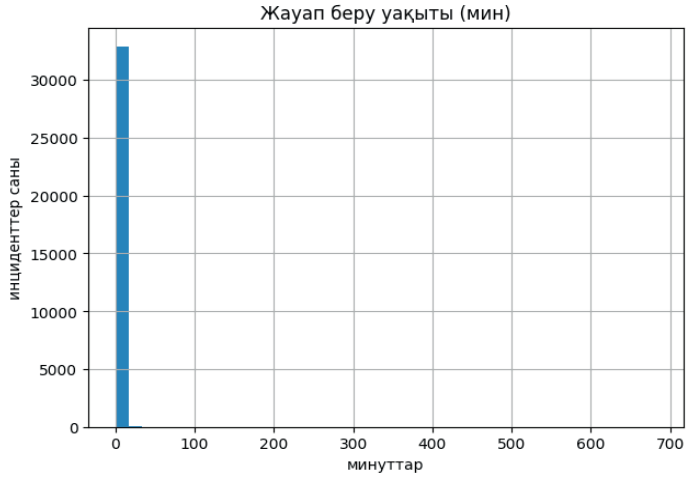
Модельдерді жасау және бағалау үшін логистикалық регрессия, Random Forest және градиентті бустинг (XGBoost) алгоритмдері қолданылды. Деректердің 20% – тестілеу үшін, ал 80% – модельдерді оқыту үшін қолданылды. Модельдердің тиімділігін бағалау үшін жіктеу көрсеткіштері (Accuracy, Precision, Recall, F1-score, ROC-AUC метрикаларымен), және регрессиялық көрсеткіштер (RMSE, MAE метрикаларымен) қолданылды.

Нәтижелерді интерпретациялау барысында ерекшеліктердің маңыздылығын талдау және параметрлердің үлесін бағалау үшін түсіндірілу мүмкін машиналық оқыту әдістері (SHAP values) қолданылды, мысалы: жауап беру уақыты, сигнализацияның болуы, қабат саны және эвакуацияның сәттілік ықтималдығы. Зерттеу Python бағдарламалау тілінде жүргізілді (pandas, scikit-learn, matplotlib, shap кітапханалары), бұл деректерді талдауды, модельдерді құруды және нәтижелерді визуализациялауды қамтамасыз етеді.

Нәтижелер

1. Алдын ала деректерді талдау

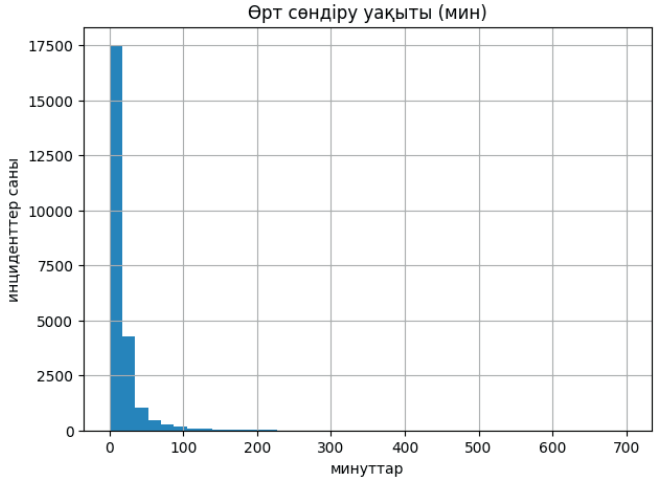
Toronto Fire Incidents деректер жиынтығына жүргізілген бастапқы талдау өрт оқиғаларының басым бөлігі тұрғын үй ғимараттарында орын алатынын және өртке қарсы қызметтердің әрекет ету уақытының салыстырмалы түрде қысқа екенін көрсетті. 1-суретте жауап беру уақытының (Response_Time_min) үлестірімі көрсетілген.



Сурет1 – Әрекет ету уақытының үлестірімі

Көптеген шақырулар 5 минут ішінде қабылданады, бұл қалалық қызметтердің жедел нормативтеріне сәйкес келеді. Алайда, 700 минутқа дейін созылған сирек жағдайлар кездеседі, олар, мүмкін, аномалды немесе қайталанатын жазбаларды көрсетеді.

2-суретте өртті сөндіру ұзақтығының (Control_Time_min) үлестірімі көрсетілген. Көптеген оқиғалар 30-40 минут ішінде аяқталады, бірақ үлестірімнің құйрығы бірнеше сағатқа созылған сөндіру жағдайларының аз мөлшерде бар екенін көрсетеді.



Сурет 2 – Өртті сөндіру уақытының үлестірімі

Мұндай асимметрия эвакуацияның күрделілігінің және адам шығынының ықтимал қауіпінің негізгі көрсеткіштері ретінде жауап беру және жою уақытын қарастыру қажеттігін растайды.Конец формы

2. Классификация нәтижелері

Жоғары эвакуациялық қаупінің (High_Casualty_Risk) ықтималдығын анықтау мақсатында Logistic regression, Random Forest, XGBoost модельдері қолданылды. Нәтижелер төмендегі 2 кестеде келтірілген.

Кесте 2 – Эвакуациялық тәуекелді болжау үшін машиналық оқыту модельдерінің классификация нәтижелері

Модель	Accuracy (классификация дәлдігі)	Precision (оң болжам-дардың дәлдігі)	Recall (толықтық модельдің сезімталдығы)	F1-score (Precision және Recall гармоникалық орташа мәні)	ROC-AUC (ROC-қисығы астындағы аудан, классификация сапасын бағалау)
Logistic regression	0.546	0.117	0.836	0.205	0.756
Random Forest	0.902	0.318	0.356	0.336	0.823
XGBoost	0.929	0.433	0.073	0.126	0.812

Логистикалық регрессия жоғары сезімталдыққа (Recall = 0.836) ие екенін көрсетті, яғни модель қауіпті жағдайлардың басым көпшілігін дұрыс анықтау қабілетіне ие, бірақ оның дәлдігі (Precision = 0.117) өте төмен. Бұл жалған дабылдардың көптігін көрсетеді, сондықтан мұндай модель алдын ала ескерту жүйелері үшін тиімді, бірақ жедел шешімдер қабылдауға жарамсыз.

Random Forest моделі Accuracy = 0.902 және ROC-AUC = 0.823 көрсеткіштерімен теңдестірілген сапаны көрсетті, ал Precision = 0.318 және Recall = 0.356 мәндері оқиғаларды анықтау мен жалған қосылуларды шектеу арасындағы жақсы үйлесімділікті көрсетеді.

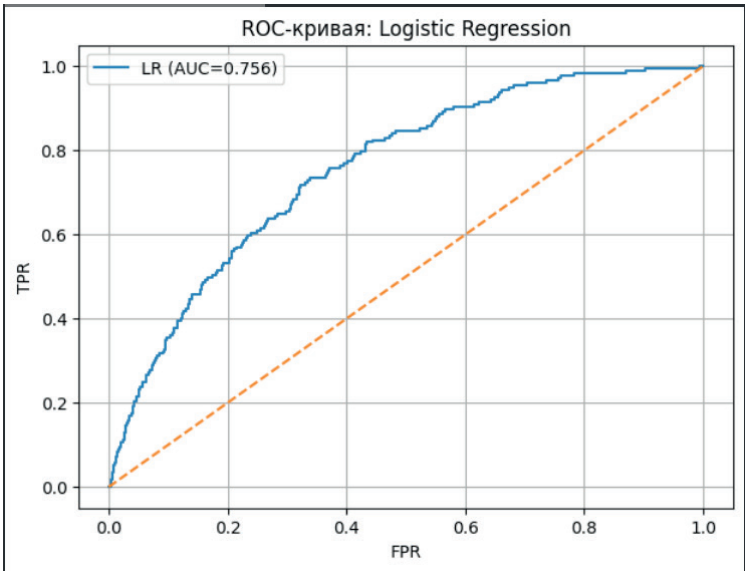
Модель күрделі сызықтық емес тәуелділіктерді жақсы анықтайды, бұл инженерлік (сигнализация жағдайы), уақыттық (тәулік уақыты, апта күні) және әлеуметтік (ғимаратты пайдалану түрі) сияқты көптеген факторлардың әсерін ескереді.

XGBoost моделі ең жоғары жалпы дәлдікті (Accuracy = 0.929) және жақсы ROC-AUC (0.812) көрсетті, бірақ оның сезімталдығы (Recall = 0.073) төмен болды – модель консервативті болжамдарға бейім және жоғары тәуекелді сирек кездесетін оқиғаларды жіберіп алуы мүмкін. Мұның себебі – кластардың тең еместігі: 12 687 жазбаның тек 887-і ғана High_Casualty_Risk санатына тиесілі.

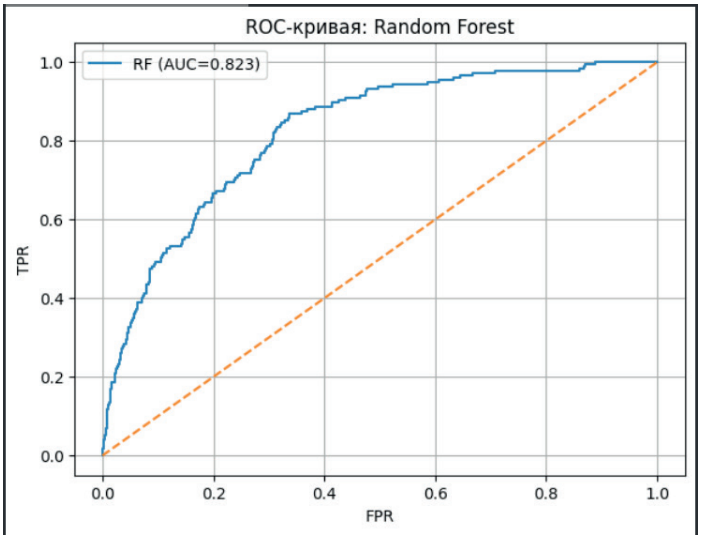
ROC қисықтарын интерпретациялау

3-5 суреттерде барлық қарастырылған модельдерге арналған ROC қисықтары көрсетілген:

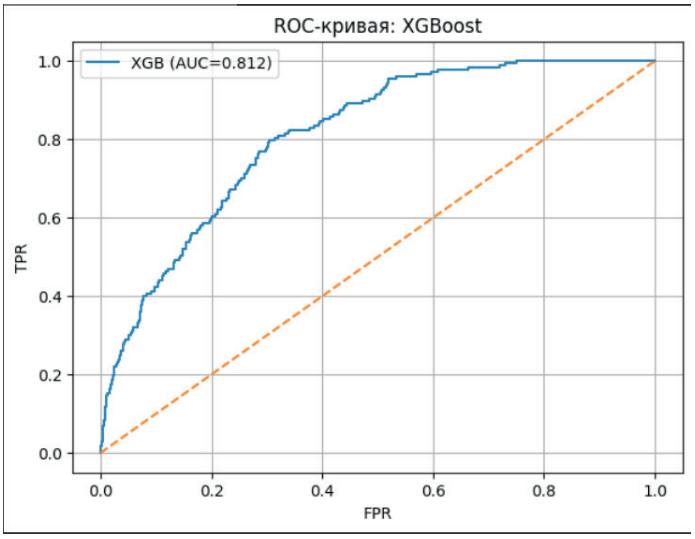
- Logistic Regression – AUC = 0.756
- Random Forest – AUC = 0.823
- XGBoost – AUC = 0.812



Сурет 3 – Логистикалық регрессия моделінің ROC-қисығы

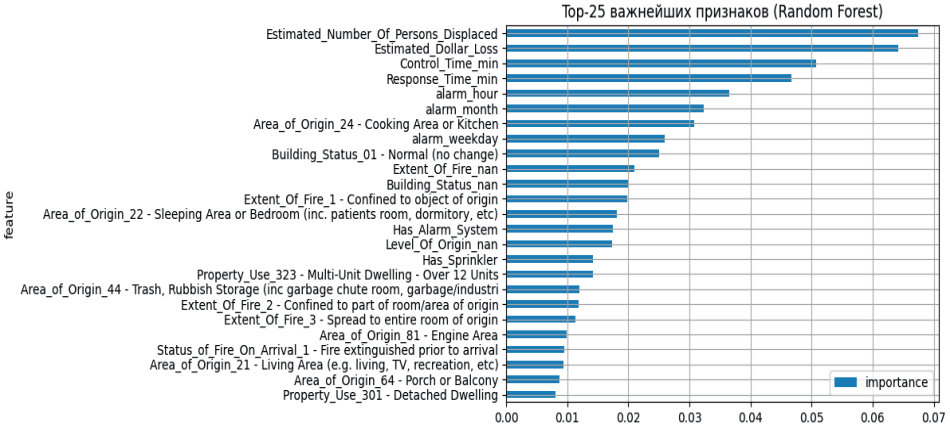


Сурет 4 – Random Forest моделінің ROC-қисығы



Сурет 5 – XGBoost моделінің ROC-қисығы

Random Forest моделі сыныптарды бөлуде және деректердегі шуға төзімділікте ең жақсы нәтиже көрсетті. Оның ROC-қисығы (4-сурет) диагональдан дерлік барлық ұзындығы бойынша жоғары орналасқан, бұл оң және теріс жағдайларды дұрыс жіктеу қабілетінің жоғары екенін растайды. XGBoost дәлдігі бойынша жақын, бірақ төмен FPR (Жалған оң нәтижелер үлесі) кезінде TPR-дің (Дұрыс оң нәтижелер үлесі) айқын төмендеуі бар, бұл басым сыныпта шамадан тыс оқытылу тенденциясын көрсетеді.



6 сурет. Random Forest моделінің маңызды белгілерінің графигі

Негізгі белгілерді талдау

Белгілердің маңыздылығының (feature importance) және SHAP-интерпретациясының (сурет 6) нәтижелері бойынша анықталды:

1. Response_Time_min – ең маңызды қауіп факторы. Өрт сөндіру қызметтерінің келу уақыты неғұрлым ұзақ болса, адам шығынының ықтималдығы соғұрлым жоғары болады.

2. Has_Alarm_System и Has_Sprinkler – жұмыс істеп тұрған дабыл және спринклер жүйелерінің болуы қауіпті азайтады. Олар болмаған жағдайда, оқиғаның High_Casualty_Risk санатына жату ықтималдығы екі еседен астам артады.

3. Extent_Of_Fire и Level_Of_Origin – өрттің таралу дәрежесі мен қабаттылығы эвакуация уақытын анықтайды: жоғарғы қабаттарда басталған өрттердің эвакуацияны қиындату ықтималдығы жоғары.

4. Уақытша белгілер (alarm_hour, alarm_weekday) – Түнгі және ерте таңғы уақытта тәуекелдің жоғары болуы реакцияның баяулауымен және көрінудің шектеулі болуымен түсіндіріледі.

Талқылау. Нәтижелер көрсеткендей, жауап беру уақыты мен дабыл жүйелерінің жағдайы сәтті эвакуацияның негізгі детерминанттары болып табылады. Тіпті Response_Time_min-ның 3-5 минутқа ұлғаюы адам шығыны ықтималдығын статистикалық тұрғыдан арттырады.

Практикалық тұрғыдан қарағанда:

- ML модельдері өрт сөндіру қызметтерінің шешім қабылдауды қолдау жүйелерінде оқиғаның күрделілігін ерте болжау үшін қолданылуы мүмкін;
- Random Forest мұндай міндеттер үшін оңтайлы модель болып табылады, ол дәлдік пен сезімталдық арасындағы тепе-теңдікті қамтамасыз етеді;
- болашақта эвакуацияны дәлірек модельдеу үшін ғимараттардың жоспарлауы, халық тығыздығы және қабаттылығы туралы деректерді пайдалана отырып, талдауды кеңейтуге болады.

Қорытынды. Жүргізілген зерттеу машиналық оқыту әдістері адамдарды биік ғимараттардан эвакуациялау процестерін талдау және болжау үшін тиімді қолданылатынын көрсетті. Toronto Fire Incidents ашық деректер жинағы негізінде эвакуацияның сәттілігіне және адам шығынының қауіп деңгейіне әсер ететін факторларды анықтауға бағытталған логистикалық регрессия, Random Forest және XGBoost модельдері жасалды және сыналды.

Зерттеу нәтижелері сәтті эвакуациялаудың ең маңызды факторлары өрт сөндірушілердің жауап беру уақыты, өрт дабылы жүйелері мен спринклерлердің болуы, сондай-ақ ғимараттың сипаттамалары – қабат саны, ауданы және өрттің шығу орны болып табылатындығын көрсетті. Пайдаланылған модельдердің ішінде Random Forest (AUC = 0.823) ең жақсы көрсеткіштерді көрсетті, дәлдік пен толықтық арасындағы оңтайл тепе-теңдікті қамтамасыз етті, бұл оны шешімдерді қолдау жүйелерінде практикалық қолдану үшін таңдаулы құралға айналдырды.

Талдау көрсеткендей, деректерді интеллектуалды талдау әдістерін пайдалану өрт қаупін бағалаудың дәлдігін арттыруға ғана емес, сонымен қатар болжамды бақылау жүйелерінің негізін құруға мүмкіндік беретінін растады. Мұндай шешімдерді қалалық құтқару қызметтерінің жұмысына біріктіру жауап беру уақытын қысқартуға, ресурстарды тиімдірек бөлуге және азаматтардың қауіпсіздік деңгейін арттыруға ықпал етеді. Бұл бағытты одан әрі дамыту ғимараттардың орналасуы туралы деректерді пайдалану, агенттік модельдеу арқылы адамдардың мінез-құлқын модельдеу және объектілердің цифрлық егіздерін құру арқылы мүмкін болады. Бұл оқиғаларды талдаудан нақты уақыттағы эвакуация сценарийлерін динамикалық болжауға және қалалық ортаның қауіпсіздігін басқарудың интеллектуалды жүйелерін қалыптастыруға көшуге мүмкіндік береді.

References

Aria (2020) Toronto Fire Incidents. Kaggle. URL: https://www.kaggle.com/datasets/phyxle/toronto-fire-incidents (in English)

Atyabi S., Kiavarz Moghaddam M., & Rajabifard A. (2019) Optimization of emergency evacuation in fire building by integrated BIM and GIS. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42. — P. 131-139. DOI: https://doi.org/10.5194/isprs-archives-XLII-4-W18-131-2019 (in English)

Bayat H., Delavar M.R., Barghi W., EslamiNezhad S.A., Hanachi P., & Zlatanova S. (2020) Modeling of emergency evacuation in building fire. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43. — P. 321-327. DOI: https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-321-2020 (in English)

Carev E., Rukomoinikov K., Makarov V., Anisimov I., & Anisimov N. (2024) Rescue system for people evacuation for highrise buildings. In E3S Web of Conferences, Vol. 486, 06008. DOI: https://doi.org/10.1051/e3sconf/202448606008 (in English)

Gao F., Du Z., Werner M., & Zhao Y. (2022) An improved optimization model for crowd evacuation considering individual exit choice preference. Transactions in GIS, 26(7). — P. 2850-2873. DOI: https://doi.org/10.1111/tgis.12984 (in English)

Gerges M., Demian P., & Adamu Z. (2021) Customising evacuation instructions for high-rise residential occupants to expedite fire egress: Results from agent-based simulation. Fire, 4(2), 21. DOI:

https://doi.org/10.3390/fire4020021 (in English)

Gravit M., Dmitriev I., & Kuzenkov K. (2018) Phased evacuation algorithm for high–rise buildings. In MATEC Web of Conferences,Vol. 245, 11012. DOI: https://doi.org/10.1051/matecconf/201824511012 (in English)

Han Z., Meng L., Mitani Y., Kawano K., Sugahara T., Taniguchi H., Honda H., & Li Z. (2025) Machine Learning-Based Assessment of Building Evacuation Vulnerability at the Pre-Disaster Stage. Sustainable Cities and Society, 130, 106571. DOI: https://doi.org/10.1016/j.scs.2025.106571 (in English)

Li N., Shi Z., Jin J., Feng J., Zhang A., Xie M., ... & Lei Y. (2024) Design of Intelligent Firefighting and Smart Escape Route Planning System Based on Improved Ant Colony Algorithm. Sensors, 24(19), 6438. DOI: https://doi.org/10.3390/s24196438 (in English)

Li Z., Qiu S., Wang X., & Zhao L. (2022) Modeling and simulation of crowd pre-evacuation decision-making in complex traffic environments. International Journal of Environmental Research and Public Health, 19(24), 16664. DOI: https://doi.org/10.3390/ijerph192416664 (in English)

Richardson O., Jalba A., & Muntean A. (2019) Effects of environment knowledge in evacuation scenarios involving fire and smoke: a multiscale modelling and simulation approach. Fire technology, 55(2). — P. 415-436. DOI: https://doi.org/10.1007/s10694-018-0743-x (in English)

Ronchi E., & Nilsson D. (2013) Fire evacuation in high-rise buildings: a review of human behaviour and modelling research. Fire science reviews, 2(1), 7. DOI: https://doi.org/10.1186/2193-0414-2-7 (in English)

Senanayake G.P., Kieu M., Zou Y., & Dirks K. (2024) Agent-based simulation for pedestrian evacuation: A systematic literature review. International Journal of Disaster Risk Reduction, 111, 104705. DOI: https://doi.org/10.1016/j.ijdrr.2024.104705 (in English)

Teng Q., Wang X., He W., Pan G., & Mao Y. (2024) An investigation into the influence of context effects on crowd exit selection under gender difference in indoor evacuation. Frontiers in Psychology, 15, 1417738. DOI: https://doi.org/10.3389/fpsyg.2024.1417738 (in English)

Testa E., Barros R.C., & Musse S.R. (2019) Crowdest: a method for estimating (and not simulating) crowd evacuation parameters in generic environments. The Visual Computer, 35(6). — P. 1119-1130. DOI: https://doi.org/10.1007/s00371-019-01684-9 (in English)

Wang Y., Ge J., & Comber A. (2022) An agent-based simulation model of pedestrian evacuation based on bayesian nash equilibrium. arXiv preprint arXiv:2211.14260. DOI: https://doi.org/10.48550/arXiv.2211.14260 (in English)

Zablotsky A., Kuperman M., & Bouzat S. (2024) Pedestrian evacuations with imitation of cooperative behavior. Physical Review E, 109(5), 054304. DOI: https://doi.org/10.1103/PhysRevE.109.054304 (in English)

Zhang L., Wu X., Lin H., Zhang M., & Liu Y. (2024) Experimental study on the synchronization mechanism and trigger characteristic density of vertical evacuation in crowds. Scientific reports, 14(1), 26182. DOI: https://doi.org/10.1038/s41598-024-77726-7 (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 218–235

https://doi.org/10.32014/2026.2518-1726.410

IRSTI 87.53.13; 20.53.00; 44.31.29;
UDC 004.94:66.074.3:662.613

© Khusain B., Telmanov M.*, Khusain A.B., Brodskiy A.R., Sass A.S., 2026.
«D.V. Sokolsky Institute of Fuel, Catalysis and Electrochemistry» JSC,
Almaty, Kazakhstan;
E-mail: merlan01tel@gmail.com

DIGITAL TWIN OF AN INTEGRATED EMISSION PURIFICATION AND DECARBONIZATION SYSTEM FOR THERMAL UNITS

Khusain Bolatbek — Candidate of technical sciences, professor of NAS RK, «D.V. Sokolsky Institute of Fuel, Catalysis and Electrochemistry» JSC, Almaty, Kazakhstan,
E-mail: b.khusain@ifce.kz, https://orcid.org/0000-0001-9588-1012;
Telmanov Merlan — Master, «D.V. Sokolsky Institute of Fuel, Catalysis and Electrochemistry» JSC, Almaty, Kazakhstan,
E-mail: merlan01tel@gmail.com, https://orcid.org/0009-0007-9193-8414;
Khusain Atabek — Master**,** «D.V. Sokolsky Institute of Fuel, Catalysis and Electrochemistry» JSC, Almaty, Kazakhstan,
E-mail: atabek.khussain@gmail.com, https://orcid.org/0000-0002-0899-6872;
Brodskiy Aleksandr — Candidate of chemical sciences, associate professor, «D.V. Sokolsky Institute of Fuel, Catalysis and Electrochemistry» JSC, Almaty, Kazakhstan,
E-mail: albrod@list.ru, https://orcid.org/0000-0001-6216-4738;
Sass Aleksandr — Candidate of chemical sciences, «D.V. Sokolsky Institute of Fuel, Catalysis and Electrochemistry» JSC, Almaty, Kazakhstan,
E-mail: alex-sass@ifce.kz, https://orcid.org/0000-0003-4049-6314.

Abstract. This paper considers a digital twin of an integrated emission purification and decarbonization system for thermal power units operating on fossil fuels. The proposed approach is based on representing a multi-module gas purification system as a virtual engineering model describing the structure of the physical system, the functional relationships between modules, and their main operating modes. Such a model makes it possible to reproduce the behavior of a real technological system in a digital environment and to use it for analysis, optimization, and prediction of technological processes. Digital twin technology relies on modern digital modeling tools that allow the operation of complex industrial systems to be studied at different stages of design and operation. The digital twin concept is based on contemporary scientific approaches to the development of digital models

of industrial systems (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018; ISO, 2021) and takes into account the technological features of flue gas purification and carbon dioxide capture processes (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011). The developed model describes an integrated structure of the gas purification system that combines the main technological elements. In particular, the system includes an electrostatic filtration module for removing particulate matter, a catalytic module for neutralizing harmful gaseous components, and a decarbonization module designed for carbon dioxide capture. The results of the study show that the use of a digital twin allows the interaction of individual modules to be analyzed, the system layout to be optimized, and the selection of technological parameters to be justified at the design stage. In addition, the digital model can be applied for simulation of technological processes, evaluation of system performance, and prediction of emission reduction indicators. The model also enables the analysis of different operating modes, assessment of system stability, and improvement of control algorithms. The practical significance of the proposed approach is confirmed by experimental and modeling studies reported in previously published works of the authors and other researchers (Khusain et al., 2018; Khussain et al., 2020; Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024; Khusain et al., 2024). The obtained results demonstrate the effectiveness of digital twins as tools for engineering analysis, technological optimization, and decision support in the development and implementation of integrated emission purification systems.

Keywords: digital twin, integrated purification system, decarbonization, gas purification, thermal power plants, CO_2 capture, process modeling

Financing. *This work was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under the project AP23488040 'Digital twin of the decarbonization process based on a system for capturing and utilizing carbon oxides from flue emissions of fossil fuel-fired thermal devices,' in accordance with contract No. 222/GF24-26 dated September 9, 2024.*

For citations: Khusain B., Telmanov M., Khusain A.B., Brodskiy A.R., Sass A.S. Digital twin of an integrated emission purification and decarbonization system for thermal units. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 218–235. DOI: https://doi.org/10.32014/2026.2518-1726.410

© Хусаин Б., Тельманов М.М.*, Хусаин А.Б., Бродский А.Р., Сасс А.С., 2026.

«Д.В. Сокольский атындағы Жанармай, катализ және электрохимия институты» АҚ, Алматы, Қазақстан.
E-mail: merlan01tel@gmail.com

ЖЫЛУ ҚОНДЫРҒЫЛАРЫНЫҢ ШЫҒАРЫНДЫЛАРЫН КЕШЕНДІ ТАЗАЛАУ ЖӘНЕ ДЕКАРБОНИЗАЦИЯЛАУ ЖҮЙЕСІНІҢ ЦИФРЛЫҚ ЕГІЗІ

Хусаин Болатбек — т.ғ.к, ҚР ҰҒА профессоры «Д.В. Сокольский атындағы Жанармай, катализ және электрохимия институты» АҚ, Алматы, Қазақстан,
E-mail: b.khusain@ifce.kz, https://orcid.org/0000-0001-9588-1012;
Тельманов Мерлан — магистр, «Д.В. Сокольский атындағы Жанармай, катализ және электрохимия институты» АҚ, Алматы, Қазақстан,
E-mail: merlan01tel@gmail.com, https://orcid.org/0009-0007-9193-8414;
Хусаин Атабек — магистр, «Д.В. Сокольский атындағы Жанармай, катализ және электрохимия институты» АҚ, Алматы, Қазақстан,
E-mail: atabek.khussain@gmail.com, https://orcid.org/0000-0002-0899-6872;
Бродский Александр — х.ғ.к., қауымдастырылған профессор - «Д.В. Сокольский атындағы Жанармай, катализ және электрохимия институты» АҚ, Алматы, Қазақстан,
E-mail: albrod@list.ru, https://orcid.org/0000-0001-6216-4738;
Сасс Александр — х.ғ.к., «Д.В. Сокольский атындағы Жанармай, катализ және электрохимия институты» АҚ, Алматы, Қазақстан,
E-mail: alex-sass@ifce.kz, https://orcid.org/0000-0003-4049-6314.

Аннотация. Бұл жұмыста қазба отынын пайдаланатын жылу энергетикалық қондырғыларынан бөлінетін зиянды шығарындыларды кешенді тазалау және декарбонизациялау жүйесінің цифрлық егізі қарастырылады. Ұсынылған тәсіл түтін газдарын тазалау және көмірқышқыл газын ұстау процестерін қамтитын мультимодульді газтазалау жүйесін виртуалды инженерлік модель түрінде көрсетуге негізделген. Мұндай модель физикалық жүйенің құрылымын, модульдер арасындағы функционалдық байланыстарды және негізгі жұмыс режимдерін сандық ортада бейнелеуге мүмкіндік береді. Цифрлық егіз технологиясы нақты өндірістік жүйенің жұмысын модельдеу, талдау және оңтайландыру үшін заманауи цифрлық құралдарды қолдануға негізделеді. Цифрлық егіз тұжырымдамасы өнеркәсіптік жүйелердің цифрлық модельдерін құруға арналған заманауи ғылыми тәсілдерге сүйенеді (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018; ISO, 2021) және түтін газдарын тазалау мен көмірқышқыл газын ұстау процестерінің технологиялық ерекшеліктерін ескереді (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011). Ұсынылған модель жүйенің негізгі технологиялық элементтерін біріктіретін интеграцияланған құрылымды сипаттайды. Атап айтқанда, жүйе құрамына қатты бөлшектерді жоюға арналған электрофильтр модулі, зиянды газдарды

бейтараптандыруға арналған каталитикалық модуль және көмірқышқыл газын ұстауға арналған декарбонизация модулі кіреді. Зерттеу нәтижелері цифрлық егізді қолдану жекелеген модульдердің өзара әсерін талдауға, жүйенің құрылымдық компоновкасын оңтайландыруға және технологиялық параметрлерді ғылыми негізде таңдауға мүмкіндік беретінін көрсетеді. Сонымен қатар цифрлық модель технологиялық процестерді имитациялық модельдеу, жүйенің тиімділігін бағалау және шығарындыларды азайту көрсеткіштерін болжау үшін қолданылуы мүмкін. Модель сонымен қатар әртүрлі жұмыс режимдерін талдауға, жүйенің тұрақтылығын бағалауға және басқару алгоритмдерін жетілдіруге мүмкіндік береді. Тәсілдің практикалық маңыздылығы эксперименттік және модельдік зерттеулер нәтижелерімен расталады (Khusain et al., 2018; Khussain et al., 2020; Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024; Khusain et al., 2024). Алынған нәтижелер цифрлық егіздерді инженерлік талдау, жүйені оңтайландыру және шешім қабылдауды қолдау құралы ретінде қолданудың тиімді екенін көрсетеді. Бұл тәсіл өнеркәсіптік экологиялық технологияларды жобалау мен дамытуда, сондай-ақ энергия және экология саласындағы инновациялық шешімдерді енгізуде маңызды рөл атқарады.

Түйін сөздер: цифрлық сыңары, кешенді тазарту жүйесі, көміртексіздендіру, газ тазарту, жылу қондырғылары, CO_2 - ні ұстау, процестерді модельдеу

© Хусаин Б., Тельманов М.М.*, Хусаин А.Б., Бродский А.Р., Сасс А.С., 2026.
АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского», Алматы, Казахстан.
E-mail: merlan01tel@gmail.com

ЦИФРОВОЙ ДВОЙНИК КОМПЛЕКСНОЙ СИСТЕМЫ ОЧИСТКИ И ДЕКАРБОНИЗАЦИИ ВЫБРОСОВ ТЕПЛОВЫХ УСТАНОВОК

Хусаин Болатбек — кандидат технических наук, профессор НАН РК, АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского», Алматы, Казахстан,
E-mail: b.khusain@ifce.kz, https://orcid.org/0000-0001-9588-1012;
Тельманов Мерлан — магистр, АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского», Алматы, Казахстан, E-mail: merlan01tel@gmail.com, https://orcid.org/0009-0007-9193-8414;
Хусаин Атабек — магистр, АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского», Алматы, Казахстан,
E-mail: atabek.khussain@gmail.com, https://orcid.org/0000-0002-0899-6872;
Бродский Александр — кандидат химических наук, ассоциированный профессор, АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского», Алматы, Казахстан,
E-mail: albrod@list.ru, https://orcid.org/0000-0001-6216-4738;
Сасс Александр — кандидат химических наук, АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского», Алматы, Казахстан,
E-mail: alex-sass@ifce.kz, https://orcid.org/0000-0003-4049-6314.

Аннотация. В работе рассматривается цифровой двойник комплексной системы очистки и декарбонизации выбросов тепловых энергетических установок, функционирующих на основе ископаемого топлива. Предлагаемый подход основан на представлении мультимодульной системы газоочистки в виде виртуальной инженерной модели, описывающей структуру физической системы, функциональные связи между её модулями и основные режимы работы. Такая модель позволяет воспроизводить поведение реальной технологической системы в цифровой среде и использовать её для анализа, оптимизации и прогнозирования технологических процессов. Концепция цифрового двойника опирается на современные научные подходы к созданию цифровых моделей промышленных систем (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018; ISO, 2021) и учитывает технологические особенности процессов очистки дымовых газов и улавливания диоксида углерода (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011). Разработанная модель описывает интегрированную структуру системы газоочистки, объединяющую ключевые технологические элементы: модуль электрофильтрации для удаления твёрдых частиц, каталитический модуль для нейтрализации вредных газовых компонентов и модуль декарбонизации для улавливания диоксида углерода. Результаты исследования показывают, что использование цифрового двойника позволяет анализировать взаимное влияние отдельных модулей системы, оптимизировать её конфигурацию и обосновывать выбор технологических параметров на этапе проектирования. Кроме того, цифровая модель может применяться для имитационного моделирования технологических процессов, оценки эффективности системы и прогнозирования показателей снижения выбросов. Модель также обеспечивает возможность анализа различных режимов работы системы, оценки её устойчивости и совершенствования алгоритмов управления. Практическая значимость предложенного подхода подтверждается результатами экспериментальных и модельных исследований, представленными в ранее опубликованных работах авторов и других исследователей (Khusain et al., 2018; Khussain et al., 2020; Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024). Полученные результаты демонстрируют эффективность использования цифровых двойников как инструмента инженерного анализа, оптимизации технологических процессов и поддержки принятия решений при разработке и внедрении комплексных систем очистки выбросов.

Ключевые слова: цифровой двойник; комплексная система очистки; декарбонизация; газоочистка; тепловые установки; улавливание CO_2; моделирование процессов

Работа выполнена при финансовой поддержке ГУ «Комитет науки Министерства науки и высшего образования Республики Казахстан» по проекту AP23488040 «Цифровой двойник процесса декарбонизации на основе системы сбора и утилизации оксидов углерода дымовых выбросов тепловых

устройств, использующих ископаемое топливо» согласно договору № 222/ ГФ24-26 от 09.09.2024 г.

Введение. Проблема снижения выбросов загрязняющих веществ и парниковых газов тепловыми установками остаётся одной из ключевых задач энергетики и промышленности. Основными компонентами дымовых газов являются твёрдые минеральные частицы, оксиды азота, диоксид серы, монооксид углерода и диоксид углерода, совокупное воздействие которых приводит к ухудшению качества атмосферного воздуха и усилению антропогенного воздействия на окружающую среду. Современные подходы к улавливанию и утилизации CO_2 рассматриваются как важный элемент стратегии декарбонизации энергетического сектора (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011).

Традиционные системы газоочистки, как правило, проектируются для удаления отдельных классов загрязняющих веществ и слабо, учитывают взаимосвязь между отдельными технологическими модулями. Это ограничивает возможности оптимизации компоновки оборудования и согласования режимов работы системы в целом. В последние годы всё большее внимание уделяется применению цифровых двойников как инструмента моделирования и анализа сложных инженерных систем (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018). Согласно международным стандартам, цифровой двойник представляет собой виртуальное представление физического объекта, связанное с ним на протяжении жизненного цикла и используемое для анализа, оптимизации и прогнозирования поведения системы (ISO, 2021).

Ранее авторами были выполнены исследования, посвящённые моделированию аэродинамики газовых потоков, процессам нейтрализации токсичных компонентов выбросов и разработке комплексных систем очистки дымовых газов (Khusain et al., 2018; Khussain et al., 2020; Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024). В более поздних работах был предложен подход к созданию цифровых двойников систем очистки с интеграцией модулей декарбонизации, ориентированный на задачи проектирования и визуализации технологических процессов (Khusain et al., 2024). Эти результаты создают основу для дальнейшего развития инженерных цифровых двойников комплексных систем очистки и улавливания CO_2.

Целью настоящей работы является разработка и описание концепции цифрового двойника комплексной системы очистки и декарбонизации выбросов тепловых установок, ориентированной на инженерное сопровождение процессов проектирования и внедрения мультимодульных систем газоочистки.

Литературный обзор. В современной научной литературе цифровые двойники рассматриваются как один из ключевых инструментов цифровизации промышленности, обеспечивающий моделирование, анализ и оптимизацию сложных технологических систем. Согласно работам (Grieves and Vickers,

2017, Tao et al., 2019; Kritzinger et al., 2018, Perno et al., 2021; Javaid et al., 2023) а также положениям стандарта (ISO, 2021), цифровой двойник представляет собой связанную с физическим объектом цифровую модель, позволяющую воспроизводить его структуру, состояние и режимы функционирования в виртуальной среде. Такой подход особенно актуален для многокомпонентных инженерных систем, где требуется учитывать взаимосвязь между отдельными модулями и изменяющимися технологическими параметрами.

Важное место в литературе занимают исследования, посвященные очистке дымовых газов и декарбонизации выбросов тепловых установок. В работах (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011) показано, что технологии улавливания CO_2 являются важным элементом снижения антропогенной углеродной нагрузки, особенно для энергетических объектов, работающих на ископаемом топливе. При этом эффективность улавливания углекислого газа напрямую зависит от предварительной подготовки газового потока, включая удаление твердых частиц и токсичных примесей. Это обосновывает необходимость комплексного рассмотрения электрофильтрационного, каталитического и декарбонизационного модулей как единой технологической системы.

Теоретической базой для описания процессов, протекающих в отдельных модулях системы, служат классические труды по химической инженерии, процессам массопереноса, теплопереноса, каталитической кинетике и сорбции. Эти подходы позволяют формировать инженерно обоснованные математические модели, пригодные для цифрового воспроизведения работы системы в различных режимах. В контексте данной статьи это особенно важно, поскольку предлагаемый цифровой двойник описывает не отдельный аппарат, а мультимодульную структуру, включающую электрофильтрацию, каталитическую нейтрализацию и улавливание CO_2 (Fogler, 2016; Levenspiel, 1999; Perry, Green, 2019).

В предыдущих наших работах были исследованы аэродинамика выбросов, свойства каталитических материалов, процессы предварительной очистки дымовых газов и подходы к созданию цифровых моделей экологических систем. Эти исследования послужили основой для разработки представленного в статье цифрового двойника комплексной системы очистки и декарбонизации выбросов (Khusain et al., 2018; Khussain et al., 2020; Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024; Khussain et al., 2022; Khussain et al., 2024; Khussain et al., 2025).

Таким образом, анализ литературы показывает, что, несмотря на значительное число работ по цифровым двойникам и отдельным технологиям газоочистки и улавливания CO_2, вопрос создания единой цифровой модели комплексной системы очистки и декарбонизации выбросов остается недостаточно раскрытым. Именно этот аспект составляет научную основу представленного исследования.

Материалы и методы

1. Комплексная система очистки и декарбонизации выбросов

Комплексная система очистки и декарбонизации выбросов тепловых установок представляет собой мультимодульную технологическую структуру, предназначенную для последовательного удаления основных компонентов загрязняющих выбросов. В состав системы входят модули удаления твёрдых минеральных частиц, модули нейтрализации газообразных загрязнителей и модули улавливания диоксида углерода. Подобная компоновка позволяет обеспечить глубокую очистку дымовых газов и создать условия для последующей декарбонизации выбросов.

Процессы удаления твёрдых частиц и нейтрализации газообразных компонентов сопровождаются сложными аэродинамическими и физико-химическими явлениями, что требует учёта взаимного влияния отдельных модулей. Ранее авторами были показаны особенности прохождения газовых потоков в системах нейтрализации и влияние режимов работы оборудования на эффективность очистки (Khusain et al., 2018). Использование каталитических систем на основе термоустойчивых носителей позволяет эффективно удалять токсичные компоненты выбросов в широком диапазоне температур (Khusain et al., 2020; Khusain et al., 2021).

Модуль улавливания диоксида углерода является завершающим элементом системы и основан на процессах сорбции и десорбции CO_2. Применение сорбционных материалов, в том числе разработанных и исследованных авторами, позволяет интегрировать процессы улавливания CO_2 в общую технологическую схему без существенного усложнения оборудования (Khusain et al., 2024; Khussain et al., 2025). Комплексный характер системы требует согласования режимов работы всех модулей, что делает использование цифрового двойника особенно актуальным.

2. Концепция цифрового двойника системы

Цифровой двойник комплексной системы очистки и декарбонизации выбросов представляет собой виртуальную инженерную модель, воспроизводящую структуру и функциональные связи физической установки. Концепция цифрового двойника основана на модульном принципе, при котором каждый элемент системы описывается в виде отдельного виртуального блока с заданными входными и выходными параметрами (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018, Perno et al., 2021).

В цифровом двойнике реализуется моделирование последовательного прохождения газового потока через все модули системы с учётом изменения температуры, состава и расхода газа. Такой подход позволяет анализировать влияние параметров отдельных модулей на эффективность системы в целом и выявлять критические режимы эксплуатации. Подобная логика моделирования соответствует современным представлениям о цифровых двойниках промышленных систем, зафиксированным в международных стандартах (ISO, 2021).

Ранее авторами был предложен и реализован подход к созданию

цифровых двойников систем очистки дымовых газов с интеграцией модуля декарбонизации, ориентированный на задачи моделирования и визуализации технологических процессов (Khusain et al., 2024). Настоящая работа развивает этот подход в направлении комплексного инженерного анализа и поддержки проектирования мультимодульных систем очистки.

Общая структура комплексной системы очистки и декарбонизации выбросов, используемой в цифровом двойнике, представлена на рисунке 1.

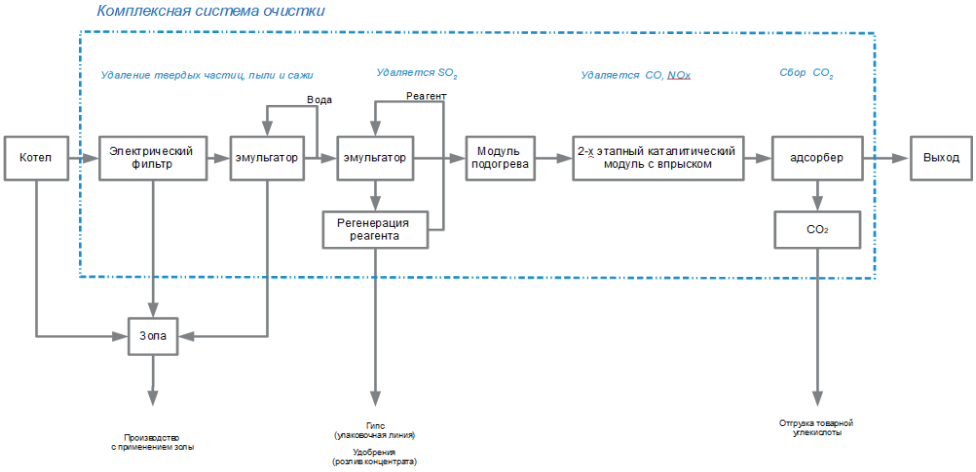


Рисунок 1 – Структурная схема комплексной системы очистки и декарбонизации выбросов тепловых установок.

3. Архитектура цифрового двойника системы

Цифровой двойник комплексной системы очистки и декарбонизации выбросов представляет собой интегрированную виртуальную среду, предназначенную для воспроизведения структуры и режимов функционирования физической установки. Архитектура цифрового двойника формируется на основе модульного принципа, при котором каждый элемент технологической системы представлен отдельным программным блоком с определёнными входными и выходными параметрами. Подобный подход соответствует современным концепциям построения цифровых двойников промышленных систем (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018).

Функциональная архитектура цифрового двойника включает несколько уровней: уровень физической системы, уровень виртуальной модели и уровень взаимодействия пользователя с системой. На уровне физической системы располагаются реальные технологические модули очистки дымовых газов и улавливания диоксида углерода. Эти модули обеспечивают удаление твёрдых частиц, нейтрализацию газообразных загрязнителей и улавливание CO_2.

На уровне виртуальной модели формируется математическое и алгоритмическое описание процессов, происходящих в каждом модуле системы. Виртуальная модель воспроизводит основные характеристики газового потока, включая температуру, давление, скорость движения газа и концентрацию отдельных компонентов газовой смеси. Такие модели широко используются при анализе технологических процессов в химической инженерии и энергетике (Fogler, 2016; Levenspiel, 1999; Perry and Green, 2019).

Уровень взаимодействия пользователя включает средства визуализации и анализа результатов моделирования. Этот уровень обеспечивает мониторинг технологических параметров, анализ режимов работы системы и оценку эффективности очистки выбросов при различных условиях эксплуатации. В соответствии с международными стандартами цифровых двойников такие системы могут использоваться для анализа, оптимизации и прогнозирования поведения технологических установок на протяжении всего жизненного цикла оборудования (ISO, 2021).

В архитектуре цифрового двойника реализуется последовательная передача данных между отдельными модулями системы. Изменение параметров газового потока после прохождения одного модуля непосредственно влияет на условия работы последующих модулей. Такой подход позволяет учитывать взаимосвязь процессов очистки и декарбонизации и обеспечивает более реалистичное моделирование функционирования всей системы.

Дополнительным элементом архитектуры является база параметров технологических процессов, включающая характеристики газовых потоков, свойства каталитических материалов и параметры сорбционных процессов. Использование такой базы данных позволяет адаптировать цифровой двойник к различным условиям эксплуатации и применять его для анализа различных вариантов компоновки системы очистки.

4. Моделирование основных технологических процессов

Моделирование технологических процессов в цифровом двойнике комплексной системы очистки и декарбонизации выбросов направлено на воспроизведение ключевых физических и химических преобразований, происходящих в каждом модуле системы. Основной задачей является получение инженерно адекватного описания процессов, достаточного для анализа режимов работы и оптимизации системы в целом, без избыточного усложнения вычислительных моделей, что соответствует классическим подходам химической инженерии (Fogler, 2016; Levenspiel, 1999; Perry and Green, 2019).

4.1 Моделирование газодинамики потока

Моделирование движения газового потока в элементах системы очистки основано на использовании уравнений сохранения массы, импульса и энергии. Эти уравнения позволяют описывать распределение параметров газового потока в различных элементах технологической системы и оценивать влияние геометрии оборудования на эффективность очистки.

Математическое описание движения газового потока в элементах системы может быть представлено с использованием уравнения баланса массы. В общем виде баланс массы для стационарного потока может быть записан следующим образом:

$$\dot{m}_{in} = \dot{m}_{out} \quad (1)$$

где,

$\dot{m}_{in}$ - массовый расход газа на входе в модуль,

$\dot{m}_{ou}$ - массовый расход газа на выходе из модуля.

Данное уравнение отражает сохранение массы при прохождении газового потока через технологический элемент системы очистки. Подобные зависимости широко применяются при инженерном моделировании процессов химической технологии и газоочистки (Fogler, 2016; Levenspiel, 1999).

Дополнительно учитывается уравнение сохранения энергии, позволяющее описывать изменение температуры газовой смеси при прохождении через элементы системы. Температурные изменения оказывают значительное влияние на эффективность каталитических реакций и процессы сорбции диоксида углерода.

Для описания движения газового потока используются зависимости, связывающие скорость потока, давление и гидравлическое сопротивление элементов системы. Эти зависимости позволяют учитывать влияние конструктивных особенностей оборудования на распределение потоков газа. Использование подобных моделей обеспечивает возможность анализа различных режимов работы системы очистки и позволяет прогнозировать изменение параметров газового потока при изменении условий эксплуатации.

4.2 Моделирование каталитических процессов

Каталитические процессы нейтрализации газообразных загрязнителей играют важную роль в обеспечении эффективности очистки дымовых газов. В цифровом двойнике эти процессы описываются с использованием кинетических моделей, отражающих зависимость скорости химических реакций от температуры, состава газовой смеси и времени контакта газа с каталитической поверхностью.

Каталитическая нейтрализация токсичных компонентов дымовых газов основана на реакциях окисления и восстановления. В частности, процессы окисления монооксида углерода и углеводородов приводят к образованию диоксида углерода и воды, тогда как восстановительные реакции способствуют удалению оксидов азота из газовой смеси.

Скорость каталитической реакции может быть описана с использованием кинетического уравнения, отражающего зависимость скорости реакции от концентрации реагентов и температуры:

$$r = kC^n \quad (2)$$

где,
r — скорость химической реакции,
k — константа скорости реакции,
C — концентрация реагента,
n - порядок реакции.

Температурная зависимость константы скорости реакции определяется уравнением Аррениуса:

$$k=k_0 e^{-\frac{E_a}{RT}} \quad (3)$$

где,
k_0 - предэкспоненциальный множитель,
E_a - энергия активации реакции,
R - универсальная газовая постоянная,
T - абсолютная температура.

Подобные кинетические зависимости широко используются при анализе каталитических процессов в химической инженерии (Fogler, 2016; Levenspiel, 1999; Perry and Green, 2019).

Кинетические параметры реакций определяются на основе экспериментальных данных и результатов исследований каталитических систем. Использование кинетических моделей позволяет оценивать степень превращения загрязняющих компонентов в зависимости от температуры и времени контакта газа с каталитическим слоем. Подобные подходы широко используются при анализе каталитических реакторов и процессов химической технологии (Fogler, 2016; Levenspiel, 1999; Perry and Green, 2019).

Важным фактором является распределение температуры в каталитическом реакторе. Повышение температуры способствует ускорению реакций, однако чрезмерные температурные значения могут приводить к деградации каталитических материалов. Поэтому моделирование температурных режимов является важным элементом цифрового двойника.

4.3 Моделирование процессов улавливания CO_2

Моделирование процессов улавливания диоксида углерода основано на описании сорбционных процессов, происходящих при взаимодействии газовой смеси с сорбционным материалом. В цифровом двойнике используются модели адсорбции и десорбции CO_2, позволяющие оценивать эффективность улавливания углекислого газа при различных температурных и давленческих условиях.

Процессы адсорбции диоксида углерода могут быть описаны с использованием изотермы Лэнгмюра, которая широко применяется для

моделирования сорбционных процессов (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011):

$$q = \frac{q_{max} bP}{1+bP} \quad (4)$$

где,

q - количество адсорбированного CO_2,

q_{max} - максимальная сорбционная ёмкость сорбента,

b - константа адсорбции,

P - парциальное давление CO_2.

Сорбционные процессы характеризуются циклическим режимом работы, включающим стадии адсорбции, регенерации сорбента и повторного использования материала. При моделировании учитываются свойства сорбционного материала, включая удельную поверхность, пористость и способность к связыванию диоксида углерода.

Подобные технологии широко рассматриваются в современных исследованиях по улавливанию и хранению углекислого газа и являются важным элементом стратегий декарбонизации энергетики (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011, Raganati et al., 2021; Hasan et al., 2024, Sharma et al., 2024). Использование математических моделей сорбционных процессов позволяет оценивать влияние параметров процесса на эффективность улавливания CO_2 и оптимизировать режимы работы оборудования.

Цифровой двойник позволяет анализировать влияние параметров сорбционного процесса на эффективность улавливания CO_2 и оценивать энергетические затраты, связанные с регенерацией сорбента. Такой подход позволяет оптимизировать параметры работы модуля декарбонизации и повысить эффективность всей системы очистки выбросов.

Моделирование процессов удаления твёрдых минеральных частиц и сажи основано на учёте аэродинамических характеристик газового потока и дисперсного состава частиц. Ранее авторами были показаны особенности прохождения газовых потоков в элементах систем нейтрализации и влияние геометрии и режимов работы оборудования на эффективность очистки (Khusain et al., 2018). Эти зависимости используются в цифровом двойнике в виде параметрических моделей, позволяющих оценивать степень очистки при различных условиях эксплуатации.

Процессы нейтрализации газообразных загрязнителей моделируются с учётом протекания каталитических реакций окисления и восстановления токсичных компонентов. В цифровом двойнике используются обобщённые зависимости, отражающие влияние температуры, времени контакта и состава газовой смеси на степень удаления оксидов азота, диоксида серы и монооксида углерода. Экспериментальные и модельные исследования каталитических систем, выполненные авторами ранее, подтверждают эффективность такого

подхода для инженерных расчётов (Khusain et al., 2020; Khusain et al., 2021; Khussain et al., 2022; Khussain et al., 2024).

Важным элементом моделирования является учёт взаимного влияния модулей системы. Изменение характеристик газового потока после одного этапа очистки непосредственно отражается на условиях работы последующих модулей, что требует согласованного моделирования всей технологической цепочки. Цифровой двойник обеспечивает возможность анализа таких взаимодействий и выявления режимов, обеспечивающих максимальную суммарную эффективность очистки и декарбонизации.

5. Использование цифрового двойника на этапах проектирования и внедрения

Применение цифрового двойника на этапах проектирования комплексной системы очистки и декарбонизации выбросов позволяет существенно повысить обоснованность инженерных решений. Виртуальная модель используется для анализа различных вариантов компоновки системы, подбора рациональных параметров оборудования и предварительной оценки эффективности очистки до физической реализации установки (Grieves and Vickers, 2017; Tao et al., 2019; Kritzinger et al., 2018; ISO, 2021).

На стадии проектирования цифровой двойник применяется для оценки влияния последовательности расположения модулей и их производительности на итоговые показатели очистки. Такой подход особенно важен для объектов с переменной нагрузкой и нестабильным составом выбросов, где традиционные методы проектирования не позволяют в полной мере учесть динамику процессов. Использование цифрового двойника снижает потребность в многоэтапных натурных испытаниях и позволяет выявлять потенциальные ограничения системы на ранних этапах (Khusain et al., 2024).

На этапе внедрения цифровой двойник используется для поддержки пусконаладочных работ и адаптации системы к реальным условиям эксплуатации. Сопоставление расчётных и фактических параметров работы установки позволяет корректировать режимы работы отдельных модулей и обеспечивать согласованное функционирование всей системы. Ранее авторами было показано, что использование цифровых двойников в задачах моделирования и визуализации процессов очистки дымовых газов способствует повышению устойчивости работы систем газоочистки (Khusain et al., 2024).

Кроме того, цифровой двойник может использоваться при модернизации и масштабировании комплексных систем очистки. На его основе возможно обоснование внедрения новых модулей, замены каталитических систем или интеграции усовершенствованных сорбционных материалов для улавливания CO_2 (Khusain et al., 2024; Khussain et al., 2025). Это обеспечивает преемственность инженерных решений и повышает гибкость применения системы в промышленных условиях.

6. Практическая значимость и области применения

Практическая значимость цифрового двойника комплексной системы очистки и декарбонизации выбросов заключается в возможности его применения для решения прикладных инженерных задач в энергетике и промышленности. Виртуальная модель позволяет обосновывать выбор технологических решений, оценивать эффективность очистки и снижать риски, связанные с проектированием и внедрением мультимодульных систем газоочистки.

Цифровой двойник может применяться при проектировании и эксплуатации тепловых электростанций, промышленных котельных, а также предприятий металлургической, нефтехимической и химической промышленности. Использование разработанных ранее каталитических и сорбционных решений позволяет адаптировать модель под конкретные условия эксплуатации и требования к степени очистки выбросов (Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024; Khussain et al., 2022; Khussain et al., 2024; Khussain et al., 2025).

Дополнительной областью применения является инженерное сопровождение эксплуатации систем очистки. Цифровой двойник может использоваться для анализа отклонений режимов работы, оценки последствий изменения состава топлива и формирования рекомендаций по оптимизации параметров эксплуатации. Это способствует снижению эксплуатационных затрат и повышению экологической эффективности промышленных объектов.

7. Экологический эффект применения системы

Применение комплексных систем очистки и декарбонизации выбросов способствует значительному снижению негативного воздействия энергетических и промышленных объектов на окружающую среду. Использование интегрированных технологических решений позволяет одновременно уменьшать выбросы твёрдых частиц, токсичных газов и парниковых газов.

Снижение выбросов загрязняющих веществ достигается благодаря последовательному прохождению дымовых газов через несколько технологических модулей. Удаление твёрдых частиц предотвращает поступление в атмосферу минеральной пыли и продуктов неполного сгорания топлива. Каталитические процессы нейтрализации обеспечивают преобразование токсичных компонентов газовой смеси в менее опасные соединения.

Особое значение имеет улавливание диоксида углерода, который является одним из основных парниковых газов. Использование технологий улавливания CO_2 рассматривается как важный элемент глобальных стратегий декарбонизации энергетики и промышленности (Bui et al., 2018; Rochelle, 2009; Wang et al., 2011, Raganati et al., 2021; Hasan et al., 2024).

Использование цифровых двойников при проектировании и эксплуатации систем очистки позволяет более точно прогнозировать экологический эффект внедрения таких технологий. Это способствует принятию обоснованных

инженерных решений и повышает эффективность экологической модернизации энергетических объектов.

Результаты исследования. В результате проведённого исследования была разработана концептуальная модель цифрового двойника комплексной системы очистки и декарбонизации выбросов тепловых установок. Созданная модель отражает структуру технологической системы и обеспечивает воспроизведение взаимодействия между её основными модулями, включая модуль электрофильтрации, каталитический модуль и модуль улавливания диоксида углерода.

Реализация цифрового двойника позволила смоделировать последовательное прохождение дымовых газов через все элементы системы и оценить изменение параметров газового потока на каждом этапе очистки. В ходе моделирования учитывались такие параметры, как температура газа, скорость потока, концентрация загрязняющих компонентов и эффективность их удаления в каждом технологическом модуле.

Полученные результаты показали, что использование интегрированной мультимодульной системы очистки обеспечивает существенное снижение концентрации загрязняющих веществ. Электрофильтрационный модуль обеспечивает эффективное удаление твердых частиц, каталитический модуль способствует нейтрализации токсичных газообразных компонентов, а модуль декарбонизации позволяет улавливать значительную часть диоксида углерода.

Применение цифрового двойника позволило также провести анализ различных режимов функционирования системы. Моделирование показало, что изменение параметров отдельных модулей оказывает значительное влияние на эффективность работы всей технологической цепочки. В частности, температура газового потока и время контакта газа с каталитическим слоем оказывают существенное влияние на эффективность каталитических реакций, тогда как свойства сорбционного материала и режим регенерации определяют эффективность улавливания CO_2.

Результаты моделирования подтверждают, что цифровой двойник может использоваться как инструмент инженерного анализа и оптимизации параметров комплексных систем очистки выбросов.

Обсуждение результатов. Полученные результаты демонстрируют значительный потенциал применения цифровых двойников при разработке и оптимизации систем очистки дымовых газов и улавливания углекислого газа. Использование виртуальной модели позволяет учитывать взаимное влияние технологических процессов, протекающих в различных модулях системы, что трудно реализовать при традиционных методах проектирования.

Одним из ключевых преимуществ цифрового двойника является возможность анализа различных конфигураций системы на ранних этапах проектирования. Виртуальное моделирование позволяет изменять параметры модулей, последовательность их расположения и режимы работы без

необходимости проведения дорогостоящих экспериментальных испытаний. Это существенно снижает затраты времени и ресурсов при разработке новых экологических технологий.

Результаты исследования согласуются с современными научными представлениями о применении цифровых двойников в промышленности, где подобные модели используются для анализа, оптимизации и прогнозирования работы сложных технологических систем (Grieves and Vickers, 2017; Tao et al., 2019). Применительно к системам газоочистки цифровые двойники позволяют учитывать взаимосвязь аэродинамических, химических и сорбционных процессов.

Следует отметить, что разработанная модель носит инженерно-ориентированный характер и предназначена прежде всего для анализа и оптимизации технологических параметров системы. Дальнейшее развитие цифрового двойника может быть связано с интеграцией данных реальных измерительных приборов и систем мониторинга, что позволит реализовать режимы оперативного контроля и адаптивного управления процессами очистки выбросов.

Таким образом, использование цифрового двойника открывает новые возможности для повышения эффективности комплексных систем очистки дымовых газов и внедрения технологий декарбонизации в энергетическом и промышленном секторе.

Заключение. В работе представлена концепция цифрового двойника комплексной системы очистки и декарбонизации выбросов тепловых установок, ориентированная на инженерное сопровождение процессов проектирования и внедрения мультимодульных систем газоочистки. Показано, что использование цифрового двойника позволяет учитывать взаимосвязь отдельных технологических модулей и анализировать их совместную работу в рамках единой виртуальной модели.

Предложенный подход обеспечивает возможность виртуальной оптимизации компоновки системы, выбора рациональных режимов эксплуатации и предварительной оценки эффективности процессов очистки и улавливания диоксида углерода. Опора на ранее выполненные исследования авторов в области газоочистки, каталитических систем, сорбентов CO_2 и цифровых двойников подтверждает практическую реализуемость и прикладную значимость предложенной концепции (Khusain et al., 2018; Khussain et al., 2020; Khusain et al., 2020; Khusain et al., 2021; Khusain et al., 2024; Khusain et al., 2024; Khussain et al., 2022; Khussain et al., 2024; Khussain et al., 2025).

Полученные результаты свидетельствуют о целесообразности применения цифровых двойников при решении задач декарбонизации и комплексной очистки выбросов тепловых установок и могут быть использованы при разработке и модернизации экологических технологий для энергетического и промышленного секторов.

References

Bui M., Adjiman C.S., Bardow A., et al. (2018) Carbon capture and storage (CCS): The way forward. Energy and Environmental Science, 11, 1062–1176. DOI: https://doi.org/10.1039/C7EE02342A (in Eng.).

Fogler H.S. (2016) Elements of Chemical Reaction Engineering. 5th ed. Pearson. ISBN: 978-0-13-388751-8 (in Eng.).

Hasan H.F., Al-Sudani, F.T., Albayati T.M., Salih I.K., Hharah H.N., Majdi H.S., Saady N.M.C., Zendehboudi S., Amari A., and Gheni S.A. (2024) Solid adsorbent material: A review on trends of post-combustion CO_2 capture. Process Safety and Environmental Protection, 182, 975–988. https://doi.org/10.1016/j.psep.2023.12.025 (in Eng.).

Grieves M., Vickers J. (2017) Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. Transdisciplinary Perspectives on Complex Systems. Springer, DOI: https://doi.org/10.1007/978-3-319-38756-7_4 (in Eng.).

Javaid M., Haleem A., and Suman R. (2023) Digital Twin applications toward Industry 4.0: A Review. Cognitive Robotics, 3, 71–92. https://doi.org/10.1016/j.cogr.2023.04.003 (in Eng.).

Khussain B., Brodskiy A., Sass A., et al. (2022) Research of the thermal effect on the Fe–Cr–Al alloy foil. Coatings, 12(9), Article 1266. https://doi.org/10.3390/coatings12091266 (in Eng.).

Khussain B., Sass A., Brodskiy A., et al. (2025) Dry carbonate sorbents for CO_2 capture from flue gases. Molecules, 30(13), Article 2859. https://doi.org/10.3390/molecules30132859 (in Eng.).

Khussain B.Kh., Zhurinov M.Zh., Sass A.S., et al. (2020) Novye nanostrukturnye katalizatory dlya neytralizatsii toksichnykh gazov v promyshlennosti [New nanostructured catalysts for neutralization of toxic gases in industry]. Izvestiya NAN RK. Seriya Khimii i Tekhnologii, 1(439), 116–122. https://doi.org/10.32014/2020.2518-1491.15 (in Eng.).

Kritzinger W., Karner M., Traar G., Henjes J., and Sihn W. (2018) Digital twin in manufacturing: A categorical literature review. IFAC-PapersOnLine, 51(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474 (in Eng.).

Khussain B. Kh., Brodskiy A. R., Sass A.S., and Telmanov M. M. (2024). Sozdanie tsifrovykh dvoynikov, vklyuchaya modul' dekarbonizatsii, pri modelirovanii sistem ochistki dymovykh gazov [Creation of digital twins, including a decarbonization module, in modeling flue gas purification systems]. Reports of the NAS RK. https://doi.org/10.32014/2023.2518-1483.318 (in Russian)

Levenspiel O. (1999) Chemical reaction engineering (3rd ed.). Wiley. (in Eng.).

Perry R.H., and Green D.W. (2019) Perry's chemical engineers' handbook (9th ed.). McGraw-Hill. (in Eng.).

Perno M., Hvam L., and Haug A. (2021) Implementation of digital twins in the process industry: A systematic literature review of enablers and barriers. Computers in Industry, 134, 103558. https://doi.org/10.1016/j.compind.2021.103558 (in Eng.).

Raganati F., Miccio F., and Ammendola P. (2021) Adsorption of Carbon Dioxide for Post-combustion Capture: A Review. Energy and Fuels, 35(16), 12845–12868. https://doi.org/10.1021/acs.energyfuels.1c01618

Rochelle G.T. (2009) Amine scrubbing for CO_2 capture. Science, 325(5948), 1652–1654. https://doi.org/10.1126/science.1176731 (in Eng.).

Tao F., Zhang H., Liu A., and Nee A.Y.C. (2019) Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), 2405–2415. https://doi.org/10.1109/TII.2018.2873186 (in Eng.).

Sharma A., Pareek V., Zhang D., et al. (2024) CO2 Post-combustion Capture: A Critical Review of Current Technologies and Future Directions. Energy and Fuels. https://doi.org/10.1021/acs.energyfuels.4c02513 (in Eng.).

Wang M., Lawal A., Stephenson P., Sidders J., and Ramshaw C. (2011) Post-combustion CO_2 capture with chemical absorption. Chemical Engineering Research and Design, 89(9), 1609–1624. https://doi.org/10.1016/j.cherd.2010.11.005(in Eng.).

ISO (2021) ISO 23247-1:2021 Automation systems and integration - Digital twin framework for manufacturing - Part 1: Overview and general principles. International Organization for Standardization. (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 236–252

https://doi.org/10.32014/2026.2518-1726.411

IRSTI 89.25
UDC 629.78

© Kulakayeva A.[1]*, Ashurov A.[2], Zhumazhanov B.[2,3], Daineko Ye.[1,4], Zylgara A.[5], 2026.

[1] JSC «International Information Technology University», Almaty, Kazakhstan;
[2] NJSC «L.N. Gumilyov Eurasian National University», Astana, Kazakhstan; [3] LLP «Ghalam», Astana, Kazakhstan;
[4]NPJSC «Satbayev University», Almaty, Kazakhstan;
[5]JSC «Kazakhstan Gharysh Sapary», Astana, Kazakhstan.
E-mail: a.kulakayeva@iitu.edu.kz

ALGORITHM FOR DETERMINING THE INITIAL ORBITAL PARAMETERS OF KAZEOSAT-1 FOR DEORBITING

Kulakayeva Aigul — PhD, Associate Professor of International Information Technology University, Almaty, Kazakhstan,
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Ashurov Abdikul — Candidate of Physical and Mathematical Sciences, Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan,
E-mail: ashurov_ae@enu.kz, http://orcid.org/0000-0001-6044-5579;
Zhumazhanov Berik — Head of the Payload and R&D Department, LLP "Ghalam"; PhD student of Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan,
E-mail: b.zhumazhanov@ghalam.kz, http://orcid.org/0000-0001-5926-9619;
Daineko Yevgeniya — PhD, Director of the Institute of Automation and Information Technology, Satbayev University; Associate Professor of International Information Technology University, Almaty, Kazakhstan,
E-mail: y.daineko@iitu.edu.kz, https://orcid.org/0000-0001-6581-2622;
Zylgara Aigerim — Engineer, JSC "Kazakhstan Gharysh Sapary", Astana, Kazakhstan,
E-mail: zylgara@gmail.com, https://orcid.org/0009-0008-9916-8685.

Abstract. The growing number of objects in low-Earth orbits increases the risk of collisions and the formation of space debris, which requires the development of controlled and technically feasible methods for deorbiting existing spacecraft. For the Republic of Kazakhstan, the creation of its own KazEOSat-1 mission completion algorithm is a prerequisite for compliance with international requirements (ISO 24113:2023E) and ensuring the safety of the national orbital infrastructure, especially given that the KazEOSat series satellites have already exceeded their estimated service life. The aim of the work is to develop an effective algorithm for controlled deorbitation of the KazEOSat–1 spacecraft, taking into account its

actual technical characteristics, orbit parameters and remaining fuel. The research is based on numerical modeling methods in Matlab using current ephemeris (TLE data). The mathematical apparatus includes the conversion of Keplerian elements into Cartesian coordinates and the calculation of multi-stage braking pulses of a low-thrust propulsion system. The simulation takes into account both active maneuvers and the effect of the aerodynamic drag of the upper atmosphere. A deorbitation algorithm has been developed and verified to ensure a controlled descent of the spacecraft from an initial altitude of 758 km. It has been established that the optimal scheme is a stepwise descent using two-pulse maneuvers lasting 600 seconds at diametrically opposite points of the orbit. The simulation confirmed that the available fuel reserve (67 kg) is sufficient to transfer the vehicle to the passive braking phase followed by combustion in dense layers of the atmosphere. The results allow us to plan the completion of satellite operations in low orbits in accordance with international standards.

Keywords: deorbitation, space debris, KazEOSat-1, satellite, spacecraft, orbit, orbital parameters

Financing. *This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan Grant No. AP26103099, titled "Development of deorbiting methods for Kazakhstan satellites using virtual reality technology"*

For citations: Kulakayeva A., Ashurov A., Zhumazhanov B., Daineko Ye., Zylgara A. Algorithm for determining the initial orbital parameters of kazeosat-1 for deorbiting. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 236–252. DOI: https://doi.org/10.32014/2026.2518-1726.411

© Кулакаева А.Е.[1]*, Ашуров А.Е.[2], Жумажанов Б.Р.[2,3], Дайнеко Е.А.[1,4], Зылғара А.Е.[5], 2026.

[1] «Халықаралық ақпараттық технологиялар университеті» АҚ, Алматы, Қазақстан;

[2] «Л.Н. Гумилев атындағы Еуразия ұлттық университеті» КеАҚ, Астана, Қазақстан;

[3]«Ghalam» ЖШС, Астана, Қазақстан;

[4]«Satbayev University» ҚеАҚ, Алматы, Қазақстан;

[5]«Қазақстан Ғарыш Сапары» АҚ, Астана, Қазақстан.

E-mail: a.kulakayeva@iitu.edu.kz

KAZEOSAT-1 ҒАРЫШ АППАРАТЫНЫҢ ДЕОРБИТАЦИЯСЫН ЖҮЗЕГЕ АСЫРУ ҮШІН БАСТАПҚЫ ОРБИТАЛЫҚ ПАРАМЕТРЛЕРІН АНЫҚТАУ АЛГОРИТМІ

Кулакаева Айгуль — PhD, Халықаралық ақпараттық технологиялар университетінің қауымдастырылған профессоры, Алматы, Казахстан,

E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Ашуров Абдикул — Физика-математика ғылымдарының кандидаты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Казахстан,
E-mail: ashurov_ae@enu.kz, http://orcid.org/0000-0001-6044-5579;
Жумажанов Берик — «Ghalam» ЖШС, Пайдалы жүктеме және ғылыми әзірлемелер бөлімінің жетекшісі; Л.Н. Гумилев атындағы Еуразия ұлттық университетінің PhD докторанты, Астана, Казахстан,
E-mail: b.zhumazhanov@ghalam.kz, http://orcid.org/0000-0001-5926-9619;
Дайнеко Евгения — PhD, Автоматтандыру және ақпараттық технологиялар институтының директоры, Satbayev University; Халықаралық ақпараттық технологиялар университетінің қауымдастырылған профессоры, Алматы, Казахстан,
E-mail: y.daineko@iitu.edu.kz, https://orcid.org/0000-0001-6581-2622;
Зылғара Айгерим — «Қазақстан Ғарыш Сапары» АҚ инженері, Астана, Казахстан,
E-mail: zylgara@gmail.com, https://orcid.org/0009-0008-9916-8685.

Аннотация. Төмен Жер орбитасындағы объектілер санының артуы соқтығысу қаупін және ғарыштық қоқыстың түзілуін күшейтеді, бұл жұмыс атқарып тұрған ғарыш аппараттарын басқарылатын әрі техникалық тұрғыдан іске асырылатын деорбитациялау әдістерін әзірлеуді талап етеді. Қазақстан Республикасы үшін KazEOSat-1 жерсерігінің миссиясын аяқтаудың жеке алгоритмін әзірлеу халықаралық талаптарды (ISO 24113:2023E) сақтау және ұлттық орбиталық инфрақұрылымның қауіпсіздігін қамтамасыз ету үшін қажетті шарт болып табылады. Бұл, әсіресе, KazEOSat сериясындағы жерсеріктердің есептік пайдалану мерзімінен асып кеткенін ескергенде маңызды. Жұмыстың мақсаты – нақты техникалық сипаттамаларын, орбита параметрлерін және қалған отын қорын ескере отырып, KazEOSat-1 ғарыш аппаратын басқарылатын деорбитациялау үшін тиімді алгоритм әзірлеу. Зерттеу Matlab ортасында өзекті эфемеридтерді (TLE деректері) пайдалана отырып сандық модельдеу әдістеріне негізделген. Математикалық аппарат Кеплер элементтерін декарт координаттарына түрлендіруді және аз тартылысты қозғалтқыш қондырғысының көпсатылы тежеу импульстерін есептеуді қамтиды. Модельдеу барысында белсенді маневрлермен қатар атмосфераның жоғарғы қабаттарындағы аэродинамикалық кедергінің әсері де ескерілді. Бастапқы биіктігі 758 км болатын ғарыш аппаратының басқарылатын төмендеуін қамтамасыз ететін деорбитация алгоритмі әзірленіп, верификациядан өтті. Орбитаның диаметрлік қарама-қарсы нүктелерінде ұзақтығы 600 секунд болатын екіимпульсті маневрлерді қолдана отырып, сатылы төмендету сұлбасы оңтайлы екені анықталды. Модельдеу нәтижелері қолда бар отын қорының (67 кг) аппаратты пассивті тежеу фазасына көшіруге және атмосфераның тығыз қабаттарында жанып кетуін қамтамасыз етуге жеткілікті екенін растады. Алынған нәтижелер төмен орбитадағы спутниктердің пайдалануын халықаралық стандарттарға сәйкес аяқтауды жоспарлауға мүмкіндік береді.

Түйін сөздер: деорбитация, ғарыштық қоқыс, KazEOSat-1, жерсерік, ғарыш аппараты, орбита, орбиталық параметрлер

© **Кулакаева А.Е.[1]*, Ашуров А.Е.[2], Жумажанов Б.Р.[2,3], Дайнеко Е.А.[1,4], Зылғара А.Е.[5] , 2026.**

[1] АО «Международный университет информационных технологий», Алматы, Казахстан;

[2]НАО «Евразийский национальный университет им. Л.Н. Гумилева», Астана, Казахстан;

[3]ТОО «Ghalam», Астана, Казахстан;

[4] НАО «Satbayev University», Алматы, Казахстан;

[5]АО «Қазақстан Ғарыш Сапары», Астана, Казахстан.

E-mail: a.kulakayeva@iitu.edu.kz

АЛГОРИТМ ОПРЕДЕЛЕНИЯ НАЧАЛЬНЫХ ОРБИТАЛЬНЫХ ПАРАМЕТРОВ KAZEOSAT-1 ДЛЯ ДЕОРБИТАЦИИ

Кулакаева Айгуль — PhD, Ассоциированный профессор Международного университета информационных технологий, Алматы, Казахстан,
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Ашуров Абдикул — кандидат физико-математических наук, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: ashurov_ae@enu.kz, http://orcid.org/0000-0001-6044-5579;
Жумажанов Берик — руководитель отдела полезной нагрузки и научных разработок, ТОО «Ghalam», PhD докторант, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: b.zhumazhanov@ghalam.kz, http://orcid.org/0000-0001-5926-9619;
Дайнеко Евгения — PhD, Директор Института автоматики и информационных технологий, Satbayev University, Ассоциированный профессор Международного университета информационных технологий, Алматы, Казахстан,
E-mail: y.daineko@iitu.edu.kz,: https://orcid.org/0000-0001-6581-2622;
Зылғара Айгерим — инженер АО «Қазақстан Ғарыш Сапары», Астана, Казахстан,
E-mail: zylgara@gmail.com, https://orcid.org/0009-0008-9916-8685.

Аннотация. Рост числа объектов на низких околоземных орбитах увеличивает риск столкновений и образования космического мусора, что обусловливает необходимость разработки управляемых и технически реализуемых методов деорбитации космических аппаратов. Для Республики Казахстан разработка собственного алгоритма завершения миссии спутника KazEOSat-1 является важным условием соблюдения международных стандартов (ISO 24113:2023) и обеспечения безопасности национальной орбитальной инфраструктуры, особенно с учётом превышения расчётного срока эксплуатации спутников данной серии. Целью работы является разработка эффективного алгоритма управляемой деорбитации космического аппарата KazEOSat-1 с учётом его технических характеристик, параметров орбиты и остатка топлива. Исследование основано на методах численного моделирования в среде MATLAB с использованием актуальных эфемерид (TLE-данных). Математический аппарат включает преобразование кеплеровых элементов орбиты в декартовы координаты и расчёт многоступенчатых

тормозных импульсов двигательной установки малой тяги. Моделирование учитывает как активные манёвры, так и влияние аэродинамического сопротивления верхних слоёв атмосферы. В результате разработан и верифицирован алгоритм деорбитации, обеспечивающий контролируемый спуск космического аппарата с начальной высоты 758 км. Установлено, что оптимальной является схема ступенчатого снижения с использованием двухимпульсных манёвров длительностью 600 секунд, выполняемых в диаметрально противоположных точках орбиты. Результаты моделирования показывают, что имеющегося запаса топлива (67 кг) достаточно для перевода аппарата в режим пассивного торможения с последующим сгоранием в плотных слоях атмосферы. Полученные результаты могут быть использованы для планирования завершения миссий космических аппаратов на низких орбитах в соответствии с международными требованиями.

Ключевые слова: деорбитация, космический мусор, KazEOSat-1, спутник, космический аппарат, орбита, орбитальные параметры

Введение. В настоящее время Республика Казахстан (РК) активно развивает свои космические технологии и к 2030 году планирует значительно расширить свое присутствие на мировом рынке космических услуг (Стратегия «Казахстан-2050»: новый политический курс состоявшегося государства. Информационно-правовая система нормативных правовых актов Республики Казахстан «Әділет». URL: https://adilet.zan.kz/rus/docs/K1200002050). Первый казахстанский спутник дистанционного зондирования Земли (ДЗЗ) высокого разрешения KazEOSat-1 был запущен 30 апреля 2014 года, а 20 июня 2014 года, за ним последовал спутник среднего разрешения KazEOSat-2. Несмотря на то, что оба спутника уже превысили свой расчетный срок службы, они продолжают работать до сих пор. Однако в ближайшем будущем возникнет необходимость их утилизации в соответствии с международными стандартами по сокращению космического мусора (Space systems – Space debris mitigation requirements (ISO 24113:2023E)). Это связано не только с выполнением глобальных обязательств в сфере космической деятельности, но и с необходимостью обеспечения безопасности работы других космических аппаратов (КА) и минимизации образования новых фрагментов на орбите.

Для РК разработка собственного метода деорбитации КА является стратегически важной задачей. Это позволит не только соответствовать международным требованиям, но и обеспечит устойчивое развитие национальной космической программы, а также минимизирует риски для отечественной и мировой орбитальной инфраструктуры.

Литературный обзор. Рост числа искусственных объектов на орбите Земли увеличивает вероятность столкновений и образования нового космического мусора. Это представляет угрозу безопасности и устойчивости будущих космических миссий. Поэтому вопросам эффективной деорбитации завершивших свою работу КА уделяется особое внимание. В последние годы

акцент был сделан на аэродинамические методы деорбитации, в частности, на использование устройств, увеличивающих аэродинамическое сопротивление, таких как мембранные паруса. Такие устройства способствуют ускоренному снижению высоты орбиты за счет взаимодействия с атмосферой. Однако остаются нерешенные проблемы, связанные с обеспечением стабильности ориентации таких систем под воздействием солнечной радиации, гравитационных возмущений и других факторов космической среды (Zhang et al., 2024; Fang et al., 2025).

Современные исследования включают методы Монте-Карло и численные термодинамические модели. Тем не менее, остаются нерешенные вопросы относительно выбора устойчивых материалов и точности моделирования взаимодействия замедлителей с атмосферой на околоземных высотах (Olave et al., 2023; Caqueo et al., 2024). Инновационным решением являются надувные замедлители, которые выделяются своей простотой, малым весом и низкими энергетическими требованиями (Xu et al., 2024; Jiao et al., 2023). Однако такие устройства подвержены риску повреждений от микрометеоритов и космического мусора, а также могут сталкиваться с проблемами надежности раскрытия при низкой плотности атмосферы.

Использование квадратных парусов с контролируемым вращением повышает предсказуемость процесса деорбитации, особенно для КА с высокой площадью поверхности относительно их массы. Однако такие решения требуют сложных и дорогостоящих систем управления ориентацией (Niccolai and Mengali, 2024; Chen et al., 2024). Пирамидальные конфигурации парусов показывают улучшенную устойчивость, но влияние внешних космических факторов остается недостаточно изученным из-за ограниченного наземного моделирования и небольшого количества реальных орбитальных экспериментов (Zhang et al., 2022; Black and Spencer, 2020)

Электродинамические тросы (ЭДТ) представляют собой безтопливный метод снижения высоты орбиты с использованием силы Лоренца. Они значительно ускоряют процесс деорбитации КА, однако создают новые проблемы, связанные с устойчивостью системы в условиях ионосферы, управляемостью троса и его уязвимостью к фрагментам космического мусора. Несмотря на эффективность данного метода, остаются нерешенными вопросы оптимизации параметров троса, выбора материалов и подтверждения его надежности на реальных орбитах (Xiao et al., 2024; Sarego et al., 2021)

В настоящее время в РК нет собственного метода деорбитации КА, который учитывал бы реальные технические характеристики и эксплуатационное состояние наших КА ДЗЗ.

Основная задача данного исследования заключается в разработке эффективного алгоритма деорбитации для КА ДЗЗ KazEOSat-1 с учетом его остаточного запаса топлива, параметров орбиты и требований по снижению рисков для орбитальной инфраструктуры.

Целью данной работы является разработка эффективного алгоритма

деорбитации КА ДЗЗ KazEOSat-1 с учетом его технических характеристик и остаточного ресурса. Это позволит минимизировать риски для орбитальной инфраструктуры и обеспечить соблюдение международных стандартов.

Материалы и основные методы. Исходные орбитальные параметры КА ДЗЗ KazEOSat-1 (элементы TLE) были получены из открытого ресурса CelesTrak. Также, использовалась информация из открытых источников о характеристиках КА, включая массу, остаток, параметры двигателя, орбитальные данные.

Объектом исследования стал казахстанский КА ДЗЗ KazEOSat-1. Для проведения моделирования и расчета деорбитации были использованы следующие исходные параметры КА (таблица 1):

Таблица 1 — Характеристики КА ДЗЗ KazEOSat-1.

Характеристика	Описание
Сухая масса КА	747 кг
Масса КА с топливом, изначальная	830 кг
Масса топлива, изначальная	83 кг
Масса КА с топливом на момент расчетов	814 кг
Масса топлива КА на момент расчета	67 кг
Срок активного существования	7 лет
Платформа	Leostar-500-XO
Размеры КА	2.10×3.70 м
Мощность	1200 Вт
Двигатель на борту	1N, однокомпонентный, гидразин.

КА ДЗЗ KazEOSat-1 находится на солнечно-синхронной низкой околоземной орбите. Основные орбитальные характеристики представлены в таблице 2:

Таблица 2 — Характеристики орбиты КА ДЗЗ KazEOSat-1

Характеристика	Описание
Позиция КА	738×742 км
Большая полуось орбиты	750 км
Наклонение орбиты	98.5 град
Классификация орбиты	Низкая околоземная
Тип орбиты	солнечно - синхронный

Для моделирования работы двигательной установки малой тяги (ДМТ) использовались технические характеристики двигателя 1N (таблица 3):

Таблица 3 — Характеристики двигателя 1N

Характеристика	Величина
Номинальная тяга	1 Н
Диапазон используемой тяги	0.320–1.1 Н
Номинальный удельный импульс	220 сек
Диапазон используемого удельного импульса	200–223 сек

Номинальный массовый расход топлива	0.44 г/с
Диапазон используемого массового расхода топлива	0.142–0.447 г/с
Минимальный порционный импульс	0.01–0.043 Н
Масса двигателя с сопловой частью	290 г
Топливо	Гидразин (N4H4)
Количество на борту КА	4 штуки
Угол между тягой и направлением движения	30 град

Все указанные параметры были использованы в качестве исходных данных для построения численной модели деорбитации. Моделирование орбитального движения и процессов деорбитации проводилось на основе законов небесной механики и динамики полета искусственных спутников Земли. Для расчета высоты орбиты, координат и скоростей КА использовались Кеплеровы элементы, которые преобразовывались в декартовы координаты.

Вся процедура моделирования и расчетов была реализована с использованием Matlab. Разработанные скрипты обеспечивали автоматизированный ввод исходных данных, расчет орбитальных параметров, определение последовательности маневров и визуализацию траекторий.

Результаты. На первом этапе расчетов был осуществлен сбор исходных орбитальных данных КА ДЗЗ KazEOSat-1, необходимых для дальнейшего моделирования траектории деорбитации. В качестве базовых данных использовались элементы TLE (two-line element set), которые регулярно обновляются в базе CelesTrak.

КА ДЗЗ KazEOSat-1 проходит над зоной радиовидимости наземных станций в среднем не менее 5–6 раз в сутки**.** Такая частота пролетов обеспечивает регулярные возможности для проведения сеансов связи и получения информации от КА, которая показана на рисунке 1.

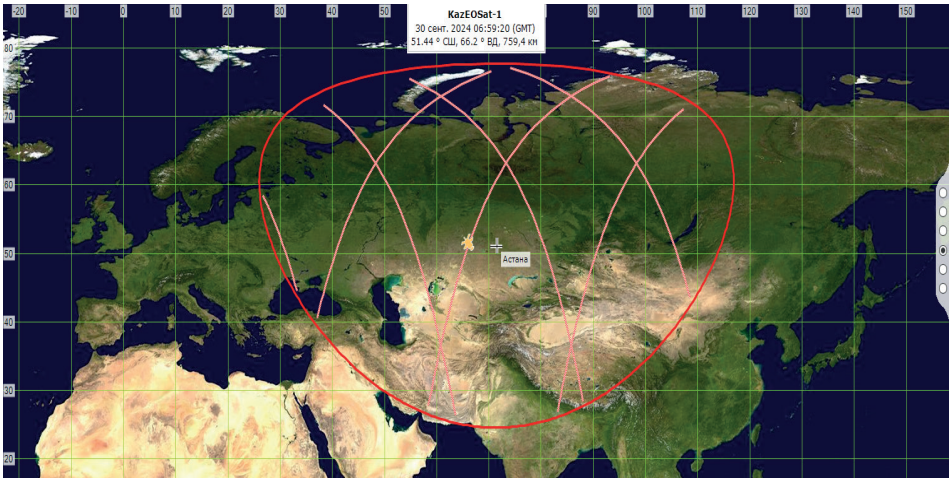


Рисунок 1 — Зона радиовидимости КА KazEOSat-1 над территорией РК

Исходные орбитальные параметры KazEOSat-1 были получены с сайта CelesTrak, где актуальные TLE для КА доступны по названию или номеру из базы данных NORAD. Для расчетов использовались элементы TLE, актуальные на 30 сентября 2024 года, 21:23:07 UTC.

KAZEOSAT 1:

1 39731U 14024A 24274.89106327 .00000983 00000+0 31265-3 0 9998.

2 39731 98.3873 349.5671 0001084 91.4249 268.7065 14.42027901548584.

На основе выбранных TLE для KazEOSat-1 была определена фактическая орбитальная позиция КА на момент начала расчетов по деорбитации. Для анализа использовалось Matlab. Рабочий процесс включал следующие этапы:

1. Поиск актуальных данных TLE.

2. Запуск Matlab и открытие скрипта для обработки орбитальных данных.

3. Загрузка файла с TLE в программу и автоматическое извлечение параметров орбиты.

4. Получение координат и скорости KazEOSat-1, необходимых для последующего моделирования.

В результате была сформирована исходная орбитальная база для проведения дальнейших численных экспериментов по моделированию деорбитации KazEOSat-1.

После ввода TLE-элементов KazEOSat-1 в скрипт Matlab были получены орбитальные параметры, соответствующие состоянию спутника на 30 сентября 2024 года, 21:23:07 UTC. В результате обработки данных были определены основные Кеплеровы элементы орбиты, которые легли в основу последующих расчетов. На основе этих элементов были вычислены пространственные координаты КА (x, y, z) и высота орбиты на выбранный момент времени.

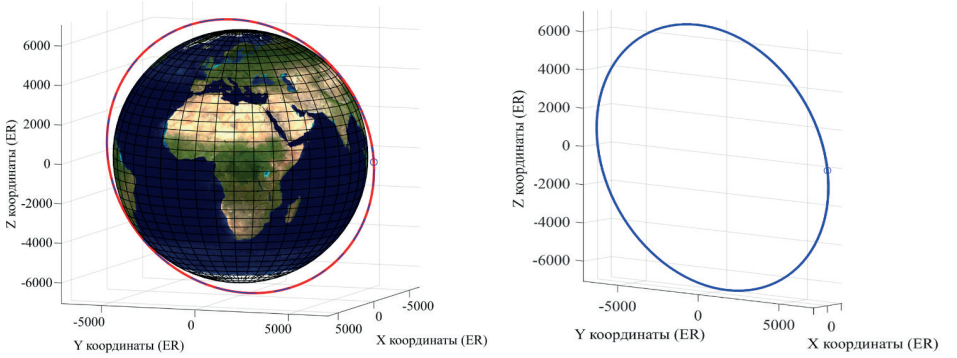
Вычисленные параметры орбиты использовались в первом скрипте для определения текущей высоты KazEOSat-1 в момент начала моделирования. Данная величина задает исходную точку для последующего расчета этапов активной фазы деорбитации КА. Для определения высоты орбиты применялась следующая формула (1):

$$h = \sqrt{(x)^2 + (y)^2 + (z)^2} - R \qquad (1)$$

где x, y, z – координаты КА относительно центра Земли; R – радиус Земли.

В результате вычислений с использованием первого скрипта и формулы (1) установлено, что начальная высота рабочей орбиты KazEOSat-1 на момент моделирования составляет 758 км. Данный параметр был принят за отправную точку для дальнейшего расчета траектории деорбитации КА.

На рисунке 2 (а) показана пространственная траектория движения КА за один полный виток, с отмеченным фактическим положением KazEOSat-1 в момент начала расчетов. Для лучшего понимания геометрии орбиты, без учета поверхности Земли визуализация приведена отдельно на рисунке 2 (б).



a *b*

Рисунок 2 — Рабочая орбита КА и местоположение к началу расчета: *а* – с изображением Земли; *б* – без изображения Земли

На рисунке 2 (б) показана траектория KazEOSat-1 за один орбитальный виток, основанная на результатах моделирования. Синяя линия обозначает путь КА вокруг Земли, а маленький синий круг указывает на его начальное положение КА на 30 сентября 2024 года, 21:23:07 UTC, которое выбрано в качестве отправной точки для всех последующих расчетов. В процессе моделирования были учтены основные орбитальные параметры и соблюдены масштабные соотношения, что обеспечивает корректность визуализации. Расчетная траектория показывает движение спутника за один оборот, начиная с указанного положения. Важно отметить, что незначительные внешние воздействия, такие как гравитационные возмущения или атмосферное сопротивление на рассматриваемой высоте, в данном анализе не учитывались из-за их минимального влияния на параметры орбиты за короткий временной промежуток.

Полученная модель орбиты KazEOSat-1 за один полный виток, основанная на актуальных исходных данных, позволяет с высокой точностью определить пространственное положение и высоту КА на начальном этапе расчетов. Данный результат является отправной точкой для следующего этапа моделирования траектории деорбитации, поскольку правильная фиксация начальных параметров критически важна для выбора и реализации оптимальной последовательности маневров по снижению орбиты КА.

Для повышения эффективности и упрощения расчетов на следующем этапе моделирования было принято допущение о постоянстве наклонения орбиты KazEOSat-1 (Δi=0). Это означает отсутствие изменения этого параметра во время маневра. Предполагается компланарный переход, при котором изменение высоты орбиты происходит без изменения наклонения. В этом случае орбиту КА удобно представлять в двумерной системе координат (x, y), что значительно облегчает расчет параметров переходных маневров. Такой подход применим в тех ситуациях, когда исходная и целевая орбиты

имеют одинаковое наклонение, что подтверждает его актуальность для рассматриваемой задачи.

Разработанный скрипт позволил рассчитать активную деорбитацию KazEOSat-1 с использованием ДМТ, установленной на борту КА. Для моделирования были использованы параметры КА, представленные в таблице 1. KazEOSat-1 работает на солнечно-синхронной низкой околоземной орбите, что является характерным для КА ДЗЗ. Основные параметры его орбиты указаны в таблице 2.

Для расчета параметров деорбитационного маневра использовалась информация из таблицы 3. На основе этих данных были определены необходимые значения импульсов и расхода топлива на различных этапах моделирования.

Для расчета последовательных этапов активной деорбитации использовались формулы (2)–(9). Поскольку KazEOSat-1 оснащен ДМТ, одномоментный перевод КА с рабочей орбиты на орбиту, где начнется пассивная фаза деорбитации невозможен. Снижение высоты происходило ступенчато, с последовательным уменьшением. В связи с этим расчетные выражения были адаптированы для многократного применения малых импульсов, что отражено в следующих формулах:

$$\Delta V_{1n} = V_{an} - V_{kp1n}, \quad (2)$$

где ΔV_{1n} – величина первого импульса скорости для торможения, определяемая разностью скоростей в апогее переходной траектории и на исходной околокруговой орбите порядка *n*, м/с.

Величина второго импульса, необходимого для перевода КА с переходной траектории на финальную орбиту, определяется следующим выражением:

$$\Delta V_{2n} = V_{kp2n} - V_{pn}, \quad (3)$$

где ΔV_{2n} – второй импульс торможения, который рассчитывается как разность скоростей в перигее переходной траектории и на конечной околокруговой орбите порядка *n*, м/с.

Суммарный импульс, необходимый для выполнения всей последовательности деорбитационных маневров, определялся как сумма всех применяемых скоростных импульсов:

$$\sum_0^n \Delta V_n = \Delta V_{1n} + \Delta V_{2n}, \quad (4)$$

Масса топлива, необходимая для перевода КА на орбиту утилизации, определялся на основе характеристик ДМТ и требуемого значения скоростного импульса. Для расчета массы топлива, затрачиваемого на выполнение первого импульса, использовалась следующая формула:

$$m_T = m_{KA} * \left[1 - e^{\frac{\Delta V}{I_{уд}}}\right], \quad (5)$$

где m_{KA} – масса КА, e – число Эйлера равное 2,7182818284, $I_{уд}$ – удельный импульс КА в м/с, ΔV – импульс скорости маневра.

В расчетах использовалось фиксированное время работы ДМТ, поэтому масса израсходованного топлива определялся с учетом данного параметра. Соответствующее выражение является модифицированным вариантом формулы (5):

$$m_T = m * \Delta t\ , \quad (6)$$

где m_T – масса топлива, кг, m – масса расхода топлива за 1 секунду работы ДМТ, Δt – время работы двигателя.

В этом случае величина скоростного импульса определяется следующим образом:

$$\Delta V = \frac{F}{m_\pi} * \Delta t,$$

где ΔV – импульс изменения скорости, м/с; F – тяга, производимая двигателем при работе, Н; – масса КА с учетом топлива, кг; Δt – рабочее время двигателя, с.

При этом необходимо учитывать, что на борту КА ДЗЗ KazEOSat-1 установлены 4 сопла ДМТ, расположенные по бокам корпуса. Угол α между вектором тяги и направлением движения КА не равен 0. Поэтому результирующая тяга определяется следующим образом:

$$F = 4 * F_H * \cos\alpha, \quad (7)$$

где $F_Н$ – номинальная тяга одного двигателя, α – угол между направлением тяги и движением КА.

При моделировании также учитывается влияние атмосферы Земли как одного из факторов, воздействующих на траекторию деорбитации КА. В связи с этим в расчеты включена сила аэродинамического сопротивления, которая оказывает тормозящее действие на КА:

$$F_A = C_{xa} S_M \frac{\rho V^2}{2}, \quad (8)$$

где $C_{ха}$ – коэффициент аэродинамического сопротивления корпуса КА; $S_М$ – эффективная площадь лобовой поверхности при снижении; ρ – плотность

атмосферы на соответствующей высоте; g – ускорение свободного падения равное 9,81 м/с²; V – текущая скорость движения КА.

Обсуждение. Таким образом, для определения величины тормозного импульса используется следующее выражение:

$$\Delta V = \frac{F+F_A}{m_\pi} * \Delta t, \qquad (9)$$

где ΔV – изменение скорости КА, м/с; F – тяга, создаваемая ДМТ, F_A – сила торможения КА об атмосферу, Н; – масса КА с учетом топлива, кг; Δt – рабочее время двигателя, с.

В рассматриваемом процессе снижение орбиты KazEOSat-1 достигается как за счет тормозных импульсов, создаваемых ДМТ, так и под воздействием аэродинамического сопротивления верхних слоев атмосферы Земли.

Процесс деорбитации KazEOSat-1 наглядно представлен на рисунке 3. Красная линия на графике отображает исходную орбиту спутника, соответствующую траектории, показанной ранее на рисунке 2 (б), но в двумерной проекции – для удобства анализа компланарных переходов. Активные этапы снижения высоты, осуществляемые с помощью включения ДМТ и расхода топлива, обозначены черной линией. Длина данного участка остается постоянной, так как время подачи импульса фиксировано. Пассивные фазы снижения, когда деорбитация КА происходит исключительно под действием аэродинамического сопротивления, выделены зеленой линией. Для визуального ориентирования на графике также отмечен радиус Земли, обозначенный синей линией.

Процесс деорбитации КА KazEOSat-1 с использованием импульсных тормозных маневров, выполняемых ДМТ, представлен на рисунке 3.

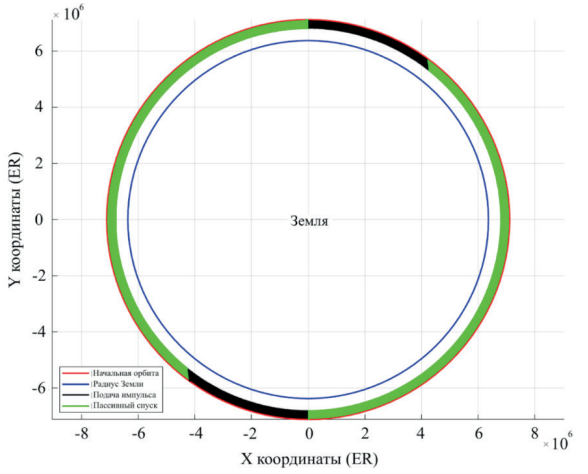


Рисунок 3 — График деорбитации КА при импульсивном торможении ДМТ

Существенное снижение высоты орбиты КА KazEOSat-1 достигнуто благодаря последовательному выполнению тормозных маневров ДМТ. Для плавного перехода на более низкую круговую орбиту были осуществлены два импульса торможения, каждый из которых длился 600 секунд. Эти импульсы использовались в диаметрально противоположных точках орбиты (с углом разделения 180°), что связано как с необходимостью восстановления работоспособности двигателя между включениями, так и с оптимизацией режима работы ДМТ в условиях заданной динамики КА.

Оба импульсных маневра выполняются за один орбитальный виток, что эффективно снижает высоту орбиты КА KazEOSat-1 в этот период. Для наглядности основные результаты моделирования представлены в компактной форме. Динамика изменения высоты орбиты отображена с интервалом в 50 км. Такой шаг позволяет акцентировать внимание на важных этапах процесса деорбитации и упрощает восприятие общей картины изменений орбитальных параметров.

В рамках работы был реализован алгоритм расчета и моделирования начальных орбитальных параметров КА KazEOSat-1 для последующей деорбитации. Основная особенность разработанного подхода заключается в адаптации традиционных методов коррекции орбит к специфике задач по снижению КА с орбиты, с учетом актуальных технических параметров, характеристик орбиты, внешних условий и доступных запасов топлива.

На рисунке 4 показана схема алгоритма, показывающая основные этапы расчета и моделирования исходных орбитальных параметров КА KazEOSat-1.

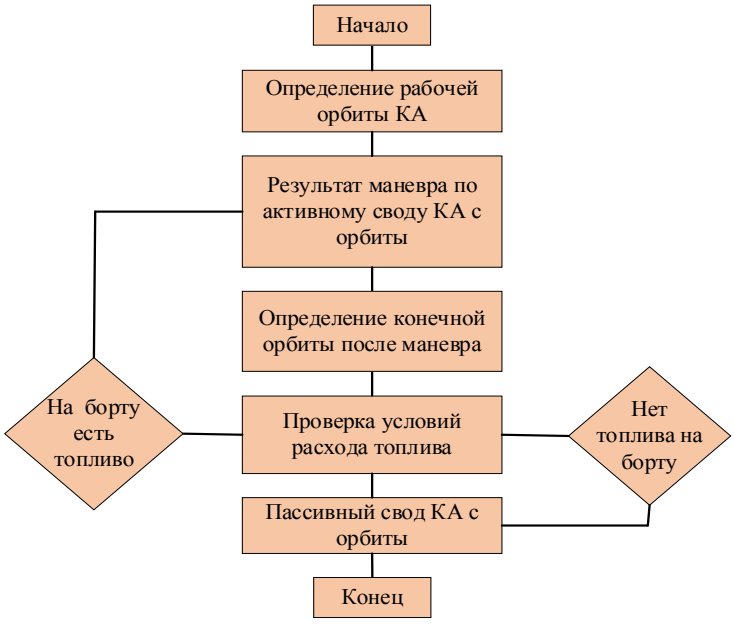


Рисунок 4 — Алгоритм расчета и моделирования начальных орбитальных параметров

Алгоритм деорбитации КА KazEOSat-1 включает несколько последовательных этапов, начиная с построения орбитальной модели на основе данных TLE. После преобразования TLE в эфемериды (координаты x, y, z и компоненты скорости *vx*, *vy*, *vz*) рассчитывается траектория КА на один виток.

Далее начинается активная фаза: определяется рабочая орбита, рассчитываются моменты и параметры подачи импульсов ДМТ. Первый импульс изменяет форму орбиты, а второй, приложенный в диаметрально противоположной точке, корректирует ее для поддержания околокруговой траектории. Этот цикл повторяется на каждом витке до исчерпания запасов топлива, что обеспечивает ступенчатое снижение высоты орбиты. На данном этапе влияние других космических объектов не учитывается, так как расчеты сосредоточены на динамике КА KazEOSat-1.

После завершения активной фазы снижения начинается пассивная фаза деорбитации, в ходе которой КА постепенно теряет высоту исключительно за счет аэродинамического сопротивления. Когда высота орбиты достигает критического значения, КА входит в плотные слои атмосферы, где происходит его торможение, разрушение и сгорание. Алгоритм охватывает полный жизненный цикл деорбитации – от расчета исходных параметров до безопасного завершения миссии и деорбитации КА с орбиты.

Данное моделирование подтвердило эффективность предложенного алгоритма для управляемой деорбитации КА KazEOSat-1. Ступенчатое снижение орбиты с использованием ДМТ и актуальных TLE обеспечивает точность расчетов и контроль параметров траектории. Основное ограничение метода связано с упрощенными допущениями и недостаточным учетом сложных внешних возмущений. Тем не менее, такой подход оправдан для практических задач с аналогичными условиями. Полученные результаты могут быть использованы для планирования деорбитации.

Заключение. В результате проведенного исследования был разработан алгоритм расчета и моделирования процесса деорбитации КА ДЗЗ KazEOSat-1, который учитывает актуальные орбитальные параметры, особенности конструкции и ограничения по остаточному запасу топлива. На основе обновленных TLE и технических характеристик КА реализована поэтапная схема деорбитации. Она включает последовательные импульсные маневры ДМТ, а также переход к пассивной фазе, когда на КА начинает действовать аэродинамическое сопротивление. Проведенный анализ показал, что предложенный подход обеспечивает эффективное и контролируемое снижение высоты орбиты, снижая вероятность образования нового космического мусора и минимизируя риски для других космических объектов. Кроме того, он соответствует международным стандартам по утилизации КА.

Уникальность предложенного алгоритма заключается в адаптации методов коррекции орбиты к задачам деорбитации КА с учетом реальных технических ограничений и внешних факторов. Полученные результаты могут быть

использованы для планирования завершения эксплуатации аналогичных КА на низких околоземных орбитах. В будущем целесообразно расширить алгоритм, интегрировав модели многоспутниковых сценариев и анализируя влияние долгосрочных возмущающих факторов на процесс деорбитации.

Важно подчеркнуть, что разработка и внедрение эффективных алгоритмов деорбитации КА имеют стратегическое значение для РК. Это связано с возрастающей ролью национальных космических программ и необходимостью соблюдения международных обязательств по снижению уровня космического мусора. Реализация предложенного подхода повысит технологическую независимость страны в управлении завершением жизненного цикла отечественных спутников и минимизирует потенциальные риски для национальной и мировой орбитальной инфраструктуры.

References

Black A., & Spencer D.A. (2020) DragSail systems for satellite deorbit and targeted reentry. Journal of Space Safety Engineering, 7(3). — P. 397-403 (in Eng.).

Caqueo N., et al. (2024) Inflatable aerodynamic decelerators for CubeSat reentry and recovery: Surface properties. Aerospace Science and Technology, 149. — P. 109151 (in Eng.).

Chen J., Chen S., Qin Y., Zhu Z., & Zhang J. (2024) Aerodynamic analysis of deorbit drag sail for CubeSat using DSMC method. Aerospace, 11(4). — P. 315 (in Eng.).

Fang G., Yun W., Liu L., Chen W., Cai J., Yan B., ... & Meng G. (2025) Advances in large-area membrane drag sail technology for standardized application on launch vehicle upper stages. Aerospace Science and Technology, 110617 (in Eng.).

International Organization for Standardization (2023) Space systems—Space debris mitigation requirements (ISO 24113:2023E) (4th ed.) (in Eng.).

Jiao N., Yang K., & Zhang J. (2023). Research on folding and inflation process of the drag balloon deorbit device. In AOPC 2023: Laser Technology and Applications; and Optoelectronic Devices and Integration, vol. 12959. — P. 194-204 (in Eng.).

Niccolai L., & Mengali G. (2024). Performance estimate of a spin-stabilized drag sail for spacecraft deorbiting. Applied Sciences, 14(2). — P. 612. DOI:10.3390/app14020612 (in Eng.).

Olave D.R., Jara N.C., Palharini R.C., Palharini R.S.A., Gaglio E., & Savino R. (2023) Inflatable aerodynamic decelerator for CubeSat reentry and recovery: Altitude effects on the flowfield structure. Aerospace Science and Technology, 138, 108358. (in Eng.).

Sarego G., Olivieri L., Valmorbida A., Brunello A., Lorenzini E. C., Tarabini Castellani L., ... & Sánchez-Arriaga G. (2021) Deployment requirements for deorbiting electrodynamic tether technology. CEAS Space Journal, 13, 567–581. DOI:10.1007/s12567-021-00349-5 (in Eng.).

Strategy "Kazakhstan-2050": a new political course of the established state, the Information and legal system of regulatory legal acts of the Republic of Kazakhstan "Adilet", as of 05/10/2025 [Strategiya «Kazakhstan-2050»: novyi politicheskii kurs sostoyavshegosya gosudarstva, Informatsionno-pravovaya sistema normativnykh pravovykh aktov Respubliki Kazakhstan «Adilet»]. [Online]. Access: https://adilet.zan.kz/rus/docs/K1200002050 (Available: 04.09.2025) (in Rus.)

Wei J., Zhang Y., Hou Y., & Tan H. (2023) Design and performance analysis of an inflatable film balloon for drag deorbiting. Journal of Tsinghua University (Science and Technology), 63(3). — P. 302-310 (in Eng.).

Xiao H., Huang J., Liu G., & Lv Y. (2020) Space environment modeling and deorbiting efficiency analysis for electro-dynamic tether system. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC). — P. 5063–5067. IEEE. DOI:10.1109/CCDC49329.2020.9164373 (in Eng.).

Xu Y., et al. (2024) Numerical simulation of rigid-flexible coupled dynamics for an inflatable sphere deorbiting device. Advances in Space Research, 74(1). — P. 373–383 (in Eng.).

Zhang J., Zhang R., & Yang K. (2022) Attitude stability analysis and configuration design of pyramid drag sail for deorbit missions. Journal of Aerospace Engineering, 35(6). — 04022084 p. (in Eng.).

Zhang R., Yang K., Zhang J., & Bi S. (2024) Overview and key technology of the membrane drag sail for low Earth orbit satellite deorbit. Space Science and Technology, 4, 0115. DOI:10.34133/space.0115 (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 253–270

https://doi.org/10.32014/2026.2518-1726.412

IRSTI 77.03.09
UDC 621.7.08

© Mimenbayeva A.B.[1], Turebayeva R. D.[2], Ospanova T.T.[2]*, Aruova A.B.[1], Naizagarayeva A.A.[3], 2026.
[1]Astana IT University, Astana, Kazakhstan;
[2]L.N. Gumilyov National Eurasian University, Astana, Kazakhstan;
[3]NCJSC «S. Seifullin Kazakh AgroTechnical Research University», Astana, Kazakhstan.
E-mail: Tleu2009@mail.ru

DEVELOPMENT AND COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR URBAN TRAFFIC PREDICTION

Mimenbayeva Aigul — Senior Lecturer, Astana IT University, Astana, Kazakhstan,
E-mail: A.Mimenbayeva@astanait.edu.kz, https://orcid.org/0000-0003-4652-470X;
Turebayeva Rakhila — Senior Lecturer, L.N. Gumilyov National Eurasian University, Astana, Kazakhstan,
E-mail: turebayeva_rd@enu.kz, https://orcid.org/0009-0006-4530-1300;
Ospanova Tleugaisha — Associate Professor, L.N. Gumilyov National Eurasian University, Astana, Kazakhstan,
E-mail: Tleu2009@mail.ru, https://orcid.org/0000-0002-1729-1321;
Aruova Aliya — Associate Professor, Astana IT University, Astana, Kazakhstan,
E-mail: a.aruova@astanait.edu.kz, https://orcid.org/0000-0001-7477-3856;
Naizagarayeva Akgul — Senior Lecturer, NCJSC «S.Seifullin Kazakh AgroTechnical Research University», Astana, Kazakhstan,
E-mail: a.naizagaraeva@kazatu.edu.kz, orcid.org/0000-0002-3616-0925.

Abstract. Urban traffic congestion is a major issue in modern cities. Rapid urbanization and increasing vehicle numbers intensify this problem. Accurate traffic prediction is essential for reducing congestion. It also improves road safety and supports intelligent transportation systems. The study aims to develop and compare machine learning models for regulating urban traffic. In addition, it aims to identify the most effective approach for predicting road traffic in advance. To achieve this goal, the study involves several tasks, including data collection and preprocessing, exploratory data analysis, feature engineering, and data normalization. Additionally, three machine learning algorithms—logistic regression, support vector machine (SVM), and Random Forest—were implemented and evaluated based on performance metrics such as accuracy, precision, recall, and

F1-score. The results of the study demonstrate that all models provide satisfactory performance; however, their effectiveness varies across traffic classes. Logistic Regression achieved an accuracy of 0.87 but showed weaker performance in the "low" traffic class. The SVC model improved classification balance and reached an accuracy of 0.88. The best performance was achieved by the Random Forest model, which demonstrated an overall accuracy of 0.99 and a weighted F1-score of 0.99, indicating high reliability and strong predictive capability across all traffic categories. The practical significance of the study lies in the potential application of the proposed model in real-time traffic management systems. The integration of machine learning models, particularly ensemble methods such as Random Forest, can support decision-making processes, optimize traffic flow, reduce congestion, and improve road safety. Furthermore, the developed approach can contribute to the advancement of intelligent transportation systems and serve as a foundation for future research involving real-time data and advanced predictive models. The findings indicate that ensemble learning methods, particularly Random Forest, provide superior accuracy and reliability for urban traffic prediction tasks. The proposed model can be integrated into real-time traffic management systems to enhance decision-making and improve urban mobility.

Keywords: Machine learning, Random forest, Support Vector Machine, Logistic Regression, Traffic classification, Urban traffic management

For citations: Mimenbayeva A.B., Turebayeva R. D., Ospanova T.T., Aruova A.B., Naizagarayeva A.A. Development and comparative analysis of machine learning models for urban traffic prediction. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 253–270. DOI: https://doi.org/10.32014/2026.2518-1726.412

© Миненбаева А.Б.[1], Туребаева А.Д.[2], Оспанова Т.Т.[2]*, Аруова А.Б.[1] , Найзағараева А.А.[3], 2026.

[1]Astana IT University, Астана, Қазақстан;

[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;

[3]«С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан.

E-mail: Tleu2009 @mail.ru

ҚАЛАЛЫҚ КӨЛІК АҒЫНЫН БОЛЖАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН ӘЗІРЛЕУ ЖӘНЕ САЛЫСТЫРМАЛЫ ТАЛДАУ

Миненбаева Айгүл — Сеньор-лектор, Astana IT University, Астана, Қазақстан,
E-mail: A.Mimenbayeva@astanait.edu.kz, https://orcid.org/0000-0003-4652-470X;

Туребаева Рахиля — техника ғылымдарының кандидаты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: turebayeva_rd@enu.kz, https://orcid.org/0009-0006-4530-1300;

Оспанова Тлеуғайша — қауымдастырылған профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: Tleu2009@mail.ru, https://orcid.org/0000-0002-1729-1321;
Аруова Алия — қауымдастырылған профессор, Astana IT University, Астана, Қазақстан,
E-mail: a.aruova@astanait.edu.kz, https://orcid.org/0000-0001-7477-3856;
Найзағараева Акгүл — PhD, аға оқытушы, «С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан,
E-mail: a.naizagaraeva@kazatu.edu.kz, orcid.org/0000-0002-3616-0925.

Аннотация. Қазіргі заманғы қалаларда қалалық көлік кептелісі – басты мәселелердің бірі болып табылады. Жедел урбанизация және көліктер санының артуы бұл мәселенің артуына әкелуде. Кептелісті азайту үшін көліктің қозғалысын дәл болжау маңызды. Сонымен қатар, бұл жол қауіпсіздігін жақсартады және ақылды көлік жүйелерін қолдайды. Зерттеудің мақсаты – қалалық көлік қозғалысын реттеуге арналған машинамен оқыту модельдерін әзірлеу және салыстыру. Сонымен қатар, зерттеу жол қозғалысын алдын ала болжауда ең тиімді тәсілді анықтауды мақсат етеді. Осы мақсатқа жету үшін зерттеу деректерді жинау және алдын ала өңдеу, зерттеушілік деректер талдауы, ерекшеліктерді инженерлеу және деректерді нормалау сияқты бірнеше тапсырманы қамтиды. Сонымен қатар, логистикалық регрессия, қолдау векторлық машинасы (SVM) және Random Forest сияқты үш машинамен оқыту алгоритмі іске асырылып, дәлдік, нақтылық, қайта табу көрсеткіші және F1-шот сияқты өнімділік көрсеткіштері бойынша бағаланды. Зерттеу нәтижелері барлық модельдердің қанағаттанарлық өнімділік көрсеткенін көрсетеді; алайда олардың тиімділігі жол қозғалысы кластары бойынша әртүрлі. Логистикалық регрессия 0,87 дәлдікке жетті, бірақ «төмен» жол қозғалысы класында әлсіз өнімділік көрсетті. SVC моделі жіктеу тепе-теңдігін жақсартып, 0.88 дәлдікке жетті. Ең жақсы нәтиже Random Forest моделінде тіркелді: ол жалпы 0.99 дәлдік пен 0.99 салмақталған F1-балл көрсетті, бұл барлық жол қозғалысы категориялары бойынша жоғары сенімділік пен мықты болжау қабілетін көрсетеді. Зерттеудің практикалық маңызы ұсынылған модельді нақты уақыттағы жол қозғалысын басқару жүйелерінде қолдану мүмкіндігінде жатыр. Машинамен оқыту модельдерін, әсіресе Random Forest сияқты ансамбль әдістерін біріктіру шешім қабылдау процестерін қолдап, қозғалыс ағынын оңтайландырып, кептелісті азайтып, жол қауіпсіздігін арттыруға мүмкіндік береді. Сонымен қатар, әзірленген тәсіл интеллектуалды көлік жүйелерін дамытуға үлес қосып, нақты уақыттағы деректер мен жетілдірілген болжамдық модельдерді қамтитын болашақ зерттеулер үшін негіз бола алады.

Түйін сөздер: машиналық оқыту, кездейсоқ орман, тірек векторлар әдісі, логистикалық регрессия, трафикті жіктеу, қалалық трафикті басқару

© Миненбаева А.Б.[1], Туребаева А.Д.[2], Оспанова Т.Т.[3]*, Аруова А.Б.[1], Найзағараева А.А.[3], 2026.
[1]Astana IT University, Астана, Казахстан;
[2] Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан;
[3] НАО «Казахский агротехнический университет им. С. Сейфуллина», Астана, Казахстан.
E-mail: Tleu2009@mail.ru

РАЗРАБОТКА И СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ГОРОДСКОГО ТРАФИКА

Миненбаева Айгуль — Сеньор-лектор, Astana IT University, Астана, Казахстан,
E-mail: A.Mimenbayeva@astanait.edu.kz, https://orcid.org/0000-0003-4652-470X;
Туребаева Рахиля — кандидат технических наук, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан,
E-mail: turebayeva_rd@enu.kz, https://orcid.org/0009-0006-4530-1300;
Оспанова Тлеугайша — ассоциированный профессор, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан,
E-mail: Tleu2009@mail.ru, https://orcid.org/0000-0002-1729-1321;
Аруова Алия — ассоциированный профессор, Astana IT University, Астана, Казахстан,
E-mail: a.aruova@astanait.edu.kz, https://orcid.org/0000-0001-7477-3856;
Найзагараева Акгул — PhD, старший преподаватель, НАО «Казахский агротехнический университет им. С. Сейфуллина», Астана, Казахстан,
E-mail: a.naizagaraeva@kazatu.edu.kz, https://orcid.org/0000-0002-3616-0925.

Аннотация. Транспортная загруженность является одной из ключевых проблем современных городов. Быстрая урбанизация и рост числа транспортных средств существенно усиливают нагрузку на дорожную инфраструктуру. Точное прогнозирование транспортных потоков играет важную роль в снижении заторов, повышении безопасности дорожного движения и развитии интеллектуальных транспортных систем. Целью исследования является разработка и сравнительный анализ моделей машинного обучения для прогнозирования городского трафика и определения наиболее эффективного подхода к его раннему прогнозированию. Для достижения поставленной цели выполнены этапы сбора и предварительной обработки данных, разведочного анализа (EDA), инженерии признаков и нормализации данных. В рамках исследования реализованы и сравнены три алгоритма машинного обучения: логистическая регрессия, метод опорных векторов (SVM) и Random Forest. Оценка эффективности моделей проводилась с использованием стандартных метрик классификации: accuracy, precision, recall и F1-score. Результаты показали, что все модели демонстрируют удовлетворительную производительность, однако их эффективность различается в зависимости от уровня интенсивности

трафика. Логистическая регрессия достигла точности 0,87, однако показала более слабые результаты для класса с низкой интенсивностью движения. Модель SVM улучшила баланс классификации и достигла точности 0,88. Наилучшие результаты продемонстрировала модель Random Forest с общей точностью 0,99 и взвешенным F1-score 0,99, что свидетельствует о высокой надёжности и способности модели эффективно прогнозировать трафик во всех категориях. Полученные результаты подтверждают, что методы ансамблевого обучения, в частности Random Forest, обеспечивают более высокую точность и устойчивость прогнозирования городского трафика. Предложенная модель может быть интегрирована в системы управления дорожным движением в режиме реального времени для оптимизации принятия решений и повышения мобильности в городах.

Ключевые слова: машинное обучение, случайный лес, метод опорных векторов, логистическая регрессия, классификация трафиков, управление городским движением

Introduction. Nowadays, due to the increasing number of vehicles and urbanization, traffic congestion has become widespread in cities around the world. Accurate and efficient prediction of traffic flow on city roads plays a crucial role in reducing the negative consequences of traffic congestion. In other words, timely and accurate forecasting of road traffic is extremely important. This helps optimize traffic flows and reduce traffic jams and delays on the roads. Traffic forecasting systems are crucial for predicting such traffic jams and hazardous situations. They help to warn drivers and regulate traffic lights. Thus, such a system reduces the likelihood of accidents, which are common on city roads. Traffic jams caused by vehicles contribute to air pollution due to the increased emissions of harmful substances from stationary vehicles. Accurate and precise traffic forecasting enables the reduction of costs resulting from traffic congestion. (Lv et al., 2015).

Machine learning has become an essential tool for analyzing and predicting traffic patterns due to its ability to process large datasets and identify complex patterns. Machine learning algorithms can analyze historical traffic data, current updates, weather conditions, and other relevant factors to generate accurate predictions (Xie et al., 2020).

The adaptability and learning abilities of machine learning models allow us to constantly improve the accuracy of predictions. Consider three popular machine learning algorithms: logistic regression, random forest, and the support vector Machine method. Logistic regression is easy to implement and interpret; it works well on small datasets and in binary classification tasks. A random forest is an ensemble method that uses multiple decision trees to improve prediction accuracy and reduce the risk of overfitting. It is effective when working with large and complex data. The support vector machine (SVM) method is especially useful for classification problems with high-dimensional data and can be adapted for nonlinear decision boundaries using sound tricks. Each of these algorithms has

unique advantages and features that make them suitable for different types of data and tasks (Gao et al., 2024).

The object of the study is traffic flows and their characteristics, such as the number of cars that affect the overall transport situation. The number of cars is measured by the number of vehicles located on a certain length of road. This parameter determines the current state and dynamics of traffic, and also helps to analyze and predict the traffic situation.

The purpose of this study is to develop, compare, and evaluate the effectiveness of various machine learning algorithms for predicting traffic based on data obtained from a specially collected dataset, as well as integrating the best model into a web application for real-time use.

Objectives of the Research Work:

- Collection and preprocessing of data, including traffic parameters.
- Development of machine learning models using logistic regression, random forest and support vector machine.
- Evaluating the effectiveness of models using accuracy, completeness, and F1-measure metrics.
- Compare model performance and identify the best model for predicting traffic.
- Integration of the selected model into a web application for real-time traffic prediction.

The results of the study can be used to further develop intelligent transport systems and improve the management of urban traffic flows. Scientifically based conclusions about the most effective methods of predicting traffic can be in demand when creating and improving traffic management technologies, which in turn helps to reduce traffic jams, improve road safety, and improve the environmental situation in cities. Thus, this research represents an important contribution to the development of the field of intelligent transport systems and can be used as a basis for further research in this area.

Literature review

Traffic flow management in the city is an important part of overseeing the city's overall infrastructure. This factor affects every aspect, from the efficiency of residents' daily city travel to the speed of emergency response. The year-on-year increase in population and the rapid growth in the number of vehicles on city roads have become a traffic congestion problem.

To address this, researchers in the field use machine learning and big data analysis methods to offer more efficient traffic prediction and management. The proposed literature review examines several recent studies that have investigated the technologies used in intelligent transportation systems in this area.

Bu (2019) explores the application of deep learning, a sophisticated machine learning approach, to predict urban traffic. The study highlights the use of big data analytics to process and analyze the vast amount of data generated by urban traffic sensors, cameras, and GPS devices. Special attention is paid to the effectiveness of convolutional neural networks (CNN) and recurrent neural networks (RNN) in modeling and forecasting traffic flows.

CNNs are excellent at processing spatial data, which makes them suitable for analyzing images from street traffic cameras, while RNNs excel at analyzing temporal data, ideal for time series such as speed and volume of traffic. The main findings indicate that deep learning models, when trained on large and diverse datasets, can predict traffic conditions with high accuracy. The study also discusses the computing requirements and infrastructure needed to implement these models in real-time applications for intelligent transportation systems. The integration of deep learning into traffic management systems requires high computational resources, since models must be trained on large amounts of data and executed in real time to process incoming information from sensors and cameras. However, with proper configuration and optimization, deep learning can be effectively applied in real time to improve urban traffic management and ensure traffic safety and efficiency.

Convolutional neural network (CNN) models can process spatially oriented data very well. Therefore, such models are well-suited for analyzing images obtained from street traffic cameras. RNN networks, on the other hand, are well-suited for analyzing data that changes over time. These networks are often used in such tasks to analyze time series of data, such as traffic speed and flow.

Furthermore, in the proposed review, the authors discussed the computational requirements necessary to implement the aforementioned models in real-time applications of intelligent transportation systems. Primarily, there are some challenges in implementing neural network models in urban traffic management systems. Specifically, training models with large volumes of data requires extensive computational resources. Information about vehicle traffic continuously received from sensors and cameras requires constant processing. However, by properly configuring and optimizing the model's architecture, deep learning can be applied efficiently in real time. This reduces the challenges in urban traffic management and enhances safety and efficiency.

Hanan Almuhalfi (2024) has conducted an in-depth study of the machine learning models used in such tasks. In the study, the author identifies three main types of these models and highlights the advantages of each. The author notes that supervised learning models are well-suited for short-term traffic forecasting. Meanwhile, unsupervised learning helps detect anomalies in irregular traffic patterns. This article notes that the effectiveness of machine learning models often depends on the crucial role of effective feature selection.

Shahzad Ahmad's (2025) research focuses on a type of deep learning model, the Stacked Autoencoder (SAE), for traffic flow prediction. This model effectively encodes the most important features of the traffic, and further notes that traffic flow models can be used for prediction by reconstructing their samples from scratch. The author used data collected under various traffic conditions during the study. Based on this data, the SAE model, Methods such as Support Vector Machines (SVM) and decision trees were considered. The experimental results proved that the SAE model demonstrated higher test accuracy than the other two

models. Despite the computational complexity, this model can be scaled well. In other words, this model was particularly suitable for collecting and processing the large amounts of data encountered in city traffic.

In the next study, Xie et al. (2020) investigated urban traffic prediction using machine learning methods based on spatio-temporal data. The study employed models such as graph convolutional networks, Bayesian networks, and ensemble methods. The study employed models such as Bayesian networks and ensemble methods. The obtained results demonstrated that integrating spatial and temporal data significantly improves the accuracy of urban traffic flow prediction. The authors propose implementing deep learning models to achieve the best results in Intelligent Transportation Systems.

In conclusion, although recent years of research in this area have demonstrated the effective use of machine learning algorithms, real-time applications demand improved rapid decision-making. In particular, when forecasting traffic flows in large city centers, the scaling issues that arise in large-scale traffic flows must be addressed. Such calculations should necessarily include weather, special city events, and infrastructural changes (Datia et al., 2024). This issue leads to the use of distributed computing methods and cloud solutions for managing large-scale traffic (Teng et al., 2024).

Furthermore, the integration of effective machine learning models into real-time road traffic management systems is essential. Moreover, it is desirable for such systems to be accessible to the general public. It would also be beneficial for them to be designed with user-friendly interfaces and equipped with visualization tools.

Therefore, to predict and monitor city traffic, machine learning algorithms require improvement, and more research is needed in this field.

Materials and methods

To address the challenges encountered in urban traffic congestion management, this study conducted an extensive analysis of the dataset and employed three machine learning methods.The dataset used in the study was obtained from official traffic flow data from the Humanitarian Data Exchange (HDX) platform, and employed for research purposes.

This study aims to leverage advanced data analytics and machine learning techniques to predict traffic patterns and optimize traffic flow in urban settings.

Exploratory Data Analysis.

EDA is the initial stage of data analysis, the purpose of which is to gain an understanding of the data structure and identify the main characteristics. In this section, we apply various visualization and statistical analysis methods.

Data visualization

- Histograms: used to display the distribution of data for each variable.
- Scatter plots: used to show the relationship between two variables in the data.
- Heat maps: allow for the visualization of the correlation between given variables.

Statistical analysis

- Mean (μ): is calculated by dividing the sum of the data in the dataset by the total number of values:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1)$$

- Standard deviation: measures the distribution of data in the dataset around their mean (Moumen et al., 2023):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (2)$$

Pearson's correlation

Pearson Correlation

Pearson's r correlation is a statistical measure that indicates the linear relationship and its direction between input variables. This measure can take values from -1 to 1. Here, a value of 1 indicates the strongest positive correlation, -1 indicates the strongest negative correlation, and 0 indicates the complete absence of a linear relationship between the variables. The covariance is normalized by the product of the standard deviations of the variables to calculate this coefficient. A high positive coefficient indicates that as one variable increases, the other increases, while a high negative coefficient indicates that as one variable increases, the other decreases.

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 (y_i - \mu_y)^2}} \quad (3)$$

Feature engineering.

After completing the EDA, we move on to feature engineering, which aims to create new features that will improve the quality of our models.

1. Creation of Temporal Features. To create time features, we extracted the hour from the timestamp to reflect the daily traffic patterns. We also extracted the day of the week to account for weekly traffic variations. In addition, the month was extracted to reflect the monthly trends.

2. Data Normalization. Bringing data to a common scale to enhance model performance.

Normalization is a technique used to scale the values of a feature to a range between 0 and 1. This is done by subtracting the minimum value of the feature and then dividing by the range (difference between the maximum and minimum values).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

Model Training.

After preparing the data, we proceed to train machine learning models. We consider three main algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM).

Logistic Regression

Logistic regression is a statistical method used to model the probability of an observation belonging to a certain category or class. It is widely used in classification problems where it is necessary to predict the probability of an observation belonging to two or more classes. Instead of a direct relationship between features and the target variable, as in conventional linear regression, logistic regression uses a logistic function to convert the sum of weighted features into probability (Tripathi et al., 2023).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree in this model (Figure 1) is trained on a random partition of the training dataset (Jedhe et al., 2022). This method prevents overfitting and enhances the model's robustness. When predicting future data, each tree in the model independently forecasts the target variable. The final prediction aggregates the predictions obtained from all trees. Then, it calculates their mode or mean value to determine the final prediction. This ensemble Random Forest method has become widely popular due to its high accuracy, versatility, and ability to work efficiently with large datasets (Razali et al., 2021).

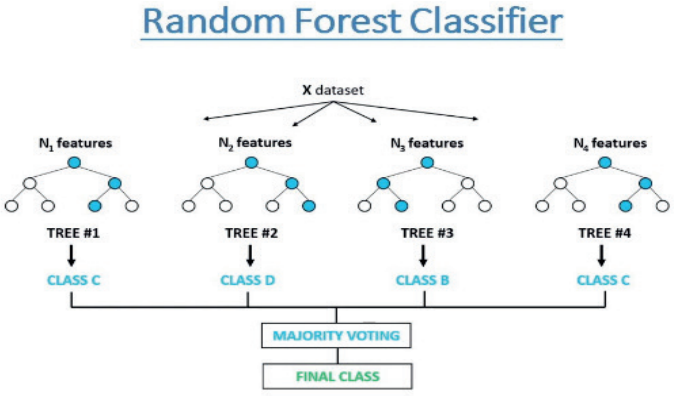


Figure 1 — Illustration of the Random Forest classification process.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) method is a machine learning algorithm that builds a hyperplane in the feature space to maximize the separation of two classes of data. It uses reference vectors - the data points closest to the separating hyperplane (Oyewola et al., 2022). SVM is effective even in the case of non-linearly separable data, due to the use of nuclear functions to transition to a higher-dimensional space (Balachander et al., 2022). The C regularization parameter in SVM controls the trade-off between maximizing gap and minimizing classification errors, making it a flexible tool for a variety of classification tasks (6).

$$f(x) = w^T x + b \quad (6)$$

Results

The results of the study present the performance of the proposed machine learning models and the analysis of traffic distribution patterns. The findings are illustrated using statistical metrics and visualizations to demonstrate differences between traffic categories.

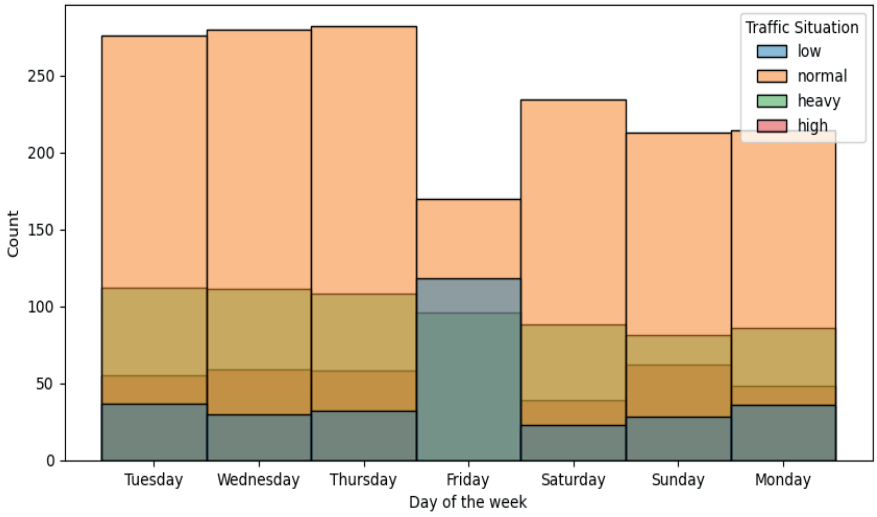


Figure 2 — Distribution of traffic situations.

The bar chart in Figure 2 displays the distribution of traffic situations across different days of the week. Each bar represents a day and is divided into segments corresponding to different traffic situations: low, normal, heavy, and high. The chart shows that traffic is predominantly normal across most days, with a notable decrease in traffic on Fridays. The data indicates variability in traffic patterns, which can be crucial for traffic prediction models. This visualization helps in understanding how traffic conditions fluctuate throughout the week, providing insights into peak and off-peak traffic days.

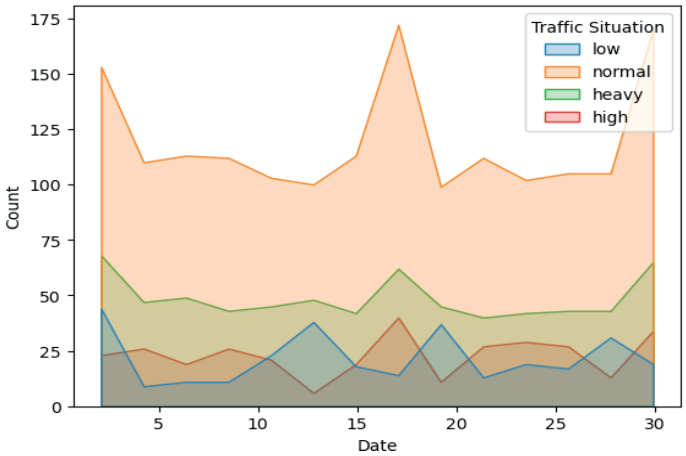


Figure 3 — Daily traffic situation over a month.

The graph (Figure 3) presents the daily traffic situation over a month, categorized into four types: low, normal, heavy, and high. The y-axis represents the count of traffic instances, while the x-axis indicates days of the month. The plot shows significant variability, particularly noticeable with 'high' traffic, which spikes dramatically towards the middle of the graph. This visualization helps in understanding the fluctuations in traffic conditions over time and can be used to predict traffic patterns or plan traffic management strategies.

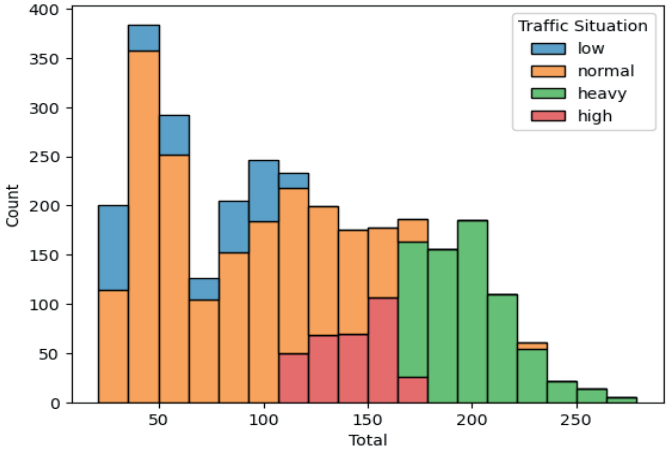


Figure 4- Distribution of traffic situations over a range of total counts.

The stacked bar (Figure 4) chart shows the distribution of traffic situations over a range of total counts. Each bar represents a total traffic count, segmented into four categories: low, normal, heavy, and high. The chart provides a clear view of how traffic conditions stack up against each other within various total count ranges. For instance, lower total counts tend to have a higher proportion of 'low' and 'normal'

traffic conditions, while higher total counts include more instances of ‘heavy’ and ‘high’ conditions. This graph is useful for analyzing the composition of traffic conditions relative to the overall traffic volume.

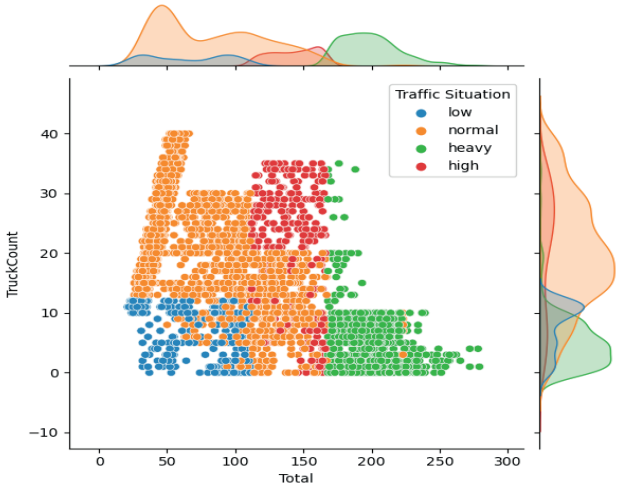


Figure 5 — Traffic situation in respective total counts and truck count.

The scatter plot in Figure 5 displays individual data points categorized by traffic situation against their respective total counts and truck counts. Each category (low, normal, heavy, high) is marked by a distinct color. The plot includes marginal histograms on the top and right, illustrating the distribution of total counts and truck counts, respectively. The scatter plot reveals a positive correlation between the total traffic counts and the severity of traffic conditions, suggesting that higher traffic volumes are often associated with more severe traffic situations.

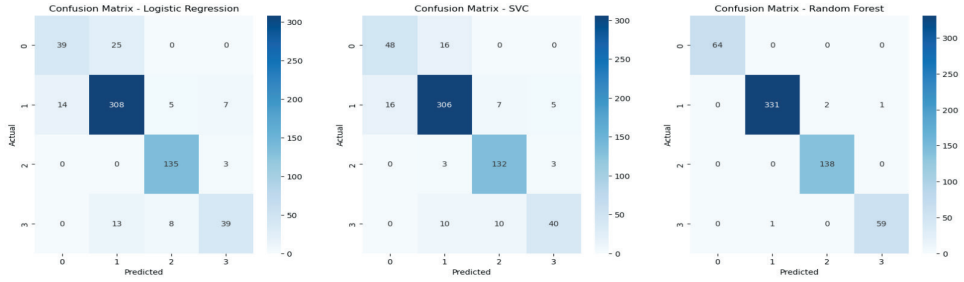


Figure 6 — Confusion matrices of models.

The image in Figure 6 shows three error matrices (confusion matrices) for three different machine learning models: logistic regression, SVC, and random forest. Each matrix shows the number of correct and incorrect predictions for four classes (0, 1, 2, and 3). For logistic regression, it can be noted that the model has difficulties classifying classes 0 and 3, often confusing them with other classes. The second

matrix for the SVC model shows that it does a better job with class 0 compared to logistic regression, but also has some errors in the classification of classes 2 and 3. The third matrix for a random forest demonstrates that this model does the best job with classification, making almost no errors for classes 0 and 2, and has the least number of errors for other classes**.**

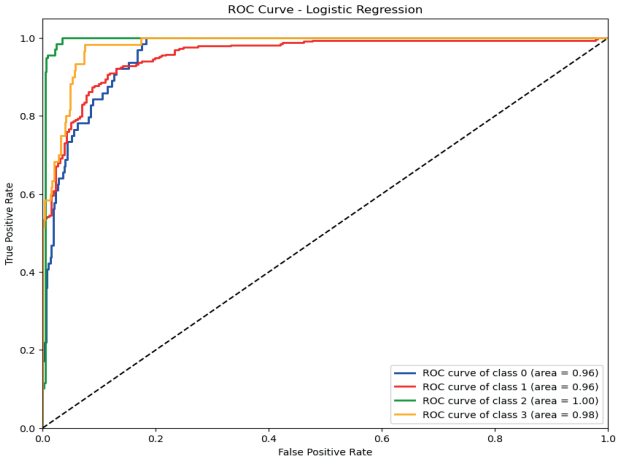


Figure 7 — ROC curves of the logistic regression model.

The graph of ROC curves in Figure 7 for logistic regression shows curves for four classes. The model shows high classification quality for all classes, with an area under the curve (AUC) from 0.96 to 1.00. Class 2 has an ideal classification with AUC=1.00, and classes 0 and 1 have AUC=0.96. This indicates that logistic regression copes well with the classification task, although for classes 0 and 1, improvements are possible.

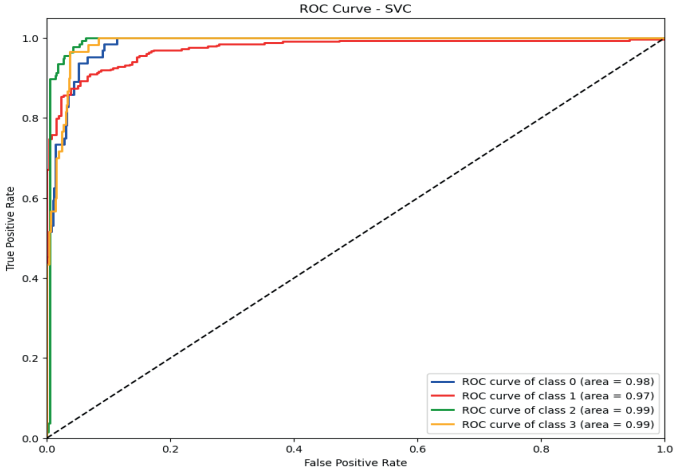


Figure 8 — ROC curves of the SVC model.

The ROC curve of the SVC model shown in Figure 8 demonstrates that it is very well classified across all given vehicle data classes. Here, the AUC values range from 0.97 to 0.99. The AUC values for class 0 were 0.98, and for class 2 they were 0.99. That is, this SVC model showed slightly better results compared to the previous logistic regression model. This SVC model is very reliable for such classification tasks.

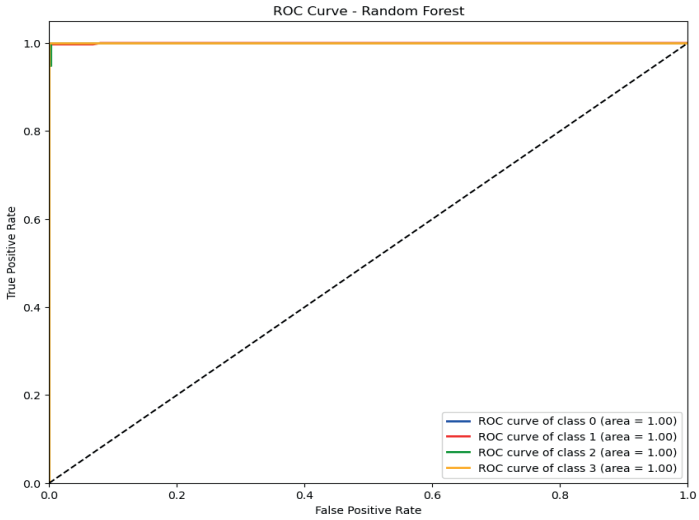


Figure 9 — ROC curves of the Random Forest model.

In the following graph in Figure 9, the ROC curves based on the results of the Random Forest model are shown. The AUC values for all classes are close to 0.99. That is, the trained model's classification quality is very high. As a result, the Random Forest model demonstrates much better discriminative ability for all classes compared to the previous two models (Figure 10).

Classification Report - Logistic Regression

	precision	recall	f1-score	support
low	0.74	0.61	0.67	64
normal	0.89	0.92	0.91	334
heavy	0.91	0.98	0.94	138
high	0.80	0.65	0.72	60
accuracy			0.87	596
macro avg	0.83	0.79	0.81	596
weighted avg	0.87	0.87	0.87	596

Classification Report - SVC

	precision	recall	f1-score	support
low	0.75	0.75	0.75	64
normal	0.91	0.92	0.91	334
heavy	0.89	0.96	0.92	138
high	0.83	0.67	0.74	60
accuracy			0.88	596
macro avg	0.85	0.82	0.83	596
weighted avg	0.88	0.88	0.88	596

Classification Report - Random Forest

	precision	recall	f1-score	support
low	1.00	1.00	1.00	64
normal	1.00	0.99	0.99	334
heavy	0.99	1.00	0.99	138
high	0.98	0.98	0.98	60
accuracy			0.99	596
macro avg	0.99	0.99	0.99	596
weighted avg	0.99	0.99	0.99	596

Figure 10 — Classification report of proposed models.

Classification Report - Logistic Regression:

Logistic regression shows accuracy (precision) and completeness (recall) for classes low (0.74 and 0.61), normal (0.89 and 0.92), heavy (0.91 and 0.98), high (0.80 and 0.65). Weighted averages (weighted avg) for accuracy, completeness and F1 measures are 0.87. The overall accuracy of the model is 0.87. The model copes best with the heavy and normal classes, but has difficulties with the low class.

Classification Report - SVC:

The SVC model shows better results compared to logistic regression for classes low (0.75 and 0.75), normal (0.91 and 0.92), heavy (0.89 and 0.96), high (0.83 and 0.67). The weighted averages for accuracy, completeness and F1 measures are 0.88. The overall accuracy of the model is 0.88. The model shows the best results for all metrics, especially for the low and high classes.

Classification Report - Random Forest:

The random forest shows the best results with accuracy and completeness for all classes equal to or almost equal to 1.00. The weighted averages for accuracy, completeness and F1 measures are 0.99. The overall accuracy of the model reaches 0.99. The model classifies all classes almost perfectly, which makes it the most effective among the presented models.

Disscussion

The results of the proposed study show that machine learning models demonstrate strong performance in predicting traffic flows. This study examined logistic regression, support vector classification (SVC), and Random Forest algorithms. Among the models considered, the Random Forest model, with an overall accuracy of 99.4%, was identified as the most effective. This result is regarded as significantly good among models used in traffic flow prediction. By accurately predicting traffic flow, it is possible to make traffic light schedules more efficient and provide drivers with real-time traffic information. This improves overall traffic flow and reduces congestion. As a result, it can lead to lower emissions and provide residents with convenient access.

The comparative analysis also highlights the importance of incorporating both spatial and temporal features in traffic modeling. The inclusion of time-based variables, such as hour and day of the week, contributed to improved classification

performance, confirming findings from previous studies on spatio-temporal traffic prediction.

Despite these promising results, several limitations should be noted. The dataset used in the study may not fully represent all possible traffic scenarios, particularly rare or extreme events such as accidents, road closures, or sudden weather changes. In addition, the models were evaluated in a controlled experimental setting, which may differ from real-world deployment conditions.

Furthermore, while Random Forest demonstrated near-perfect performance, such high accuracy may indicate potential overfitting or dataset bias, and therefore requires further validation on larger and more diverse datasets. Future research should focus on incorporating real-time data streams, external influencing factors (e.g., weather conditions, special events), and exploring advanced deep learning approaches to enhance model generalization and scalability.

Conclusion

In this study, machine learning models were used to predict urban traffic flow, and their performance was comparatively analyzed. The provided data were trained, tested, and evaluated using logistic regression, support vector machine (SVM), and Random Forest algorithms. Overall, the results of all three models were good. The logistic regression model used first yielded satisfactory results. However, this model is limited in handling more complex traffic patterns.

The SVC model's results were somewhat better than those of the first model, providing a more balanced classification. Among the models used, the Random Forest model proved the most reliable and highest in accuracy.

The obtained results determine the effectiveness of ensemble learning methods for traffic forecasting tasks in large cities. In addition, it was determined that the combined use of spatial and temporal data increases the model's performance. The proposed approach can be effectively applied in real-time traffic management systems to support decision-making, optimize traffic flow, and reduce congestion in urban environments.

However, in this regard, it is also necessary to note several limitations. The dataset used in this study may not cover all possible traffic scenarios on the road. In other words, situations such as sudden road closures caused by major accidents may not be accounted for. Additionally, the study did not account for external factors such as sudden weather changes, special city events, or road repairs. Incorporating these factors into the models could enhance their predictive accuracy and reliability.

Future research will focus on incorporating additional factors, such as weather and real-time data, to improve prediction accuracy. It is also important to explore scaling and advanced deep learning models for processing big city data.

References

Almukhalfi H., Noor A., & Noor T.H. (2024) Traffic management approaches using machine learning and deep learning techniques: A survey. Engineering Applications of Artificial Intelligence, 133(Part B), Article 108147. https://doi.org/10.1016/j.engappai.2024.108147 (in Eng.).

Balachander B., Kandasamy M., Dornadula V.H.R., Nirmal M., & Alanya-Beltran J. (2022) Integrating machine learning approaches in SDN for effective traffic prediction using correlation analysis. Communications in Computer and Information Science. — P. 611–622. https://doi.org/10.1007/978-3-031-07012-9_51 (in Eng.).

Bui K.-H. N., Yi H., Jung H., & Seo, J. (2019) Big data analytics-based urban traffic prediction using deep learning in ITS. In Proceedings of the International Conference on Artificial Intelligence (ICAI) (in Eng.).

Datia N., Pato M. P., Vaz, J., & Pires, J. M. (2024) Road traffic flow prediction with visual analytics. Studies in Computational Intelligence. — P. 293–318. https://doi.org/10.1007/978-3-031-46549-9_10 (in Eng.).

Deekshetha H. R., Shreyas Madhav A. V., & Tyagi A. K. (2022) Traffic prediction using machine learning. Lecture Notes on Data Engineering and Communications Technologies. — P. 969–983. https://doi.org/10.1007/978-981-16-9605-3_68 (in Eng.).

Gao M., Wei Y., Xie Y., & Zhang Y. (2024). Traffic prediction with self-supervised learning: A heterogeneity-aware model for urban traffic flow prediction. Mathematics, 12(9). — 1290 p. https://doi.org/10.3390/math12091290 (in Eng.).

Jedhe A., & Chopade S. (2022) Traffic prediction for intelligent transportation system using machine learning. International Journal for Research in Applied Science and Engineering Technology, 10(9). — P. 244–257. *https://doi.org/10.22214/ijraset.2022. 46604* (in Eng.).

Lv Y., Duan Y., Kang W., Li Z., & Wang F.-Y. (2015) Traffic flow prediction with big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 16(2). — P. 865–873. *https://doi.org/10.1109/TITS.2014.2345663* (in Eng.).

Moumen I., Abouchabaka J., & Rafalia N. (2023). Adaptive traffic lights based on traffic flow prediction using machine learning models. International Journal of Electrical and Computer Engineering, 13(5). — 5813 p. https://doi.org/10.11591/ijece.v13i5.pp5813-5823 (in Eng.).

Nashaat H., Mohammed N. H., Abdel-Mageid S. M., & Rizk R. Y. (2024) Machine learning-based cellular traffic prediction using data reduction techniques. IEEE Access. https://doi.org/10.1109/access.2024.3392624 (in Eng.).

Razali N.A.M., Shamsaimon N., Ishak K.K., Ramli S., Amran M.F.M., & Sukardi S. (2021) Gap, techniques and evaluation: Traffic flow prediction using machine learning and deep learning. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00542-7 (in Eng.).

Srivastava A., Aryan N., & Khurana M. (2023) A survey paper on traffic prediction using machine learning. International Journal for Research in Applied Science and Engineering Technology, 11(5). — P. 44–48. https://doi.org/10.22214/ijraset.2023.51390 (in Eng.).

Teng T., Jagadeesh G. R., Kunal T., & Chung C. C. (2024) Traffic state prediction of perturbed and non-perturbed traffic scenarios. Lecture Notes in Networks and Systems. — P. 129–144. https://doi.org/10.1007/978-3-031-47721-8_8 (in Eng.).

Tripathi A. N., & Sharma B. (2023). A deep review: Techniques, findings and limitations of traffic flow prediction using machine learning. Lecture Notes in Mechanical Engineering. — P. 21–33. https://doi.org/10.1007/978-981-99-2921-4_3 (in Eng.).

Wang Q., Chen C., Zhang L., Song X., Li H., Zhao Q., Niu B., & Gu J. (2024) Period extraction for traffic flow prediction. Lecture Notes in Computer Science. — P. 490–501. https://doi.org/10.1007/978-981-97-0798-0_29 (in Eng.).

Xie P., Li T., Liu J., Du S., & Yang X. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. Information Fusion, 59. — P. 1–12. https://doi.org/10.1016/j.inffus.2020.01.002 (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 271–286

https://doi.org/10.32014/2026.2518-1726.413

IRSTI 20.23.29
UDC 378.147:004.9

© Naumenko V.V.*, Mukanova Zh.A., Kiseleva O.V., Maintser D.A., Nerezov A.K., 2026.
Turan University, Almaty, Kazakhstan.
E-mail: v.naumenko@turan-edu.kz

THE USE OF REAL-TIME POLLING TO IMPROVE STUDENT ACADEMIC PERFORMANCE

Naumenko Vitaliy — PhD, Turan University, Almaty, Kazakhstan,
E-mail: v.naumenko@turan-edu.kz, https://orcid.org/0009-0005-9393-6733;
Mukanova Zhanna — PhD, Turan University, Almaty, Kazakhstan,
E-mail: zhanna.mukanova.83@mail.ru, https://orcid.org/0000-0001-6506-9007;
Kiseleva Olga — PhD, Turan University, Almaty, Kazakhstan,
E-mail: o.kisseleva@turan-edu.kz, https://orcid.org/0000-0001-8637-1416;
Maintser Dmitry — Master's degree, Turan University, Almaty, Kazakhstan,
E-mail: d.maintser@turan-edu.kz, https://orcid.org/0009-0006-7024-9700;
Nerezov Alexey — Turan University, Almaty, Kazakhstan,
E-mail: alexeynerezov@gmail.com, https://orcid.org/0009-0009-6440-311X.

Abstract. In the context of the digital transformation of higher education and the expansion of blended and online learning formats, universities require tools that provide immediate feedback and support improvements in student performance. Real-time polling systems represent one such tool, enabling instructors to assess students' understanding during lectures, seminars, or practical sessions and to adapt instruction accordingly. The relevance of the study is driven by the need for universities to adopt technologically accessible and methodologically sound tools to enhance student engagement and improve the quality of education. The study employs a literature review of research on active learning, formative assessment, and audience response systems, combined with a descriptive analysis of a practical case involving the implementation of a web-based platform for real-time polling and testing in a university setting. The functional features of the system, its architecture, and its data analytics capabilities are examined. The findings indicate that systematic use of real-time polls increases student engagement, enables early detection of knowledge gaps, and allows instructors to adjust the pace and structure of teaching. Integration of the platform with learning analytics and artificial intelligence tools makes it possible to track response patterns,

identify challenging topics, and support data-informed pedagogical decisions. These mechanisms contribute to more stable academic performance and reduce the accumulation of learning difficulties over time. The practical significance of the study lies in demonstrating a model for integrating real-time polling into a university's digital learning environment. The proposed approach can be used to enhance instructional quality, monitor student progress, and support data-driven management of the educational process, thereby contributing to more adaptive and evidence-based higher education practices.

Keywords: real-time polling, higher education, formative assessment, active learning, learning analytics, digital learning environment, student performance

For citations: Naumenko V.V., Mukanova Zh.A., Kiseleva O.V., Maintser D.A., Nerezov A.K. The use of real-time polling to improve student academic performance. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. *271–286. DOI: https://doi.org/10.32014/2026.2518-1726.413*

© Науменко В.В. *, Муканова Ж.А., Киселева О.В., Майнцер Д.А., Нерезов А.К., 2026.

«Тұран» университеті, Алматы, Қазақстан.
E-mail: v.naumenko@turan-edu.kz

БІЛІМ АЛУШЫЛАРДЫҢ ҮЛГЕРІМІН АРТТЫРУ ҮШІН REAL-TIME САУАЛНАМАЛАРЫН ҚОЛДАНУ

Науменко Виталий — PhD, «Тұран» университеті, Алматы, Қазақстан,
E-mail: v.naumenko@turan-edu.kz, https://orcid.org/0009-0005-9393-6733;
Муканова Жанна — PhD, «Тұран» университеті, Алматы, Қазақстан,
E-mail: zhanna.mukanova.83@mail.ru, https://orcid.org/0000-0001-6506-9007;
Киселева Ольга — PhD, «Тұран» университеті, Алматы, Қазақстан,
E-mail: o.kisseleva@turan-edu.kz, https://orcid.org/0000-0001-8637-1416;
Майнцер Дмитрий — магистр, «Тұран» университеті, Алматы, Қазақстан,
E-mail: d.maintser@turan-edu.kz, https://orcid.org/0009-0006-7024-9700;
Нерезов Алексей — «Тұран» университеті, Алматы, Қазақстан,
E-mail: alexeynerezov@gmail.com, https://orcid.org/0009-0009-6440-311X.

Аннотация. Жоғары білім беруді цифрландыру және аралас оқыту форматтарының кеңеюі жағдайында студенттердің оқу белсенділігі мен үлгерімін арттыруға мүмкіндік беретін жедел кері байланыс құралдарының маңызы артып келеді. Осындай құралдардың бірі – real-time сауалнамалар, олар оқытушыға сабақ барысында білім алушылардың материалды түсіну деңгейін жедел анықтауға және оқыту процесін бейімдеуге мүмкіндік береді. Зерттеудің өзектілігі жоғары оқу орындарында оқыту сапасын арттыруға бағытталған технологиялық және әдістемелік негізделген шешімдерге деген қажеттілікпен айқындалады. Зерттеу барысында белсенді оқыту, қалыптастырушы бағалау және аудиториялық жауап жүйелері бойынша

ғылыми әдебиеттерге талдау жасалды, сондай-ақ жоғары оқу орнында real-time сауалнамалар мен тестілеуді ұйымдастыруға арналған веб-платформаны енгізудің практикалық мысалы қарастырылды. Жүйенің функционалдық мүмкіндіктері, архитектурасы және деректерді талдау құралдары сипатталды. Нәтижелер real-time сауалнамаларды жүйелі қолдану студенттердің оқу процесіне қатысуын арттырып, білімдегі олқылықтарды ерте анықтауға және оқыту қарқынын реттеуге мүмкіндік беретінін көрсетті. Платформаны оқу аналитикасы және жасанды интеллект элементтерімен біріктіру жауаптардың динамикасын талдауға, күрделі тақырыптарды анықтауға және педагогикалық шешімдерді деректер негізінде қабылдауға жағдай жасайды. Бұл өз кезегінде үлгерімнің тұрақтануына және оқу барысында қателердің жинақталу қаупінің төмендеуіне ықпал етеді. Зерттеудің практикалық маңызы real-time сауалнамаларды университеттің цифрлық білім беру ортасына енгізу үлгісін ұсынумен анықталады. Ұсынылған тәсіл оқыту сапасын арттыруға, оқу нәтижелерін мониторингтеуге және деректерге негізделген білім беру процесін басқару жүйесін қалыптастыруға мүмкіндік береді.

Түйін сөздер: real-time сауалнамалар, жоғары білім, қалыптастырушы бағалау, белсенді оқыту, оқу аналитикасы, цифрлық білім беру ортасы, студенттердің үлгерімі

© Науменко В.В. *, Муканова Ж.А., Киселева О.В., Майнцер Д.А., Нерезов А.К., 2026.
Университет «Туран», Алматы, Казахстан.
E-mail: v.naumenko@turan-edu.kz

ПРИМЕНЕНИЕ ОПРОСОВ В РЕЖИМЕ REALTIME ДЛЯ ПОВЫШЕНИЯ УСПЕВАЕМОСТИ ОБУЧАЮЩИХСЯ

Науменко Виталий — PhD, Университет «Туран», Алматы, Казахстан,
E-mail: v.naumenko@turan-edu.kz, https://orcid.org/0009-0005-9393-6733;
Муканова Жанна — PhD, Университет «Туран», Алматы, Казахстан,
E-mail: zhanna.mukanova.83@mail.ru, https://orcid.org/0000-0001-6506-9007;
Киселева Ольга — PhD, Университет Туран, Алматы, Казахстан,
E-mail: o.kisseleva@turan-edu.kz, https://orcid.org/0000-0001-8637-1416;
Майнцер Дмитрий — магистр, Университет «Туран», Алматы, Казахстан,
E-mail: d.maintser@turan-edu.kz, https://orcid.org/0009-0006-7024-9700;
Нерезов Алексей — Университет «Туран», Алматы, Казахстан,
E-mail: alexeynerezov@gmail.com, https://orcid.org/0009-0009-6440-311X.

Аннотация: В условиях цифровизации высшего образования и распространения смешанных форм обучения возрастает потребность в инструментах, обеспечивающих оперативную обратную связь и повышение академической успеваемости студентов. Одним из таких инструментов являются опросы в режиме реального времени, позволяющие преподавателю

диагностировать уровень усвоения материала непосредственно в ходе занятия и адаптировать процесс обучения. Актуальность исследования обусловлена необходимостью внедрения технологически доступных и методически обоснованных решений для повышения учебной активности и качества образовательного процесса в высших учебных заведениях. В работе использованы методы анализа научной литературы по активному обучению, формирующему оцениванию и системам аудиторного отклика, а также описательный анализ практического кейса внедрения веб-платформы для проведения опросов и тестирования в режиме реального времени. Рассмотрены функциональные возможности системы, её архитектура и аналитические инструменты обработки данных. Результаты исследования показывают, что регулярное использование опросов в режиме реального времени способствует повышению вовлечённости студентов, своевременному выявлению пробелов в знаниях и адаптации темпа обучения. Интеграция платформы с инструментами учебной аналитики и элементами искусственного интеллекта позволяет анализировать динамику ответов, выявлять проблемные темы и поддерживать принятие педагогических решений. Эффект внедрения проявляется в стабилизации академической успеваемости и снижении риска накопления учебных затруднений. Практическая значимость исследования заключается в демонстрации модели интеграции опросов в режиме реального времени в цифровую образовательную среду вуза. Предложенный подход может быть использован для повышения качества обучения, мониторинга образовательных результатов и формирования данных-ориентированной системы управления образовательным процессом.

Ключевые слова: real-time опросы, высшее образование, формирующее оценивание, активное обучение, учебная аналитика, цифровая образовательная среда, успеваемость студентов

Введение. Применение опросов в режиме реального времени в образовательном процессе вуза в последние годы рассматривается как один из эффективных инструментов повышения учебной активности и успеваемости обучающихся, особенно в условиях цифровизации и массового внедрения смешанных и дистанционных форм обучения (Bond et al., 2020; Martin and Bolliger, 2018). Под real-time опросами в университетском контексте понимаются цифровые интерактивные инструменты, позволяющие преподавателю оперативно собирать ответы студентов в ходе лекции, семинара или практического занятия, анализировать их в режиме реального времени и немедленно адаптировать ход обучения (Kay and LeSage, 2009). В отличие от традиционных форм контроля знаний, таких как тесты по завершении темы или итоговые экзамены, real-time опросы встроены непосредственно в учебный процесс и выполняют не столько контролирующую, сколько диагностическую и формирующую функцию.

С теоретической точки зрения использование опросов в реальном времени

опирается на концепции активного обучения и формирующего оценивания. Современные исследования в области педагогики высшей школы показывают, что пассивное восприятие лекционного материала сопровождается быстрым снижением концентрации внимания, тогда как регулярное вовлечение студентов в микроактивности способствует удержанию внимания и более глубокому усвоению знаний. Real-time опросы выступают механизмом внешней активации познавательной деятельности, побуждая обучающихся регулярно соотносить новый материал со своими текущими представлениями и принимать осознанное участие в учебном процессе. С точки зрения когнитивной психологии это снижает эффект иллюзии понимания, когда студент субъективно считает материал усвоенным, не проходя проверку на уровне применения и интерпретации (Roediger and Karpicke, 2006).

Практическая значимость real-time опросов в вузе особенно проявляется на лекционных занятиях с большим количеством обучающихся, где традиционные формы обратной связи затруднены (Kay and LeSage, 2009). Использование цифровых платформ для мгновенных опросов позволяет преподавателю получить агрегированную картину понимания материала всей аудиторией в течение нескольких секунд. Например, после объяснения ключевого теоретического положения преподаватель может задать концептуальный вопрос с несколькими вариантами ответов, а результаты сразу отобразятся в виде распределения. Если значительная доля студентов выбирает неверный вариант, это служит сигналом о необходимости дополнительного пояснения или изменения темпа изложения. Таким образом, учебный процесс становится адаптивным не на уровне курса, а на уровне конкретного занятия.

С технической стороны реализация real-time опросов в вузе опирается на веб-ориентированные клиент-серверные решения, доступные с мобильных устройств и персональных компьютеров. Наиболее распространённые платформы, такие как Kahoot, Mentimeter и Poll Everywhere, используют модель, при которой преподаватель создаёт вопросы через веб-интерфейс, а студенты подключаются к сессии с помощью браузера или мобильного приложения. Ответы передаются на сервер в режиме реального времени, агрегируются и визуализируются для преподавателя и, при необходимости, для аудитории. Такая архитектура не требует сложной локальной установки и легко масштабируется для больших потоков, что делает её особенно привлекательной для университетской среды (Wang and Tahir, 2020).

С точки зрения повышения успеваемости ключевым является не сам факт использования опросов, а способ их педагогического и методического встраивания. Эмпирические исследования в высшем образовании показывают, что регулярное применение формирующего оценивания может приводить к статистически значимому росту итоговых результатов обучения, особенно в дисциплинах с высокой когнитивной нагрузкой, таких как математика, инженерные науки и экономика. Real-time опросы способствуют

выявлению пробелов в знаниях на ранних этапах, когда коррекция ещё возможна без накопления системных ошибок (Roediger and Karpicke, 2006). Например, в курсах высшей математики оперативная проверка понимания промежуточных шагов доказательства или преобразования формул позволяет предотвратить ситуацию, при которой студент теряет нить рассуждений и перестаёт понимать дальнейший материал.

Особую роль real-time опросы играют в семинарских и практических занятиях, где они используются для активизации дискуссии и коллективного анализа. Преподаватель может предложить проблемный вопрос, не имеющий однозначного ответа, и затем обсудить распределение мнений в группе. В вузовском образовании такой подход особенно ценен, поскольку формирует навыки аргументации, критического мышления и научной дискуссии. Технические возможности современных платформ позволяют использовать не только закрытые вопросы, но и короткие текстовые ответы, которые затем группируются по смыслу или обсуждаются выборочно. Это делает возможным вовлечение даже тех студентов, которые в традиционном формате редко вступают в устный диалог (Wang and Tahir, 2020).

В контексте дистанционного и гибридного обучения real-time опросы приобретают дополнительное значение как инструмент поддержания учебной дисциплины и присутствия. Отсутствие физического контакта с аудиторией повышает риск снижения вовлечённости и формального участия в занятиях. Интерактивные опросы в режиме реального времени позволяют преподавателю отслеживать активность обучающихся, а студентам — ощущать свою включённость в общий процесс. Практика показывает, что регулярные интерактивные элементы снижают долю пассивных слушателей и положительно влияют на посещаемость онлайн-занятий, что косвенно отражается и на итоговой успеваемости.

С технической и аналитической точки зрения данные, собираемые через real-time опросы, представляют собой ценный источник информации для мониторинга учебного процесса. При систематическом использовании результаты опросов могут сохраняться и анализироваться в динамике, позволяя выявлять типичные трудности, сравнивать группы и оценивать эффективность изменений в преподавании. В вузах с развитой цифровой инфраструктурой такие данные интегрируются с системами управления обучением, формируя основу для учебной аналитики. При этом важно учитывать требования к защите персональных данных и анонимизации, особенно при использовании опросов для диагностических целей.

Таким образом, применение опросов в режиме real-time в образовательном процессе вуза представляет собой эффективный инструмент повышения успеваемости обучающихся за счёт усиления обратной связи, адаптации преподавания и роста учебной вовлечённости. Их педагогическая ценность обусловлена сочетанием активного обучения и формирующего оценивания, а технологическая — доступностью, масштабируемостью и

возможностью аналитической обработки данных. В условиях современного университета real-time опросы перестают быть вспомогательным элементом и становятся частью системного подхода к управлению качеством обучения, ориентированного на своевременное выявление проблем и поддержку устойчивых образовательных результатов.

Литературный обзор. Эмпирическая база, позволяющая обосновать применение опросов в режиме реального времени как средства повышения успеваемости в вузе, опирается на два взаимодополняющих массива результатов: исследования по активному обучению в целом и специализированные исследования по системам аудиторийного отклика, включая «кликеры» и современные веб-платформы опросов (Theobald et al., 2020). В метаанализе по активному обучению в STEM-дисциплинах показано, что переход от преимущественно лекционного формата к активным практикам в среднем повышает результаты экзаменов примерно на шесть процентных пунктов, а вероятность неуспешного завершения курса (fail rate) заметно выше в традиционных лекционных группах; авторы также приводят суммарный эффект на уровне около 0,47 стандартного отклонения по показателям экзаменов и концептуальных тестов, что в педагогических исследованиях трактуется как практически значимый эффект (Freeman et al., 2014). Сходные выводы о влиянии активного обучения на академические результаты и сокращение разрывов в успеваемости между различными группами студентов получены и в более поздних метааналитических работах (Theobald et al., 2020). Важно, что в ряде включённых исследований активные элементы занимали лишь небольшую долю аудиторного времени, что делает именно micro-интервенции, к которым относятся и real-time опросы, методологически релевантными для массовых университетских курсов с ограниченным временем на взаимодействие (Schneider and Preckel, 2017).

В более узком сегменте работ по аудиторийным системам ответа метаанализ Hunsu и соавторов показывает статистически значимый, хотя и умеренный, прирост когнитивных результатов при использовании clicker-подобных технологий, а также более выраженные эффекты по некогнитивным переменным, включая вовлечённость и отношение студентов к обучению; это принципиально для вуза, где именно вовлечённость и регулярность учебных действий часто выступают промежуточным механизмом, через который меняется итоговая успеваемость (Hunsu et al., 2016). Дополнительные эмпирические исследования подтверждают, что использование систем аудиторийного отклика повышает интерактивность занятий и способствует более глубокому осмыслению материала при условии включения обсуждения и повторного голосования (Mayer et al., 2009; Blasco-Arcas et al., 2013). Обзор Caldwell по крупным аудиториям подчёркивает, что влияние «кликеров» на экзаменационную успеваемость может быть от нейтрального до положительного и существенно зависит от педагогического дизайна, прежде всего от того, организуются ли обсуждение, интерпретация распределения

ответов и повторное голосование, а не просто фиксация результата (Caldwell, 2007).

Дополнительный аргумент в пользу систематического, а не эпизодического применения таких инструментов дают современные обзоры, в том числе систематическая работа 2025 года в рамках PRISMA-подхода, где анализируется совокупность исследований по ARS в высшем образовании и обсуждаются как эффекты, так и методологические ограничения измерений мотивации и знаний (Serrada-Sotil et al., 2025). В контексте цифровых платформ опросов исследования показывают, что веб-ориентированные системы (например, Kahoot и аналогичные инструменты) положительно влияют на вовлечённость, мотивацию и участие студентов в учебной деятельности, что косвенно связано с повышением академических результатов (Wang and Tahir, 2020; Licorish et al., 2018).

Практическая логика влияния real-time опросов на успеваемость в вузе раскрывается через механизм «быстрой диагностики» и «немедленной коррекции», который в учебной аналитике можно описать как замыкание короткого контура обратной связи между ошибкой и интервенцией. Теоретические модели формирующего оценивания показывают, что своевременная обратная связь способствует саморегуляции обучения и повышению результатов за счёт раннего выявления затруднений (Nicol and Macfarlane-Dick, 2006; Shute, 2008). В традиционном курсе значительная часть ошибок обнаруживается поздно — на контрольной, рубежном тесте или экзамене, когда образовательный ущерб уже накоплен. При опросах в реальном времени преподаватель получает сигнал о неверном понимании в момент формирования концепта, что позволяет либо повторно объяснить ключевой шаг, либо предложить контрпример, либо организовать краткую дискуссию «почему этот вариант неверен». На крупных потоках это особенно ценно, потому что даже при высокой квалификации преподавателя без цифрового инструмента он наблюдает лишь небольшую долю студентов и часто переоценивает степень понимания аудитории. Эффект здесь проявляется не только в росте средних баллов, но и в снижении дисперсии результатов: студенты, которые «теряются» в середине темы, получают шанс на раннюю коррекцию, а значит снижается вероятность провала на последующих разделах, что согласуется с выводами метаанализа по активному обучению о снижении доли неуспевающих в активных форматах (Freeman and др., 2014; Schneider and Preckel, 2017).

Материалы и методы. На прикладном уровне для вузов характерно несколько устойчивых кейсов, которые технически различаются по типу вопроса, каналу доставки и способу обработки данных. В теоретически насыщенных дисциплинах, например в математическом анализе, теории вероятностей, микроэкономике или электродинамике, наиболее результативны концептуальные вопросы на «узловых местах» объяснения, когда проверяется не вычислительная техника, а понимание смысла

определения или условия применимости теоремы (Prince, 2004). В таких курсах real-time опрос обычно строится как двухтактная процедура: сначала индивидуальный ответ, затем короткое обсуждение в парах или мини-группах и повторное голосование; преподаватель анализирует распределение ответов, комментируя типичные ошибки. Именно такой дизайн, а не сам факт клика, в обзорах по ARS рассматривается как объяснение различий в эффекте на экзамен (Caldwell, 2007; Mayer et al., 2009).

В прикладных курсах, например в программировании и анализе данных, опросы могут быть встроены в практику отладки: студентам предлагается выбрать причину ошибки, корректную асимптотику алгоритма или интерпретацию метрики модели, а затем преподаватель показывает, как по распределению ответов диагностируется «массовая» ошибка мышления. В гуманитарных и управленческих программах опросы в реальном времени используются как средство структурирования дискуссии: студенты оценивают аргументы по заданным критериям или выбирают интерпретацию кейса. Обязательный эффект проявляется в повышении качества аргументации и вовлеченности, что соответствует данным метаанализов о значимости некогнитивных эффектов. ARS нередко выражены сильнее когнитивных (Hunsu et al., 2016; Blasco-Arcas and др., 2013). Дополнительные исследования показывают, что интерактивные системы опросов усиливают участие студентов и глубину обсуждения при условии включения элементов peer discussion и повторного голосования (Kay and LeSage, 2009; Licorish et al., 2018).

Технический аспект внедрения real-time опросов в вузе имеет принципиальное значение, поскольку определяет надежность, масштабируемость и пригодность данных для последующего анализа. Современные платформы, как правило, реализованы в виде облачных веб-приложений с архитектурой «преподаватель – сервис – студенты»: обучающиеся подключаются через браузер или мобильное приложение, а передача ответов осуществляется по протоколам с низкой задержкой, чаще всего с использованием WebSocket или аналогичных механизмов push-обновления (Wang and Tahir, 2020). Низкая латентность важна не только с технической, но и с дидактической точки зрения: мгновенное отображение результатов позволяет преподавателю сразу интерпретировать распределение ответов и встроить корректирующее объяснение в ход занятия, сохраняя причинно-следственную связь между вопросом, ошибкой и ее разбором (Nicol and Macfarlane-Dick, 2006; Shute, 2008). Для крупных потоков ключевым фактором является устойчивость к пиковым нагрузка: облачные сервисы обеспечивают ее за счет балансировки и масштабирования, тогда как при on-premise внедрении вузу необходимо самостоятельно решать задачи планирования, особенно при одновременном использовании в нескольких корпусах.

Результаты. Отдельное направление практики связано с внедрением опросов в цифровую образовательную среду университета. Когда такие

инструменты используются изолированно, они дают быстрый эффект в рамках конкретного занятия, но не позволяют проводить долгосрочный анализ. Однако при интеграции с LMS через LTI или API открываются более широкие возможности: результаты опросов можно связать с темами курса, активностью студентов и итоговыми оценками, выявлять проблемные темы и строить модели раннего предупреждения. В этом случае опросы в реальном времени перестают быть лишь интерактивным элементом и становятся частью системы управления качеством обучения, поскольку данные становятся сопоставимыми между группами и семестрами.

В университетской практике всё чаще применяется гибридный подход: для текущей активности используются простые облачные сервисы, а для оценочных мероприятий — встроенные инструменты LMS, что обеспечивает требования к идентификации, журналированию и хранению данных. С исследовательской точки зрения это важно, так как успеваемость начинает рассматриваться не как итоговый результат, а как динамическая траектория обучения. Опросы выступают в роли частых индикаторов понимания, позволяя количественно оценивать влияние изменений в преподавании (Caldwell, 2007).

Если опросы в реальном времени встроены в структуру занятий как средство, регулярно формирующее диагностику, то их эффект становится объяснимым и воспроизводимым, поскольку таким образом увеличивается частота извлечения знаний из памяти и снижается задержка между между ошибкой и ее исправлением. Согласно метааналитическим исследованиям по активному обучению связаны с устойчивым ростом учебныъ результатов и снижением числа неуспевающих обучающихся.

На рисунке 1 представлен интерфейс административной панели цифровой системы. Она предназначена для создания, управления и анализа онлайн-тестирований и опросов, что позволяет перейти от теоретического обсуждения real-time опросов к конкретному прикладному применению в образовательном процессе. Данная система демонстрирует, каким образом принципы интерактивного формирующего оценивания и оперативной обратной связи могут быть реализованы в виде полноценного программного решения, интегрируемого в учебную инфраструктуру университета. Наличие инструментов для быстрого создания тестов, мониторинга количества участников и управления процедурами опроса в режиме реального времени создаёт техническую основу для внедрения real-time опросов непосредственно в аудиторные и онлайн-занятия. В дальнейшем рассматриваемый кейс демонстрирует, как разработанная система используется для проведения оперативных опросов обучающихся, сбора данных об уровне понимания учебного материала и последующего анализа результатов с целью повышения успеваемости и адаптации образовательного процесса под реальные потребности студентов.

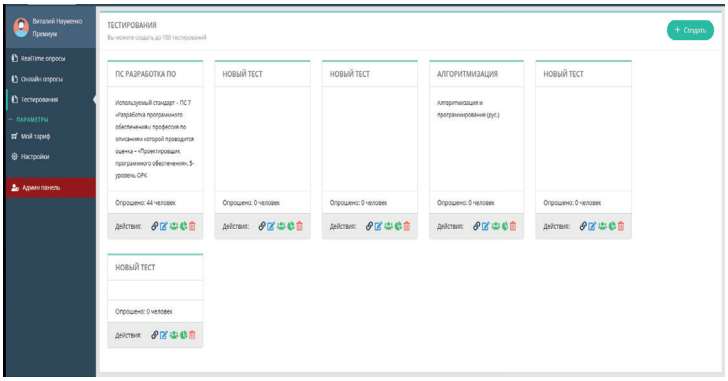


Рисунок 1 – Административная панель с возможностью создания и редактирования тестирования.

Административная панель отечественной веб-платформы eee.kz предназначена для создания и управления онлайн-тестированием и опросами в режиме реального времени и позволяет преподавателю быстро создавать и проводить такие опросы, затрачивая на это минимальное время. В левой части экрана расположено вертикальное навигационное меню, включающее разделы «Realtime опросы», «Онлайн опросы», «Тестирования», «Параметры», «Мой тариф», «Настройки» и «Админ панель». Рабочая область платформы содержит список карточек созданных тестов. Каждая карточка содержит название теста, краткое описание, информацию о количестве опрошенных участников, а также набор функциональных кнопок для редактирования, копирования, просмотра, экспорта и удаления теста. Кнопка «Создать», расположенная в верхней части интерфейса, предназначена для добавления нового теста. Интерфейс выполнен в минималистичном стиле с акцентом на четкую визуальную структуру и удобную навигацию, что обеспечивает простое управление тестовыми материалами, эффективный мониторинг активности и оперативное администрирование оплайн-опросов.

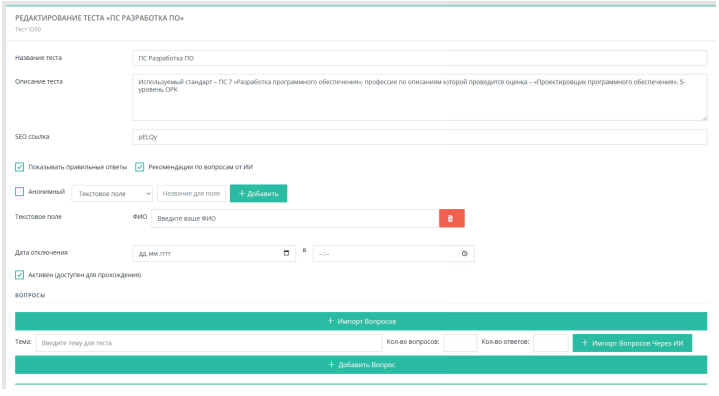


Рис. 2 – Шаблон для добавления тестовых заданий.

На рисунке 2 представлен интерфейс редактирования теста в административной панели веб-платформы онлайн-тестирования. В верхней части экрана отображается заголовок «Редактирование теста», а также уникальный идентификатор теста. Центральная область интерфейса содержит форму ввода и редактирования основных параметров тестирования. Поля формы включают: название теста, описание, SEO-ссылку, параметры отображения правильных ответов и рекомендаций на основе искусственного интеллекта, а также настройки анонимности и сбора персональных данных. Имеется возможность добавлять пользовательские текстовые поля. Например, для указания фамилии, имени и отчества участника. Также система позволяет установить дату и время автоматического отключения теста, а также управлять его активностью с помощью соответствующего переключателя. В нижней части интерфейса расположен блок работы с вопросами, включающий инструменты импорта, автоматического генеративного добавления заданий с использованием технологий искусственного интеллекта, а также ручного создания и редактирования структуры теста. Такая организация интерфейса позволяет обеспечивать гибкость настроек контента, автоматизацию подготовки материалов и контроль проведения онлайн-оценивания.

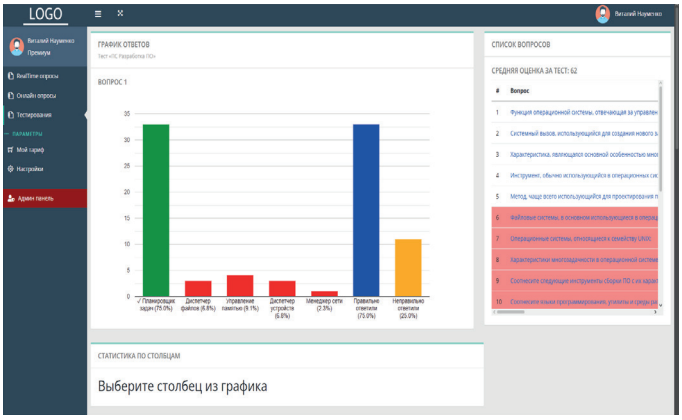


Рис 3. – Статистика прохождения тестирования.

На рисунке 3 представлен интерфейс аналитического модуля для визуализации и анализа результатов прохождения теста. В центральной части интерфейса представлен график распределения ответов по первому вопросы теста, который отражает абсолютное количество и процентное соотношение выбранных вариантов. Цветовое выделение столбцов обеспечивает наглядное сравнение популярности ответов и позволяет быстро выявлять наиболее типичные ошибки.

В правой части интерфейса размещен список всех вопросов теста в указанием их порядковых номеров, что обеспечивает удобную навигацию и быстрый доступ к статистике по каждому заданию. В верхней части

панели отображается средний балл за тест, отражающий общий уровень успешности его выполнения. В нижней части интерфейса расположен блок дополнительной аналитики, позволяющий выбирать отдельные элементы графика для более детального анализа. Така организация интерфейса обеспечивает комплексную оценку результатов тестирования, способствует выявлению проблемных тем и повышает эффективность педагогической диагностики и последующей корректировки учебного процесса.

На рисунке 4 представлена панель управления группами пользователей в системе онлайн-тестирования. В верхней части интерфейса расположена кнопка создания новой группы, позволяющая добавлять дополнительные категории пользователей. Ниже приведен сводный список всех существующих групп, связанных с данным тестом, что обеспечивает полный контроль над распределением доступа. Такая организация интерфейса обеспечивает гибкое управление контингентом обучающихся, позволяет оперативно формировать выборки для проведения администрирования онлайн-тестирования в образовательной среде.

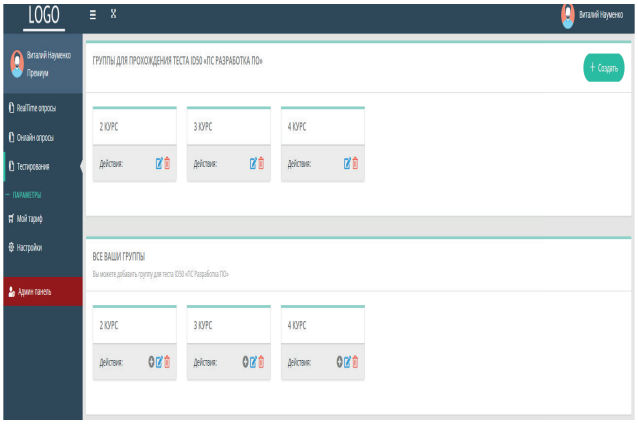


Рис 4. – Разделение тестируемых по группам.

Обсуждение. Функциональность разработанной системы не ограничивается проведением опросов и тестирования в режиме реального времени, а расширяется за свет интеграции инструментов анализа данных и методов искусственного интеллекта. Это позволяет рассматривать ее как средство мониторинга и аналитической поддержки образовательного процесса. Данные, собираемые в ходе опросов, формируют репрезентативный массив информации об учебной активности и уровне усвоение материала, который может использоваться для автоматизированного анализа успеваемости как на уровне отдельного занятия, так и в рамках всего курса. В отличие от традиционных форм контроля, ориентированных на итоговую оценку, система обеспечивает накопление детализированных данных о промежуточных результатах, включая ответы на концептуальные вопросы,

динамику ошибок и скорость реагирования, что создаёт основу для более глубокого и объективного анализа учебного процесса.

Интеграция методов искусственного интеллекта позволяет системе автоматически выявлять закономерности в результатах опросов и интерпретировать их с точки зрения усвоения учебного материала. На уровне отдельного занятия алгоритмы анализа данных позволяют выявлять темы и концепты, вызывающие наибольшие затруднения у обучающихся. Это дает преподавателю возможность оперативно корректировать объяснение материала, возвращаться к проблемным аспектам и предлагать дополнительные примеры. При накоплении данных за несколько занятий или модулей система формирует целостный профиль усвоения дисциплины, отражающий сильные и слабые стороны групп.

С точки зрения индивидуальной аналитики система представляет инструменты для отслеживания учебной траектории каждого студента. На основе результатов real-time опросов и тестирований, а также их динамики, алгоритмы машинного обучения могут классифицировать обучающихся по уровню освоения материала, выделять группы с высоким, низким и средним уровнем подготовки. Такая классификация носит диагностический характер и используется для своевременного выявления рисков снижения успеваемости и планирования адресных педагогических вмешательств.

Практическая значимость данной функциональности заключается в переходе от реактивного к проактивному управлению учебным процессов, Система не только фиксирует низкие результаты, но и анализирует их контекст, включая повторяемость ошибок, нестабильность ответов и снижение активности. Это позволяет выявлять обучающихся, испытывающих трудности, на ранних этапах, когда корректирующие меры наиболее эффективны. Одновременно система помогает идентифицировать сильных студентов, создавая условия для индивидуализации обучения, вовлечение их в проектную и исследовательскую деятельность, а также применения элементов взаимного обучения в группе.

Заключение. Проведенный анализ показывает, что опросы в режиме реального времени являются эффективным инструментом повышения учебной активности и успеваемости студентов за счет усиления формирующей обратной связи и сокращения времени между допущенной ошибкой и ее корректировкой. Их эффективность определяется не столько самим использованием цифровых платформ, сколько качеством педагогического дизайна: постановкой концептуально значимых вопросов в ключевые моменты объяснения, обсуждением полученных результатов и адаптацией хода занятия на основе обратной связи. В этом контексте real-time опросы выступают инструментом оперативного управления учебным процессов и способствует более устойчивому усвоению материала.

Рассмотренный прикладной кейс демонстрирует, что интеграция системы онлайн-опросов в цифровую образовательную среду вуза позволяет

использовать ее не только для текущей диагностики, но и в качестве инструмента учебной аналитики. Сбор и обработка данных об ответах обучающихся создают основу для выявления типичных затруднений, отслеживая динами освоения дисциплины и раннего выявления рисков снижения успеваемости. Применение методов искусственного интеллекта дополнительно расширяет эти возможности, обеспечивая автоматизированный анализ результатов и формирование рекомендаций по корректировке преподавания.

Практическая значимость системы заключается в переходе от эпизодного контроля знаний к непрерывному мониторингу и поддержке учебных достижений. Это способствует принятию более обоснованных педагогических решения и формированию модели управления качеством обучения, основанной на данных. Вместе с тем эффективность внедрения во многом зависит от уровня развития цифровой инфраструктуры и методической готовности преподавателей. Перспективным направлением дальнейших исследований является проведение эмпирических работ, направленных на количественную оценку влияния real-time опросов и аналитических модулей на успеваемость и вовлеченность обучающихся.

References

Blasco-Arcas L., Buil I., Hernández-Ortega B., & Ses, F.J. (2013) Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. Computers & Education, 627 — P. 102-110. https://doi.org/10.1016/j.compedu.2012.10.019 (in Eng.).

Bond M., Bedenlier S., Marín V.I., & Händel M. (2021) Emergency remote teaching in higher education: Mapping the first global online semester. International journal of educational technology in higher education, 18(1), 50. https://doi.org/10.1186/s41239-021-00282-x (in Eng.).

Caldwell J.E. (2007) Clickers in the large classroom: Current research and best-practice tips. CBE—Life Sciences Education, 6(1). — P. 9-20. https://doi.org/10.1187/cbe.06-12-0205 (in Eng.).

Freeman S., Eddy S L., McDonough M., Smith M.K., Okoroafor N., Jordt H., & Wenderoth M.P. (2014) Active learning increases student performance in science, engineering, and mathematics. Proceedings of the national academy of sciences, 111(23). — P. 8410-8415. https://doi.org/10.1073/pnas.1319030111 (in Eng.).

Hunsu N.J., Adesope O., & Bayly D.J. (2016) A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. Computers & Education, 94. — P. 102-119. https://doi.org/10.1016/j.compedu.2015.11.013 (in Eng.).

Kay R.H., & LeSage A. (2009) Examining the benefits and challenges of using audience response systems: A review of the literature. Computers & Education, 53(3). — P. 819-827. https://doi.org/10.1016/j.compedu.2009.05.001 (in Eng.).

Licorish S.A., Owen H.E., Daniel B., & George J.L. (2018) Students' perception of Kahoot!'s influence on teaching and learning. Research and Practice in Technology Enhanced Learning, 13(1). — P. 1-23. (in Eng.).

Martin F., & Bolliger D.U. (2018) Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. Online learning, 22(1), 205-222. (in English).

Mayer R.E., Stull A., DeLeeuw K., Almeroth K., Bimber B., Chun D., ... & Zhang H. (2009) Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. Contemporary educational psychology, 34(1). — P. 51-57. https://doi.org/10.1016/j.cedpsych.2008.04.002 (in Eng.).

Nicol D.J., & Macfarlane-Dick D. (2006) Formative assessment and self-regulated learning: A

model and seven principles of good feedback practice. Studies in higher education, 31(2). — P. 199-218. https://doi.org/10.1080/03075070600572090 (in Eng.).

Prince M. (2004) Does active learning work? A review of the research. Journal of engineering education, 93(3). — P. 223-231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.xDigital Object Identifier (DOI) (in Eng.).

Roediger III H.L., & Karpicke J.D. (2006) Test-enhanced learning: Taking memory tests improves long-term retention. Psychological science, 17(3). — P. 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x (in Eng.).

Schneider M., & Preckel F. (2017) Variables associated with achievement in higher education: A systematic review of meta-analyses. Psychological bulletin, 143(6). — 565 p. https://doi.org/10.1037/bul0000098 (in Eng.).

Serrada-Sotil J., Huertas Martínez J.A., & Granado-Peinado M. (2025) Do audience response systems truly enhance learning and motivation in higher education? A systematic review. Humanities and Social Sciences Communications, 12(1). — 1767 p. https://doi.org/10.1057/s41599-025-06042-w (in Eng.)

Shute V.J. (2008) Focus on formative feedback. Review of educational research, 78(1). — P. 153-189. https://doi.org/10.3102/0034654307313795 (in Eng.).

Theobald E.J., Hill M.J., Tran E., Agrawal S., Arroyo E.N., Behling S., ... & Freeman S. (2020) Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. Proceedings of the National Academy of Sciences, 117(12). — P. 6476-6483. https://doi.org/10.1073/pnas.1916903117 (in Eng.).

Wang A.I., & Tahir R. (2020) The effect of using Kahoot! for learning–A literature review. Computers & Education, 149. — 103818 p. https://doi.org/10.1016/j.compedu.2020.103818 (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 287–315

https://doi.org/10.32014/2026.2518-1726.414

IRSTI 28.23.25
UDC 004.89

© Nazyrova A.E.[1*], Kaderkeyeva Z.K.[1], Bekmanova G.T.[1], Milosz M.[2], Lamasheva Zh.[1], 2026.
[1] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
[2] Lublin University of Technology, Poland.
E-mail: nazyrova_aye_1@enu.kz

TRANSFORMATION OF EDUCATION THROUGH DIGITAL TECHNOLOGIES: ADVANCING STUDENT ACADEMIC PERFORMANCE ACROSS LEARNING STAGES

Nazyrova Aizhan — PhD, senior lecturer of Institute of Digital Sciences and Artificial Intelligence, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: nazyrova_aye_1@enu.kz, https://orcid.org/0000-0002-9162-6791;
Kaderkeyeva Zulfiya — master of technical sciences, senior lecturer of Institute of Digital Sciences and Artificial Intelligence, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: kaderkeyeva_zk@enu.kz, https://orcid.org/0000-0001-8237-8549;
Bekmanova Gulmira — candidate of technical sciences, PhD, associate professor of Institute of Digital Sciences and Artificial Intelligence L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: bekmanova_gt@enu.kz, https://orcid.org/0000-0001-8554-7627;
Milosz Marek — Ph.D, professor of Lublin University of Technology, Lublin, Poland,
E-mail: m.milosz@pollub.pl, https://orcid.org/0000-0002-5898-815X;
Lamasheva Zhanar — PhD, senior lecturer of Institute of Digital Sciences and Artificial Intelligence, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: lamasheva_zhb@enu.kz, https://orcid.org/0000-0002-9535-2636.

Abstract. Academic performance is considered an indicator reflecting the quality of higher education and is associated with graduates' academic mobility and employment opportunities. This study examines differences in students' academic results across three levels of education – bachelor's, master's, and doctoral programs – within the context of the digitalization of the educational process, including the use of online platforms, electronic courses, and digital forms of assessment. The aim of the research is to determine whether differences in average grades are related to the level of study and the characteristics of the disciplines undertaken. For statistical comparison of the groups, analysis of variance (ANOVA) and Tukey's post hoc test were applied, which made it possible to identify differences between educational levels and specify which groups differed from one another.

The results indicate that students enrolled in specialized master's programs and doctoral studies demonstrate higher average grades compared to bachelor's students and those in academic-pedagogical master's programs. This distribution may be explained by differences in educational objectives and the structure of academic activities: at advanced levels, there are more research-oriented tasks, a greater proportion of independent work, and increased requirements for planning skills, source analysis, and academic writing. Additional analysis revealed the influence of course complexity on academic performance. Subjects with a strong analytical or technical component, requiring mathematical background and systematic thinking, are characterized by lower average grades. Applied digital disciplines focused on mastering tools and completing practical assignments show higher results. The findings can be used to adjust curricula, distribute academic workload, develop assessment criteria, and select support strategies for students at different levels of education.

Keywords: academic performance, educational level, digital education, higher education, ANOVA, curriculum design

Financing. *This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant №BR28713531 Intelligent digital system of higher and postgraduate education organizations Smart.EDU).*

For citations: M Nazyrova A.E., Kaderkeyeva Z.K., Bekmanova G.T., Milosz M., Lamasheva Zh. Transformation of education through digital technologies: advancing student academic performance across learning stages. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. *287–315. DOI: https://doi.org/10.32014/2026.2518-1726.414*

© Назырова А.Е.[1*], Кадеркеева З.К.[1], Бекманова Г.Т.[1], Милош М.[2], Ламашева Ж.Б.[1], 2026.

[1]Л.Н. Гумилев атындағы Еуразия Ұлттық Университеті, Астана, Қазақстан;
[2] Люблин Технологиялық Университеті, Польша.
E-mail: nazyrova_aye_1@enu.kz

ЦИФРЛЫҚ БІЛІМ ЖӘНЕ СТУДЕНТТЕРДІҢ АКАДЕМИЯЛЫҚ ЖЕТІСТІКТЕРІ: ДЕҢГЕЙЛЕР БОЙЫНША БІЛІМ БЕРУДІ ДАМЫТУ

Назырова Айжан — PhD, Цифрлық ғылымдар және жасанды интеллект институтының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, E-mail: nazyrova_aye_1@enu.kz, https://orcid.org/0000-0002-9162-6791;
Кадеркеева Зульфия — магистр, Цифрлық ғылымдар және жасанды интеллект институты аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, E-mail: kaderkeyeva_zk@enu.kz, https://orcid.org/0000-0001-8237-8549;

Бекманова Гүлмира — PhD, Цифрлық ғылымдар және жасанды интеллект институты қауымд. доценті, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: bekmanova_gt@enu.kz, https://orcid.org/0000-0001-8554-7627;
Милош Марек — PhD, Люблин технологиялық университетінің профессоры, Люблин, Польша,
E-mail: m.milosz@pollub.pl, https://orcid.org/0000-0002-5898-815X;
Ламашева Жанар — PhD, Цифрлық ғылымдар және жасанды интеллект институтының аға оқытушысы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: lamasheva_zhb@enu.kz, https://orcid.org/0000-0002-9535-2636.

Аннотация. Академиялық үлгерім жоғары білім сапасын көрсететін көрсеткіш ретінде қарастырылады және түлектердің академиялық мобильділігі мен жұмысқа орналасу мүмкіндіктерімен байланысты. Бұл зерттеуде білім берудің үш деңгейі – бакалавриат, магистратура және докторантура – бойынша студенттердің оқу нәтижелеріндегі айырмашылықтар білім беру процесін цифрландыру жағдайында талданады. Зерттеу онлайн-платформаларды, электрондық курстарды және цифрлық бағалау түрлерін қолдануды қамтиды. Зерттеудің мақсаты орташа бағалардағы айырмашылықтардың оқу деңгейіне және пәндердің сипаттамаларына байланысты екенін анықтау болып табылады. Топтарды статистикалық салыстыру үшін дисперсиялық талдау (ANOVA) және Тьюкидің апостериорлық тесті қолданылды, бұл білім деңгейлері арасындағы айырмашылықтарды анықтауға және нақты қай топтар арасында айырмашылық бар екенін көрсетуге мүмкіндік берді. Нәтижелер бейінді магистратура бағдарламалары мен докторантурада оқитын студенттердің орташа бағалары бакалаврлар мен академиялық-педагогикалық магистратура студенттеріне қарағанда жоғары екенін көрсетті. Мұндай бөлініс оқытудың мақсаттары мен оқу қызметінің құрылымындағы айырмашылықтармен түсіндіріледі: жоғары деңгейлерде зерттеу сипатындағы тапсырмалар көп, дербес жұмыс үлесі жоғары және жоспарлау, дереккөздермен жұмыс істеу, академиялық жазу дағдыларына қойылатын талаптар артады. Қосымша талдау пәндердің күрделілігінің академиялық үлгерімге әсерін анықтады. Аналитикалық немесе техникалық құрамдасы басым, математикалық дайындық пен жүйелі ойлауды талап ететін пәндер бойынша орташа балл төменірек болады. Қолданбалы цифрлық пәндер, құралдарды меңгеруге және практикалық тапсырмаларды орындауға бағытталған, жоғары нәтижелер көрсетеді. Алынған қорытындылар оқу жоспарларын түзетуге, оқу жүктемесін бөлуге, бағалау критерийлерін әзірлеуге және білім деңгейіне сәйкес студенттерді қолдау тәсілдерін таңдауға пайдаланылуы мүмкін.

Түйін сөздер: академиялық үлгерім, білім беру деңгейі, цифрлық білім беру, цифрлық білім беру, ANOVA, оқу бағдарламасын әзірлеу

© Назырова А.Е.[1*], Кадеркеева З.К.[1], Бекманова Г.Т.[1], Милош М. [2], Ламашева Ж.Б.[1], 2026.
[1]Евразийский национальный университет им. Л.Н. Гумилева,
Астана, Казахстан;
[2]Люблинский технологический университет, Польша.
E-mail: nazyrova_aye_1@enu.kz

ЦИФРОВОЕ ОБРАЗОВАНИЕ И АКАДЕМИЧЕСКАЯ УСПЕВАЕМОСТЬ УЧАЩИХСЯ: РАЗВИТИЕ ОБРАЗОВАНИЯ МЕЖДУ УРОВНЯМИ

Кадеркеева Зульфия — магистр, старший преподаватель Института цифровых наук и искусственного интеллекта, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: kaderkeyeva_zk@enu.kz, https://orcid.org/0000-0001-8237-8549;
Назырова Айжан — PhD, старший преподаватель Института цифровых наук и искусственного интеллекта, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: nazyrova_aye_1@enu.kz, https://orcid.org/0000-0002-9162-6791;
Бекманова Гульмира — PhD, ассоциированный доцент Института цифровых наук и искусственного интеллекта, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: bekmanova_gt@enu.kz, https://orcid.org/0000-0001-8554-7627;
Милош Марек — PhD, профессор Люблинского политехнического университета, Люблин, Польша,
E-mail: m.milosz@pollub.pl, https://orcid.org/0000-0002-5898-815X;
Ламашева Жанар — PhD, старший преподаватель Института цифровых наук и искусственного интеллекта, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: lamasheva_zhb@enu.kz, https://orcid.org/0000-0002-9535-2636.

Аннотация. Академическая успеваемость является важным индикатором качества подготовки в системе высшего образования и тесно связана с академической мобильностью и последующим трудоустройством выпускников. В работе анализируются различия в учебных результатах студентов на трёх уровнях высшего образования - бакалавриат, магистратура и докторантура - в условиях цифровизации образовательного процесса, включая использование онлайн-платформ, электронных курсов и цифровых форм контроля знаний. Цель исследования заключается в выявлении различий в академической успеваемости в зависимости от уровня обучения и характеристик изучаемых дисциплин. Для статистического анализа применены методы дисперсионного анализа (ANOVA) и апостериорный тест Тьюки, позволяющие определить наличие статистически значимых различий между группами и уточнить, между какими уровнями образования они проявляются. Результаты показали, что обучающиеся магистратуры (особенно профильной) и докторантуры демонстрируют более высокие средние академические показатели по сравнению со студентами бакалавриата и академико-педагогической магистратуры. Это объясняется различиями в

целях обучения и структуре образовательного процесса: на старших уровнях увеличивается доля исследовательской деятельности, самостоятельной работы и требований к академическим навыкам, включая анализ, планирование и научное письмо. Дополнительный анализ выявил влияние сложности дисциплин на уровень успеваемости. Курсы с выраженной аналитической или технической направленностью, требующие развитых математических и системных компетенций, характеризуются более низкими средними баллами. В то же время прикладные цифровые дисциплины, ориентированные на освоение инструментов и выполнение практических заданий, демонстрируют более высокие результаты. Полученные результаты могут быть использованы при совершенствовании образовательных программ, перераспределении учебной нагрузки, разработке критериев оценивания и формировании механизмов академической поддержки студентов на различных уровнях обучения.

Ключевые слова: академическая успеваемость, уровень образования, цифровое образование, высшее образование, ANOVA, разработка учебной программы

Introduction. Students' academic performance is commonly regarded as one of the central indicators of educational effectiveness in higher education. It reflects not only how successfully students meet curriculum requirements, but also how well educational institutions support learning, adaptation, and intellectual development. Academic achievement is closely connected with later professional opportunities, access to advanced study, and long-term career development. For this reason, the factors that shape student performance remain an important area of research in education, psychology, and pedagogy. Among these factors, educational level deserves particular attention, since the conditions, expectations, and forms of learning differ substantially between undergraduate, master's, and doctoral study.

At the undergraduate level, students are usually introduced to the foundations of disciplinary knowledge and are required to adapt to university learning practices, assessment systems, and increasing academic independence. In master's programs, learning often becomes more specialized, with greater emphasis on analytical work, critical interpretation of knowledge, and the application of research skills in academic or professional contexts. Doctoral education, in turn, is generally characterized by a strong focus on original research, independent inquiry, and sustained scholarly productivity. These structural differences suggest that academic performance may vary across levels of higher education not only because of differences in student ability, but also because of changing academic demands, learning environments, and forms of self-regulation.

Despite the relevance of this issue, the relationship between educational level and academic performance has not been fully clarified in the existing literature. Many studies in educational psychology examine factors such as motivation, self-

regulated learning, personality, engagement, and digital learning behavior, yet fewer studies directly compare academic outcomes across distinct stages of higher education within one analytical framework (Dent et al., 2016). As a result, there remains a gap in understanding whether differences in performance are associated with the level of study itself or with related variables that become more visible at different educational stages.

Some findings suggest that students in advanced academic programs may achieve stronger results because they have already passed earlier stages of academic selection and typically possess more developed learning strategies. Research on self-regulation and personality supports the view that students who continue into master's and doctoral study often demonstrate stronger goal-setting, planning, persistence, and metacognitive awareness, all of which can contribute to academic success (Mammadov, 2022). In this sense, higher educational levels may be associated with improved performance because students become more capable of managing complex tasks, working independently, and integrating theoretical knowledge with research-oriented practice.

At the same time, this relationship is not necessarily linear. Graduate and doctoral students also face a number of pressures that may negatively affect performance. These include a heavier reading and writing load, the need to combine study with employment, higher expectations for originality and autonomy, and increased responsibility for research outcomes. In many cases, postgraduate students must balance academic tasks with professional or family obligations, which may limit the time and energy available for coursework and assessment preparation. As some studies indicate, the complexity of graduate education may create conditions in which stronger academic maturity is offset by greater workload and stress (Berlanga et al., 2025). Therefore, the assumption that students at more advanced levels always perform better should be examined carefully rather than accepted automatically.

Another important dimension of the problem concerns the changing organization of higher education in recent years. The growing use of digital platforms, learning management systems, blended instruction, and online learning formats has transformed the way students engage with course materials, instructors, and peers. These changes have affected students differently depending on their educational level, since undergraduate, master's, and doctoral students often use digital tools for different purposes and with different degrees of independence. For example, undergraduates may rely more heavily on structured guidance, while graduate students may use digital environments more flexibly for research, collaboration, and self-directed study. This makes it necessary to examine academic performance not in isolation, but in relation to the broader learning conditions that now shape higher education (Mulyaningsih et al., 2022).

Against this background, the present study aims to determine the extent to which educational level influences academic performance among students in higher education. More specifically, the research seeks to compare academic outcomes across bachelor's, master's, and doctoral programs using a systematic

statistical approach. In addition to identifying whether differences exist, the study also aims to explore possible explanations for those differences in light of the literature on self-regulated learning, student adaptation, educational engagement, and academic workload. By focusing on educational level as a central explanatory variable, the study contributes to ongoing discussions about fairness in assessment, curriculum design, and the development of learning support systems for students at different stages of academic progression.

Objectives and Hypotheses

The objectives of the study are threefold. First, it seeks to provide a structured review of scholarly literature addressing academic performance, educational level, and the methodological approaches commonly used to analyze group differences in higher education research. Second, it aims to assemble and examine student grade data representing different stages of university education. Third, it applies one-way Analysis of Variance (ANOVA) together with Tukey's post-hoc test in order to identify whether statistically significant differences in academic performance exist among the selected groups and, if so, between which specific levels these differences occur. The statistical analysis is supported by visual representations of the data in order to improve interpretability and provide a clearer understanding of score distributions across groups.

The study is based on the following hypotheses:

H_0: Educational level does not exert a statistically significant influence on academic performance.

H_1: Educational level significantly affects academic performance, with master's and doctoral students expected to demonstrate higher academic results than undergraduate students.

These hypotheses are grounded in previous research showing that students at higher educational levels tend to possess stronger cognitive, metacognitive, and self-regulatory competencies, which are often associated with better academic outcomes (Dent et al., 2016). At the same time, the literature does not provide a fully consistent picture. While some authors emphasize the positive role of maturity, academic experience, and developed learning strategies, others point to the negative impact of workload, role conflict, and pressure in postgraduate education (Mammadov, 2022). This mixed evidence makes the present comparison both relevant and necessary.

The study also has practical significance. Understanding how academic performance differs across educational stages may help universities reconsider how they organize teaching, assessment, and academic support. If performance differences are observed, they may indicate the need for more adaptive learning strategies, differentiated instructional models, or targeted interventions for specific student groups. Such findings are especially relevant in the context of digitally mediated higher education, where student autonomy, engagement, and access to learning resources play an increasingly important role. In this sense, the research is expected to contribute not only to statistical description, but also to discussions

about how higher education institutions can respond more effectively to the needs of students at different stages of study.

The novelty of the present study lies in its integrated comparison of academic performance across three educational levels within a single methodological framework. By combining ANOVA with Tukey's post-hoc procedure and visual analysis, the study offers a structured way to identify both general and pairwise differences in student outcomes. This approach allows for a more detailed interpretation of educational-level effects and may support future work on curriculum modernization, academic advising, and the design of personalized learning environments (Berlanga et al., 2025; Mulyaningsih et al., 2022).

Literature Review. The selection of statistical techniques for analyzing students' academic performance is shaped by a number of methodological considerations widely discussed in contemporary scholarly literature. In recent years, the demand for rigorous quantitative approaches aimed at ensuring the objectivity and reliability of educational data analysis has grown substantially. This tendency is connected with the broader development of evidence-based educational research, where conclusions are expected to rest on transparent analytical procedures and clearly interpretable results. Among the most widely recognized tools are Analysis of Variance (ANOVA) and post hoc procedures such as Tukey's Honest Significant Difference (HSD) test, both of which are regarded as appropriate and reliable methods for detecting statistically significant differences across multiple comparison groups. Their methodological characteristics make them particularly suitable for assessing the influence of educational level on academic achievement (Akpen et al., 2024).

ANOVA has become a standard analytical instrument in comparative educational research due to its ability to evaluate differences in mean values across more than two groups simultaneously. This is especially relevant in studies where student populations are divided into several categories, such as undergraduate, master's, and doctoral levels. For instance, Evans and Taylor (2025) employed ANOVA to examine variations in students' academic performance while considering numerous contextual factors, including level of study (Iglesias-Pradas et al., 2021). One of the notable advantages of ANOVA is its capacity to reduce the inflation of Type I errors by eliminating the need for repeated pairwise comparisons of means. In this respect, it offers a more methodologically consistent approach than conducting several separate tests across the same dataset.

Moreover, ANOVA incorporates both between-group and within-group variance, thereby offering a more comprehensive analytical framework compared to basic mean-difference tests such as the independent-samples t-test (Witteveen et al., 2020). This feature is important in educational research, where differences in student performance may reflect not only distinctions between academic levels but also substantial variation within each level. By accounting for these two sources of variance, ANOVA allows the researcher to determine whether the observed differences in average performance are likely to reflect a meaningful group effect

rather than random fluctuation. In our research, ANOVA served as the principal statistical tool for verifying whether meaningful performance differences existed among undergraduate, master's, and doctoral students.

When ANOVA indicates statistically significant outcomes, an additional step is required to determine which specific groups differ from one another. For this purpose, Tukey's HSD test is considered one of the most effective post hoc procedures. It is widely used in educational and psychological research because it enables multiple pairwise comparisons while preserving an acceptable level of statistical control. Xu et al. (2023) demonstrate that Tukey's method provides reliable pairwise comparisons across multiple groups while maintaining a high degree of control over Type II errors and minimizing the likelihood of false-positive results. In the present study, this test enabled a more precise differentiation of academic performance among undergraduate students, master's students in both academic-pedagogical and professional programs, and doctoral candidates. This step was necessary because the identification of a general difference across groups does not, by itself, explain the structure of that difference.

Contemporary literature on educational statistics also underscores the role of graphical data visualization in improving interpretability. Statistical significance alone does not always provide a complete understanding of how the data are distributed or how strongly groups differ in practical terms. For this reason, researchers often complement formal tests with visual methods that make patterns in the data easier to observe. Mishra et al. (2025) and Juarros-Basterretxea et al. (2024) highlight the utility of boxplots, histograms, and related visual tools in illustrating data distribution patterns, including variability, skewness, and the presence of outliers. Such instruments help reveal whether group means are supported by relatively consistent observations or influenced by dispersion and extreme values. In this research, visual representations were employed not merely as supplementary illustrations but as integral components that reinforce and contextualize the results of statistical testing.

The use of visualization is particularly relevant in the analysis of academic performance, since grade distributions often vary across student groups and may contain features that are not immediately visible in summary statistics alone. For example, two groups may have similar mean values while differing substantially in terms of spread, concentration of scores, or the presence of unusually high or low results. Graphical presentation helps identify these features and supports a more careful interpretation of quantitative findings. In this way, visual analysis serves as a bridge between statistical output and substantive educational interpretation.

Overall, the synthesis of findings from recent academic publications allows us to conclude that the combined application of ANOVA, Tukey's HSD test, and complementary visualization techniques constitutes an appropriate methodological framework for analyzing differences in academic performance across educational levels. This integrated approach provides statistically reliable results, supports a clearer understanding of performance patterns, and offers useful implications for

the ongoing refinement of higher education curricula and assessment practices. It also makes it possible to connect quantitative findings with broader discussions about student progression, academic demands, and the conditions that shape learning outcomes at different stages of higher education.

Materials and main methods. The empirical basis of the study consisted of a dataset containing detailed information on students' educational levels, the academic disciplines they completed, and their final grades. The sample encompassed four distinct groups of learners: undergraduate students, master's students enrolled in academic-pedagogical programs, master's students pursuing professional tracks, and doctoral candidates (PhD programs). This group structure made it possible to compare academic performance across several stages of higher education within a single analytical framework. The primary objective of the statistical analysis was to identify systematic differences in academic performance across these educational categories and to determine whether such differences were statistically meaningful.

The dataset was organized in a way that allowed the examination of both general and subject-specific patterns of performance. Educational level was treated as the main independent grouping variable, while the final academic grade served as the dependent variable. In addition, the academic discipline completed by the student was used as an auxiliary analytical factor in order to explore whether differences in performance were related not only to the level of study, but also to the characteristics of particular courses. This made it possible to consider academic achievement from two perspectives: first, as a function of educational level, and second, as a function of subject-related complexity.

Only completed course records with available final grades were included in the analysis. Observations with missing or incomplete grade information were excluded in order to preserve the comparability of the groups and avoid distortion in the statistical calculations. This procedure helped ensure that all analyzed records reflected completed academic performance and could be interpreted within the same grading framework. The use of cleaned and comparable data was especially important because the study involved several educational levels and a wide range of subjects with different instructional formats.

As an initial step, descriptive statistics were computed, including the calculation of mean final grades for each category of students. This provided a preliminary indication of performance variation across educational levels and made it possible to identify broad patterns before applying inferential procedures. Descriptive analysis was also used to summarize average grades across individual disciplines, which later served as the basis for identifying relatively more difficult and less difficult subjects. In this way, the first stage of analysis established the empirical background for subsequent statistical testing.

Following this, Analysis of Variance (ANOVA) was employed to test for the presence of overall group differences. ANOVA was selected due to its suitability for comparing more than two independent groups and its ability to partition variance

into between-group and within-group components. This feature is particularly relevant for educational research, where differences in performance may exist not only between educational levels but also within each student group. By comparing these two sources of variation, ANOVA makes it possible to assess whether the observed differences in mean academic performance are likely to reflect a real group effect rather than random fluctuation.

Before conducting the ANOVA, the main assumptions underlying this procedure were taken into account. These included the independence of observations, approximate normality of grade distributions within groups, and the relative homogeneity of variances. Although ANOVA is generally considered robust to moderate deviations from normality, attention to these assumptions was necessary in order to support the validity of the analysis and strengthen confidence in the interpretation of results. The level of statistical significance was set at $p < 0.05$.

Upon detection of statistically significant differences, Tukey's post hoc test was applied to conduct pairwise comparisons among the educational levels. This procedure enabled the identification of specific group pairs in which performance disparities were statistically significant, thereby offering a more granular understanding of the academic differences observed among undergraduates, master's students of different orientations, and doctoral students. While ANOVA determines whether at least one significant difference exists among the means, Tukey's test clarifies where exactly these differences are located. This was necessary for the present study, since the comparison involved more than two categories and required a detailed interpretation of intergroup relationships.

To enhance interpretability and ensure clarity of findings, graphical visualization techniques were incorporated. Histograms and boxplots were used to display the distribution of final grades across all categories, illustrating central tendencies, variability, and potential outliers. These visual tools supported the statistical analysis by showing whether the mean values were accompanied by concentrated or dispersed score distributions. In the context of educational data, such visualizations are particularly useful because they make it easier to identify internal heterogeneity within student groups and to assess whether some courses or educational levels are characterized by greater inconsistency in outcomes than others.

The integration of descriptive statistics, ANOVA, Tukey's post hoc test, and graphical visualization provided a comprehensive analytical framework for examining the relationship between educational level and academic performance. This approach allowed the study not only to determine whether performance differences existed, but also to interpret how these differences were distributed across student categories and subjects. As a result, the methodology supported both statistical rigor and educational interpretability.

Mean Grade by Educational Level

As part of this research, the final academic grades of students enrolled in different levels of higher education were examined. Table 1 presents the mean scores for undergraduate students, master's students from both academic-pedagogical and professional tracks, and doctoral (PhD) candidates.

Table 1 – Mean Final Grade of Students by Educational Level

Educational Level	Mean Grade
Undergraduate (Bachelor's program)	74.8
Master's (Scientific & Pedagogical)	75.1
Master's (Specialized / Professional)	84.7
Doctoral Studies (PhD program)	85.1

The descriptive results presented in Table 1 indicate that average academic performance differs across the four educational categories. The mean grades of bachelor's students and students enrolled in the scientific-pedagogical master's program are very close, suggesting that these two groups demonstrate broadly similar levels of academic achievement. In contrast, students in the specialized or professional master's track and doctoral candidates show visibly higher average scores. This pattern suggests that academic outcomes are associated not only with progression to a higher formal level of education, but also with the orientation and internal structure of the academic program.

The observed differences may reflect several educational factors. Students in specialized master's and doctoral programs often work in more focused academic environments, where narrower specialization, clearer professional or research goals, and stronger academic self-regulation contribute to better performance. At the same time, the similarity between bachelor's students and students in the scientific-pedagogical master's track may indicate that these two groups share certain structural features in terms of curriculum organization, instructional logic, or assessment format. Thus, even at the level of descriptive statistics, the results point to the need for a differentiated interpretation of graduate education rather than treating all master's programs as a single category. The distribution of these mean values is illustrated in Figure 1.

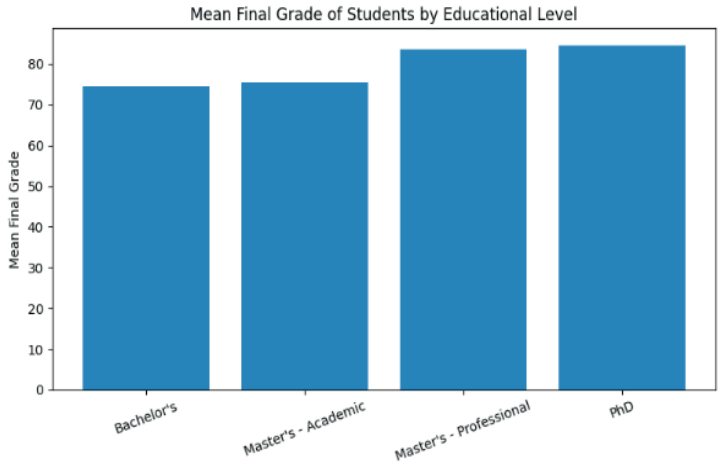


Figure 1 – Average final score of students by education level.

An analysis of variance (ANOVA) was then performed to determine whether the observed differences among the educational level groups were statistically meaningful. The resulting p-value ($p < 0.001$) indicates that academic performance varies significantly across the four educational categories. In other words, the differences in mean grades shown in the descriptive statistics are unlikely to be random and may be interpreted as reflecting a meaningful association between educational level and students' academic outcomes.

At the same time, the ANOVA result alone does not indicate which specific groups differ from one another. A statistically significant overall test only shows that at least one intergroup difference exists. For this reason, a post hoc comparison procedure was required in order to determine where the significant differences were located. Tukey's post hoc multiple comparison test was therefore applied as the next analytical step. The outcomes of this procedure are summarized in Table 2.

Table 2 – Tukey Post Hoc Multiple Comparison Test Results

Group Pair	Difference in Means	Adjusted p-value
Bachelor's vs PhD	10.25	< 0.001
Bachelor's vs Master's (Scientific–Pedagogical)	0.24	0.945
Bachelor's vs Master's (Specialized)	9.84	< 0.001
PhD vs Master's (Scientific–Pedagogical)	–10.00	< 0.001

The findings from Tukey's test confirm that educational level exerts a statistically significant influence on students' academic performance. The most substantial difference is observed between bachelor's students and doctoral candidates, with doctoral students achieving markedly higher final grades. This result may be associated with the selective nature of doctoral admissions, the narrower specialization of PhD study, and the research-centered character of doctoral education. These features typically require and reinforce a high degree of academic independence, sustained motivation, and analytical maturity, which may contribute to better academic outcomes.

The comparison between bachelor's students and students enrolled in the scientific-pedagogical master's track reveals no statistically significant difference. This finding is important because it suggests that the scientific-pedagogical master's program may remain relatively close to the undergraduate level in terms of curriculum demands, assessment logic, or general learning expectations. Such a result may indicate that the transition from bachelor's study to this particular master's track does not involve a strong enough shift toward advanced theoretical preparation or research-based learning. From a curriculum perspective, this may justify reconsidering the structure of the program in order to strengthen its academic distinctiveness.

A different pattern is observed in the case of the specialized master's program. Students in this category significantly outperform bachelor's students, which may be explained by the more applied and professionally oriented character of the

program. It is also possible that this group includes students with stronger prior preparation, clearer career goals, or relevant practical experience. These factors may positively influence their engagement with course content and contribute to more stable and higher academic performance.

Moreover, the absence of a strong distinction between the specialized master's track and doctoral studies suggests that both groups occupy a relatively high-performance segment within the dataset. Although doctoral education differs from professional master's study in depth and research orientation, both groups may share common characteristics such as stronger self-regulation, clearer academic purpose, and greater capacity for independent work. This does not imply that the programs are equivalent in content, but it does indicate that their students reach similarly high levels of measured academic achievement.

Overall, the Tukey analysis shows that the most visible performance differences emerge between undergraduate education and the more advanced, specialized forms of graduate study. At the same time, the findings point to internal diversity within master's education itself. This suggests that educational level should not be treated as a purely formal category; the substantive orientation of a program also appears to matter for academic outcomes. These results support the need for a more differentiated view of graduate education and may be useful for improving program design, especially in tracks where performance remains close to the undergraduate level.

Analysis of Subject Complexity

The analysis of differences in students' academic performance across educational levels demonstrated that average grades are influenced not only by the level of education but also by the characteristics and complexity of the disciplines studied. To obtain a deeper understanding of how subject-specific factors affect academic outcomes, an additional examination of course difficulty was conducted based on students' average final grades.

For the purpose of analysis, all subjects were grouped according to their relative complexity. Two categories were identified:

(1) disciplines with the lowest average scores, interpreted as the most challenging (Table 3), and

(2) disciplines with the highest average scores, considered less difficult for students (Table 4).

It should be noted that, in the present study, subject complexity was not measured directly through an external rating scale or expert judgment. Instead, it was interpreted indirectly on the basis of students' average academic results. Courses with lower average grades were treated as relatively more difficult, while courses with higher average grades were considered comparatively easier for the student groups represented in the dataset. This operational approach does not claim to capture every dimension of course difficulty, but it provides a practical basis for comparing patterns of subject-specific performance within the available empirical material.

At the same time, average grade should not be interpreted as a perfect or exclusive indicator of objective course complexity. Academic results in a given subject may also be shaped by assessment design, teaching methods, students' prior preparation, grading practices, or the balance between theoretical and practical tasks. Therefore, the interpretation of low average grades as evidence of higher difficulty should be made with caution and in relation to the broader instructional context. This reservation is important because some courses may appear difficult not only because of their content, but also because of the way learning outcomes are assessed.

The findings show that courses related to programming, engineering technologies, and database management exhibit relatively high average grades in some cases, while other analytically demanding technical subjects demonstrate much lower averages. This pattern suggests that technical disciplines are not uniform in their effect on academic performance. Applied and practice-oriented courses may support stronger outcomes when students can rely on hands-on tasks, concrete problem-solving procedures, and clearer performance criteria. By contrast, highly theoretical, algorithmic, or systems-oriented subjects may create more difficulties because they require stronger abstract reasoning, deeper conceptual understanding, and more extensive independent preparation.

In contrast, subjects with low average grades warrant closer attention. Their complexity may arise from the inherent cognitive demands of the course content, insufficient foundational knowledge among students, or shortcomings in pedagogical approaches, instructional materials, or assessment design. In some cases, the low results may reflect the fact that students encounter these disciplines before they have fully mastered prerequisite concepts. In other cases, the difficulty may be associated with fragmented course structure or a mismatch between teaching methods and assessment requirements. For this reason, the interpretation of difficult subjects should take into account not only the nature of the content itself but also the broader educational conditions in which the course is delivered.

Future research will aim to examine the variance and distribution of grades for each discipline in order to pinpoint subjects with the largest performance disparities. Such analysis will make it possible to determine whether variability arises from structural curriculum issues, inconsistent assessment practices, or differences in instructional quality. These insights may serve as a basis for refining course design and improving support strategies for students encountering difficulties in specific subject areas.

Table 3 – The most difficult subjects by average final score

№	Discipline	Level	Avg. Score
1	Introduction to Distributed Systems	Bachelor	5.2
2	Machine Learning	Bachelor	7.8
3	Programming Languages 2	Bachelor	8.5

4	Fundamental Algorithms	PhD	12.4
5	System Software	Bachelor	34.7
6	Microcontroller Programming	Bachelor	36.1
7	Distributed Intelligent Systems	PhD	38.9
8	Fundamentals of Cybersecurity	Bachelor	40.3
9	Mobile Sensor Data Processing	Bachelor	42.6
10	IT Infrastructure	Bachelor	47.5

Table 4 – The easiest subjects by average final score

№	Discipline	Level	Avg. Score
1	Database	Bachelor	99.6
2	Database Management Tools	Bachelor	99.8
3	Internet of Things	Bachelor	100.0
4	Computer Engineering	Bachelor	98.9
5	Microprocessor & Embedded Systems	Bachelor	98.7
6	Artificial Intelligence	Master (prof.)	98.5
7	Higher Mathematics	Master (prof.)	97.2
8	Theoretical Mechanics	PhD	96.8
9	Numerical Methods	Master (sci.)	96.5
10	Data Analysis	Bachelor	95.9

The analysis of disciplines with the highest average final scores revealed a group of subjects in which students demonstrated consistently strong academic performance. The top-performing courses were predominantly concentrated within the bachelor's curriculum, including Database, Database Management Tools, and Internet of Things, each showing average scores close to or equal to 100. Such results suggest that these subjects are either well aligned with students' prior preparation or characterized by clearly structured content, practice-oriented learning tasks, and assessment methods that allow students to perform with a high degree of accuracy.

Additionally, subjects such as Computer Engineering and Microprocessor & Embedded Systems also showed high performance, with average scores above 98. These disciplines typically involve applied technical skills and hands-on laboratory activities, which may contribute to higher levels of student engagement and improved mastery of course material. Where learning tasks are concrete, iterative, and supported by practical feedback, students may find it easier to demonstrate competence in assessment settings.

Among graduate programs, the professional master's disciplines Artificial Intelligence (98.5) and Higher Mathematics (97.2) demonstrated strong results as well. This suggests that students in professional-oriented tracks possess sufficient prior knowledge and motivation, enabling them to perform successfully in

advanced, specialized subjects. It may also indicate that these programs are more tightly aligned with students' academic interests and professional goals, which can positively affect learning persistence and final outcomes.

At the doctoral level, Theoretical Mechanics showed an average score of 96.8, indicating that even highly theoretical and conceptually complex subjects may result in high academic performance when taken by well-prepared and research-oriented students. Similarly, in the scientific master's program, the course Numerical Methods showed strong outcomes (96.5), reflecting students' ability to work with mathematical models and computational approaches. These findings suggest that high performance in difficult subjects is possible when students possess adequate background knowledge and when course expectations are clearly structured.

Overall, the analysis indicates that subjects with high average grades tend to share several characteristics: relatively well-organized curricula, applied or practice-intensive learning tasks, clear links between instruction and assessment, and student cohorts with strong initial preparation and sustained motivation. These findings draw attention to the role of course design and assessment clarity in shaping academic outcomes and suggest that successful performance often depends not only on the content of the subject, but also on how learning is organized and evaluated.

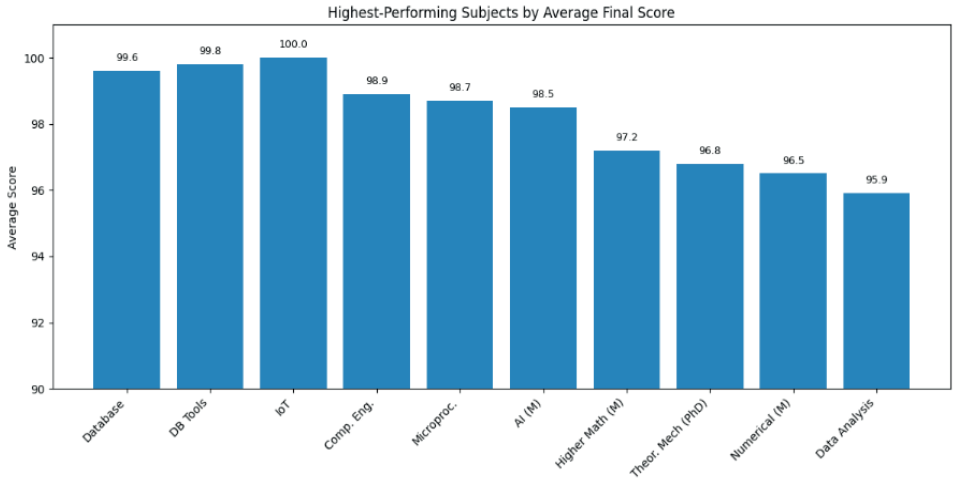


Figure 2 – Average final score of students by education level.

Figure 3 presents the distribution of average student scores across various subjects. The difficult subjects (highlighted in red), such as Machine Learning, Fundamental Algorithms, Microcontroller Programming, and Fundamentals of Cybersecurity, demonstrate comparatively low performance levels, with average scores ranging from 0 to 40 points. These results indicate that such disciplines involve a high degree of conceptual and cognitive complexity for the students represented in the dataset.

In contrast, the less difficult subjects (shown in green), including Database, Internet of Things, Artificial Intelligence, and Higher Mathematics, are characterized by very high student achievement, with average scores of approximately 98–100 points. This suggests that students perform more confidently in these courses, possibly because the material is delivered in a more structured way, because the assessment procedures are clearer, or because the learners have stronger prior preparation in these areas.

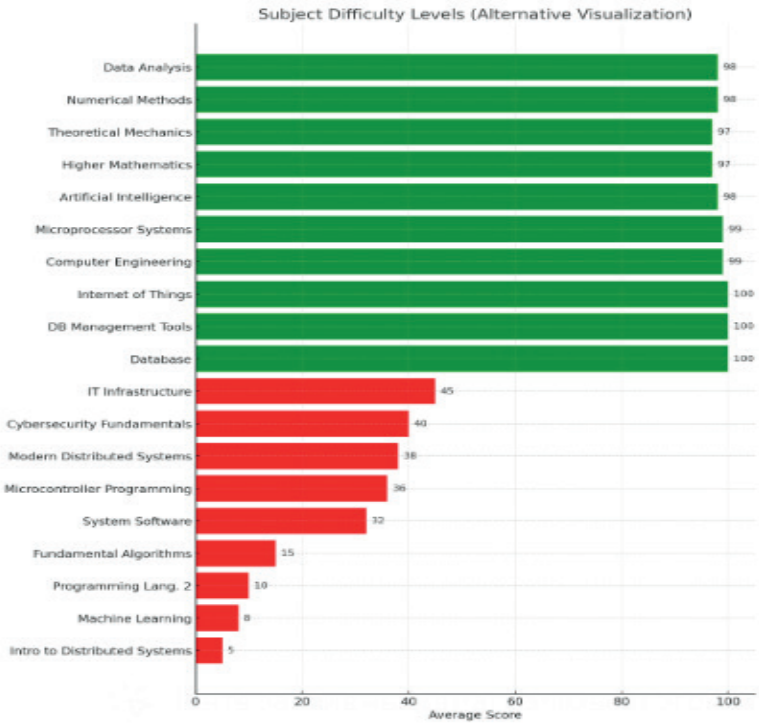


Figure 3 – Distribution of the most difficult and easiest subjects by mean score.

A comparison between the lowest-performing and highest-performing disciplines reveals a visible contrast in the character of academic outcomes. The subjects classified as more difficult are largely associated with algorithmic thinking, abstract reasoning, systems-level understanding, and technically demanding problem-solving. These courses often require students to integrate several layers of knowledge at once, apply theoretical principles to unfamiliar situations, and work with complex computational or engineering concepts. Such demands may explain why average scores in these disciplines are substantially lower.

By contrast, the courses with the highest average scores appear to be more favorable in terms of assessment outcomes. In some cases, this may reflect the practical orientation of the subject, where students can demonstrate their knowledge through structured tasks, technical exercises, or hands-on projects. In other cases,

the high average scores may indicate better alignment between teaching methods and expected learning outcomes. It is also possible that these courses attract or are taken by students who already possess a stronger academic foundation in the relevant field.

Another important aspect of subject-specific analysis concerns the internal variability of grades. A low mean score does not necessarily imply that all students performed poorly; rather, it may indicate that the subject produces a wider dispersion of outcomes, including both very weak and relatively strong results. Similarly, a high average score may conceal meaningful variation if a smaller subgroup of students performs below the majority. For this reason, the interpretation of subject difficulty should take into account not only average values but also the spread of grades, the concentration of scores around the mean, and the presence of outliers. This is why graphical tools such as boxplots are particularly useful in the present study.

The analysis of subject-specific performance demonstrates that students experience considerable challenges when studying complex technical disciplines, while applied and practice-oriented subjects are associated with notably higher academic outcomes. This contrast suggests that the cognitive load, theoretical depth, and algorithmic complexity inherent to certain technical courses significantly affect students' ability to achieve high scores. These findings indicate that some disciplines require a greater amount of independent study, stronger foundational knowledge, and more sustained academic effort than others.

From a practical perspective, the results are relevant for curriculum design and assessment policy. Subjects associated with lower average grades and wider dispersion may require additional instructional support, revised learning materials, or more transparent evaluation criteria. At the same time, the strong results observed in some applied disciplines suggest that well-structured courses with clearly communicated expectations can support consistently positive learning outcomes. A deeper analysis of these differences may help universities identify where academic difficulties are most concentrated and how teaching and assessment practices can be adjusted to support students more effectively.

In this regard, further examination of grade variability across disciplines will be particularly important. Identifying subjects with the widest dispersion of scores will make it possible to determine which courses are most challenging for learners in practice, not only in terms of average performance but also in terms of inequality of outcomes. High variability may indicate differences in students' initial preparation, methodological shortcomings, or instructional formats that require refinement. Such analysis can provide a basis for improving teaching methods, assessment strategies, and curriculum structure in order to make student achievement more stable and more comparable across disciplines.

Results. The analysis shows that educational level has a statistically significant effect on academic performance. Doctoral students and those enrolled in specialized (professional) master's programs demonstrate higher average grades

than bachelor's students and students in scientific–pedagogical master's tracks. This pattern suggests that academic achievement is related not only to the formal stage of higher education, but also to the orientation of the program and the learning conditions associated with it. In many cases, students who continue into professional master's or doctoral study have already passed through a process of academic selection and may therefore represent groups with stronger subject preparation, clearer educational goals, and greater readiness for advanced study.

These differences may be linked to stricter selection procedures, narrower specialization, higher academic motivation, and stronger research skills. Students at more advanced stages of education are often expected to manage complex tasks independently, work with specialized literature, and engage in analytical or research-based activities with less external guidance. Such demands may gradually strengthen their ability to organize learning, maintain focus over longer periods, and apply knowledge in more structured and purposeful ways. As a result, higher average performance among doctoral students and students in specialized master's programs may reflect the cumulative effect of previous academic experience together with the more focused nature of their studies.

At the same time, the lack of significant differences between specialized master's and doctoral programs suggests that performance at these stages may be close to an upper limit within the given assessment system. This does not mean that the two groups are identical in terms of academic content or educational purpose. Rather, it indicates that both groups achieve similarly high measured outcomes, despite the fact that doctoral study usually involves a stronger research orientation and a higher level of academic independence. One possible explanation is that the grading system used in the analyzed context may not fully capture qualitative differences between these advanced groups, particularly when most students already perform at a consistently high level.

The similarity in results between bachelor's students and scientific–pedagogical master's students points to possible weaknesses in how the scientific–pedagogical track is designed and assessed. This finding is important because a master's program would generally be expected to demonstrate a clearer shift toward greater academic complexity, stronger theoretical grounding, and deeper engagement with research-related tasks. If students in this track perform at nearly the same level as undergraduates, it may indicate that the curriculum does not provide enough differentiation in rigor or research training compared to the bachelor's level. It may also suggest that learning outcomes, instructional methods, and assessment procedures are not yet sufficiently aligned with the expectations typically associated with graduate education.

Another possible interpretation is that the scientific–pedagogical master's track may function more as a continuation of previous study than as a qualitatively distinct academic stage. In this case, the transition from bachelor's to master's education may remain limited in practical terms, especially if students continue to work within familiar formats of coursework and evaluation. This would reduce

the likelihood of substantial differences in average grades and may partly explain the statistical similarity between these groups. From the perspective of educational design, such a result points to the need for closer examination of curriculum content, progression logic, and the balance between pedagogical preparation and research-oriented learning.

Subject complexity also matters. Courses that require strong analytical and algorithmic thinking, such as Machine Learning, Fundamental Algorithms, and Microcontroller Programming, tend to have lower average scores, suggesting higher cognitive load and greater conceptual difficulty. These disciplines often require students to work with abstract models, sequential reasoning, and technically demanding problem-solving procedures. In addition, they may presuppose a level of prior preparation that not all students possess to the same degree. As a result, such subjects can produce lower average grades and greater variation in performance, especially when students differ in foundational knowledge or in their ability to adapt to mathematically or computationally complex material.

In contrast, database and internet-technology subjects show higher average grades, which may reflect clearer practical tasks, less abstract content, or more accessible assessment formats. In applied disciplines, students often benefit from direct links between theory and practice, more visible problem structures, and a stronger sense of how course tasks relate to real technical applications. This may make learning more manageable and allow students to demonstrate achievement more confidently during assessment. Higher grades in such subjects do not necessarily imply that they are simple, but they may indicate that the organization of instruction and evaluation is more transparent or better matched to student expectations.

Even in high-performing subjects, grade variation remains noticeable, indicating differences in prior preparation and/or limited standardization in assessment across groups. This point is important because high average performance alone does not guarantee consistency in student achievement. A course may have a strong overall average while still containing subgroups of students who perform substantially better or worse than others. This internal variation can reflect unequal entry-level competencies, different patterns of engagement, or differences in access to academic support. It may also indicate that some assessment formats allow stronger students to excel while leaving weaker students with fewer opportunities to recover or demonstrate partial understanding.

High dispersion is especially visible in programming and data-analysis disciplines, and outliers in the form of very low or very high scores may signal issues in course organization, uneven teaching quality, or gaps in assessment methods. In some cases, outliers may reflect genuinely exceptional performance or substantial learning difficulties. In other cases, they may point to inconsistencies in grading criteria, unequal instructional support across groups, or differences in how assignments and examinations are structured. For this reason, variability in grades should be treated as a meaningful analytical result rather than as a secondary

feature of the data. A closer examination of dispersion can help identify courses where academic outcomes are less stable and where improvements in teaching or evaluation may be needed.

Overall, the results indicate that academic performance in higher education is shaped by a combination of educational level, program orientation, and subject-specific demands. The findings suggest that higher achievement is more common in advanced and specialized programs, while lower and more variable performance is observed in courses with stronger analytical and technical complexity. At the same time, the results show that curriculum structure and assessment design remain important factors in understanding why some groups and disciplines produce more favorable outcomes than others.

Discussion. The overall pattern supports the idea that higher educational levels are associated with stronger learning regulation, deeper specialization, and more developed analytical competencies, which helps explain why PhD students and specialized master's students perform better (Dent & Koenka, 2016; Broadbent & Poon, 2015; Milkova et al., 2024). Prior research has repeatedly shown that academic success is connected with self-regulated learning, planning, monitoring of progress, and the ability to manage complex tasks over time. These qualities tend to become more visible as students move into advanced stages of education, where independent study and sustained academic effort are required more consistently. In this sense, the present findings are in line with the broader view that educational progression is often accompanied by the development of stronger academic habits and more effective learning strategies, especially under conditions of digital and technology-supported learning (Mena-Guacas et al., 2025).

Related work also connects academic progression with metacognitive and research-oriented skill development (Mammadov, 2022). Students who continue into specialized master's and doctoral programs are often better equipped to evaluate their own learning, identify gaps in understanding, and adjust their study strategies when working with difficult material. These forms of metacognitive control may contribute to more stable and higher performance, especially in programs where academic expectations are clearly defined and closely linked to professional or research goals. The results of the present study support this interpretation by showing that more advanced and specialized educational tracks are associated with stronger average outcomes.

The absence of significant differences between specialized master's and doctoral students can be interpreted as a sign that both groups already demonstrate stable high performance once key competencies are formed (Iglesias-Pradas et al., 2021). This pattern may suggest that, beyond a certain stage of academic development, differences in measurable performance become less pronounced because both groups function at a similarly high level within the grading system. It is also possible that the assessment format used in institutional practice is better suited to distinguishing between lower and middle levels of achievement than between already high-performing graduate groups. Therefore, the similarity

between specialized master's and doctoral students should not necessarily be understood as evidence of equivalent academic depth, but rather as an indication of comparable performance within the available measurement framework.

At the same time, the fact that bachelor's and scientific–pedagogical master's students show similar outcomes suggests that the scientific–pedagogical master's curriculum may not sufficiently increase complexity, research depth, or academic rigor. This is one of the more important findings of the study because it raises questions about the internal differentiation of graduate education. If a master's track does not produce a clear shift in academic performance compared with undergraduate study, this may point to limits in curriculum progression, insufficient emphasis on independent inquiry, or a continued reliance on instructional and assessment formats that remain close to those used at the bachelor's level. In the context of current higher education transformation, this issue also relates to the broader need to align program design with emerging technological and pedagogical demands (Mena-Guacas et al., 2025; Zhang et al., 2025).

Studies on program competitiveness also emphasize curriculum redesign and stronger specialization as factors that improve outcomes (Berlanga & Corti, 2025). In this context, the present results may be interpreted as support for revising the scientific–pedagogical master's track in order to strengthen its academic identity. Such revision could involve deeper theoretical preparation, a larger research component, clearer connections between coursework and scholarly practice, and more explicit expectations regarding independent analytical work. These changes may help the program become more distinct from undergraduate education and may contribute to stronger academic results in future cohorts. Recent studies on digital transition in higher education also suggest that curriculum redesign should increasingly account for AI-assisted and technology-enhanced teaching models (Zhang et al., 2025).

Differences across subjects are consistent with research showing that algorithmic and abstract courses often create higher cognitive load and lead to lower averages and larger variability (Mulyaningsih et al., 2022). The present findings fit this pattern. Courses that depend heavily on formal reasoning, mathematical abstraction, or programming logic tend to require sustained concentration and a strong command of prerequisite knowledge. When these foundations are uneven, academic performance is likely to become more dispersed. This helps explain why certain technical subjects in the dataset combine lower average grades with wider variation in results. At the same time, recent work on AI-assisted engineering teaching suggests that the integration of digital instructional frameworks may help reduce some of these difficulties by improving the structure and support of technically complex learning environments (Zhang et al., 2025).

The observed grade dispersion in technical disciplines also fits discussions in STEM assessment research. Variability can reflect uneven starting skill levels, instructor differences, and inconsistent grading, especially in practical work (Xu et al., 2023; Witteveen & Attewell, 2020; Tempelaar et al., 2015; Conijn et al.,

2017). In applied and technical courses, assessment often includes project-based tasks, laboratory assignments, coding activities, or open-ended problem solving. These formats can provide valuable evidence of student competence, but they may also introduce greater variation if grading criteria are interpreted differently across instructors or if students receive unequal levels of methodological support. As a result, dispersion in grades may reflect not only student ability but also differences in course implementation. This point is also consistent with recent findings that digital competence and ICT-supported instruction can influence academic outcomes by shaping how students engage with course requirements and learning tasks (Milkova et al., 2024).

Another point worth considering is that high variation in grades may indicate that some subjects are functioning as points of academic stratification within the curriculum. In such cases, technically demanding courses do not simply measure knowledge; they also separate students according to prior preparation, adaptability, and persistence. This may be educationally meaningful, but it also creates a need for more transparent support systems and more clearly structured assessment practices. Where very low and very high results coexist within the same discipline, the issue may lie partly in student heterogeneity, but it may also suggest that course design could be improved to make expectations more accessible and learning trajectories more consistent. Broader reviews of educational transformation through emerging technologies likewise underline the importance of redesigning learning environments so that they support different student groups more effectively (Mena-Guacas et al., 2025).

From a broader perspective, the discussion of educational level and subject complexity suggests that academic performance should not be interpreted as the product of one factor alone. It emerges through the interaction of student characteristics, curriculum structure, instructional methods, and assessment design. Higher educational levels may support stronger achievement because they attract and develop students with greater academic maturity, but this effect is also shaped by the kinds of courses students take and by the way institutions organize learning. In this sense, the present findings support a multidimensional understanding of academic performance in higher education, including the role of digital transition, technological support, and evolving teaching models (Milkova et al., 2024; Zhang et al., 2025).

Taken together, the discussion indicates that stronger performance among doctoral and specialized master's students is consistent with existing literature on self-regulation, specialization, and advanced academic skill development. At the same time, the similarity between bachelor's and scientific–pedagogical master's students points to a possible need for curriculum revision in order to strengthen the distinctiveness and rigor of that track. The subject-level findings further suggest that technical and analytically demanding courses require particular attention in terms of teaching support, grading transparency, and instructional design. These conclusions may be useful for universities seeking to improve both program

structure and the comparability of academic outcomes across student groups, especially in the context of ongoing educational transformation driven by emerging technologies (Mena-Guacas et al., 2025; Zhang et al., 2025).

Limitations of the Study. Despite the analytical insights provided by the present study, several limitations should be acknowledged when interpreting the results.

First, the empirical data used in this research were derived from a single institutional context. As a result, the observed patterns in academic performance may reflect the specific characteristics of the university, its grading culture, curriculum structure, and student population. Consequently, the findings should not be generalized to all higher education institutions without caution. Future studies involving multiple universities or international datasets could provide a broader comparative perspective.

Second, the relatively high academic performance observed among doctoral students may partly reflect the effect of academic selection. Admission to doctoral programs typically involves rigorous evaluation of candidates' academic background, research potential, and prior achievements. Therefore, doctoral students may represent a group that has already undergone substantial academic filtering. This factor may contribute to higher average grades and should be considered when interpreting differences between educational levels.

Third, the study relies on final course grades as the primary indicator of academic performance. Although grades are widely used as a measurable outcome in educational research, they may not fully capture all aspects of learning. Differences in grading practices, instructor expectations, and assessment formats across courses may affect the comparability of scores. In addition, some disciplines rely more heavily on project-based or practical assignments, while others emphasize theoretical examinations, which may influence the distribution of grades.

Another limitation concerns the interpretation of subject complexity. In this study, course difficulty was inferred from average student performance rather than from an external evaluation of curriculum difficulty or instructional design. While this approach provides useful insights into patterns of student achievement, it does not necessarily reflect the intrinsic difficulty of the subject itself.

Future research may address these limitations by incorporating additional variables such as prior academic preparation, student engagement, learning analytics data from digital platforms, or qualitative analysis of course design and teaching methods.

Recommendations for Improving Educational Programs

Based on the findings of the statistical analysis, several recommendations can be proposed for improving curriculum design and instructional practices across different levels of higher education.

The results indicate that academic performance among students in scientific–pedagogical master's programs remains relatively close to the bachelor's level.

This suggests that the curriculum of this track may require stronger differentiation in terms of academic rigor and research orientation. One possible improvement is the earlier integration of research-oriented learning activities, including project-based research tasks, methodological training, and participation in supervised research projects.

Another important aspect concerns technically complex disciplines that demonstrate lower average academic performance. Courses such as Machine Learning, Distributed Systems, and Microcontroller Programming require strong analytical thinking and substantial prior preparation. To support student learning in these areas, instructors may consider incorporating additional learning support mechanisms, including step-by-step problem-solving sessions, interactive digital learning tools, and adaptive learning platforms that allow students to practice complex concepts at their own pace.

Improving transparency and consistency in assessment procedures is also essential. Greater standardization of grading criteria and clearer communication of learning expectations may reduce grade dispersion and improve comparability of academic outcomes across groups. This is particularly relevant for technical disciplines where project-based evaluation and programming assignments may introduce variability in grading practices.

Finally, curriculum planners may benefit from regularly analyzing academic performance data in order to identify subjects with unusually high variability or consistently low results. Such monitoring can help institutions detect potential weaknesses in course design and provide targeted academic support to students encountering difficulties.

Table 5 – Suggested Curriculum Improvement Measures

Educational Area	Identified Problem	Proposed Improvement Strategy	Expected Academic Outcome
Scientific–pedagogical master's programs	Academic performance close to bachelor's level, limited differentiation in curriculum rigor	Introduce research-based learning modules, expand methodological training, integrate supervised research projects and thesis preparation earlier in the curriculum	Stronger analytical skills, improved research competence, clearer distinction between undergraduate and graduate academic levels
Analytical and algorithmic technical disciplines (e.g., Machine Learning, Algorithms)	Lower average scores and higher grade dispersion due to cognitive complexity	Implement scaffolded instruction, increase problem-based learning sessions, provide additional digital learning resources and coding practice environments	Reduced learning difficulties, improved conceptual understanding, more stable academic performance
Assessment and grading practices	Grade variability across courses and instructors	Develop standardized grading rubrics, clarify evaluation criteria, and implement transparent assessment frameworks across departments	Greater comparability of grades and improved fairness of academic evaluation

Academic performance monitoring	Lack of systematic analysis of student performance across disciplines	Establish institutional learning analytics systems to track grade patterns, identify high-risk subjects, and monitor curriculum effectiveness	Data-driven curriculum improvement and earlier identification of learning difficulties
Student academic support	Uneven student preparation in technically complex courses	Provide preparatory modules, tutoring programs, and adaptive digital learning systems for foundational subjects	Increased student readiness and improved academic outcomes in complex disciplines

The recommendations presented in Table 5 summarize the main practical implications derived from the statistical analysis conducted in this study. The results indicate that differences in academic performance across educational levels are not determined solely by student ability, but are also influenced by curriculum structure, instructional approaches, and assessment practices. Therefore, improving educational programs requires a systematic approach that combines curriculum redesign, enhanced instructional support, and data-informed monitoring of academic outcomes.

In particular, the findings highlight the need to strengthen the scientific–pedagogical master's curriculum so that it more clearly differs from the bachelor's level in terms of research training and analytical depth. Introducing research-based learning activities and methodological courses may help students develop stronger academic competencies and better prepare them for advanced academic or professional work. At the same time, the results emphasize the importance of additional instructional support in technically demanding disciplines, where lower average scores and higher variability suggest that students face greater cognitive challenges.

The implementation of standardized assessment frameworks and learning analytics systems may also contribute to improving transparency and consistency in grading practices. By systematically analyzing academic performance data, universities can identify disciplines with persistent learning difficulties and implement targeted improvements in teaching methods and curriculum design. In this way, the integration of empirical evidence into educational planning can support more effective learning environments and promote sustainable improvement in student academic outcomes.

Conclusion. At the same time, the results also indicate that the scientific–pedagogical master's track remains relatively close to the bachelor's level in terms of performance. This finding deserves particular attention, since a master's program would normally be expected to reflect a higher level of academic complexity and a clearer transition toward advanced theoretical and research-oriented work. The observed similarity may suggest that the current curriculum of the scientific–pedagogical track does not provide enough differentiation from undergraduate study in terms of content, assessment logic, or academic demands. In this regard,

the results support the idea that this program could benefit from revision. Such revision may include a stronger research component, more consistent theoretical preparation, clearer learning outcomes, and better alignment between course content, teaching methods, and assessment tasks.

The study also shows that academic performance is influenced by the nature of the subjects students take. Courses with a stronger analytical, mathematical, or algorithmic component tend to produce lower average scores and greater variation in results. This may reflect the higher cognitive demands of such disciplines, as well as differences in students' prior preparation and confidence when working with abstract or technically complex material. By contrast, more applied subjects, including courses related to databases and internet technologies, generally show higher average grades. However, even within these better-performing disciplines, noticeable variation remains, which suggests that strong average performance does not necessarily mean that outcomes are uniform across all students.

An important conclusion of the study is that differences in academic outcomes should be considered not only at the level of student groups, but also at the level of curriculum design and assessment practice. Where grade dispersion is high and outliers are frequent, the issue may lie partly in uneven preparation, but it may also be linked to differences in course organization, grading standards, or the clarity of performance criteria. For this reason, one of the practical implications of the study is the need for more standardized and transparent approaches to assessment, especially in technical and analytically demanding disciplines. Clearer criteria, more consistent grading procedures, and closer alignment between learning objectives and evaluation methods could help reduce unnecessary variability and make the comparison of results across groups more meaningful.

Overall, the study supports the conclusion that educational level remains a relevant predictor of academic performance, but this relationship is not isolated from program structure, subject difficulty, and assessment design. The results suggest that academic success in higher education should be understood as the outcome of several interacting factors, including the level of study, the type of curriculum, and the characteristics of the disciplines being taught. In practical terms, these findings may be useful for improving curriculum development, revising master's programs with weaker differentiation, and strengthening support for students in courses that are associated with lower performance and wider score dispersion.

References

Akpen C.N., Asaolu S., Atobatele S., Okagbue H., Sampson S. (2024) Impact of online learning on student's performance and engagement: A systematic review. Discover Education, 3:205. DOI: 10.1007/s44217-024-00253-0 (in English).

Berlanga V., Corti F. (2025) Impact of scholarships on university academic performance: A comparative analysis of students with and without scholarships. Frontiers in Education, 10:1554073. DOI: 10.3389/feduc.2025.1554073 (in English).

Broadbent J., Poon W.L. (2015) Self-regulated learning strategies and academic achievement

in online higher education learning environments: A systematic review. The Internet and Higher Education, 27:1–13. DOI: 10.1016/j.iheduc.2015.04.007 (in English).

Conijn R., Snijders C., Kleingeld A., Matzat U. (2017) Predicting student performance from LMS behavior: Variability across courses and prediction models. Computers in Human Behavior, 68:93–104. DOI: 10.1016/j.chb.2016.11.023 (in English).

Dent A.L., Koenka A.C. (2016) The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. Educational Psychology Review, 28(3):425–474. DOI: 10.1007/s10648-015-9320-8 (in English).

Iglesias-Pradas S., Hernández-García Á., et al. (2021) Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: A case study. Computers in Human Behavior, 119:106713. DOI: 10.1016/j.chb.2021.106713 (in English).

Juarros-Basterretxea J., Aonso-Diego G., Postigo Á., Montes-Álvarez P., Menéndez-Aller Á., García-Cueto E. (2024) Post-hoc tests in one-way ANOVA: The case for normal distribution. Methodology, 20(2):84–99. DOI: 10.5964/meth.11721 (in English).

Kuh G.D., Kinzie J., Buckley J.A., Bridges B.K., Hayek J.C. (2008) What matters to student success: A review of the literature. National Postsecondary Education Cooperative. DOI: 10.1109/FIE.2005.1612154 (in English).

Mammadov S. (2022) Big five personality traits and academic performance: A meta-analysis. Journal of Personality, 90(2):222–255. DOI: 10.1111/jopy.12663 (in English).

Mena-Guacas A.F., López-Catalán L., Bernal-Bravo C., Ballesteros-Regaña C. (2025) Educational transformation through emerging technologies: Critical review of scientific impact on learning. Education Sciences, 15(3):368. DOI: 10.3390/educsci15030368 (in English).

Milkova E., Moldoveanu M., Krcil T. (2024) Sustainable education through information and communication technology: A case study on enhancing digital competence and academic performance of social science higher education students. Sustainability, 17(10):4422. DOI: 10.3390/su17104422 (in English).

Mishra P., Singh U., Pandey C.M., Mishra P., Pandey G. (2019) Application of student's t-test, analysis of variance, and covariance. Annals of Cardiac Anaesthesia, 22(4):407–411. DOI: 10.4103/aca.ACA_94_19 (in English).

Mulyaningsih T., Dong S., Miranti R., Daly A., Purwaningsih Y. (2022) Targeted scholarship for higher education and academic performance: Evidence from Indonesia. International Journal of Educational Development, 88:102510. DOI: 10.1016/j.ijedudev.2021.102510 (in English).

Richardson M., Abraham C., Bond R. (2012) Psychological correlates of university students' academic performance: A systematic review and meta-analysis. Psychological Bulletin, 138(2):353–387. DOI: 10.1037/a0026838 (in English).

Tempelaar D., Rienties B., Nguyen Q. (2015) Learning analytics in the classroom: Improving student performance through integrated analytics models. Computers & Education, 87:98–110. DOI: 10.1016/j.compedu.2015.04.010 (in English).

Witteveen D., Attewell P. (2020) The STEM grading penalty: An alternative to the "leaky pipeline" hypothesis. Science Education, 104(4):714–735. DOI: 10.1002/sce.21580 (in English).

Xu Z., Zhao Y., Zhang B., Liew J., Kogut A. (2023) A meta-analysis of the efficacy of self-regulated learning interventions on academic achievement in online and blended environments. Behaviour & Information Technology, 42(16):2911–2931. DOI: 10.1080/0144929X.2022.2151935 (in English).

Zhang Y., Zhang M., Wu L., et al. (2025) Digital transition framework for higher education in AI-assisted engineering teaching. Science & Education, 34:933–954. DOI: 10.1007/s11191-024-00575-3 (in English).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 316–336

https://doi.org/10.32014/2026.2518-1726.415

IRSTI 28.23.37
UDC 004.89

© Oralbekova D.[1*], Mamyrbayev O. [1], Akhmediyarova A.[2], Kassymova D. [3], Alibiyeva Z. [2], 2026.
[1] Institute of information and computational technologies, Almaty, Kazakhstan;
[2] Satbayev University, Almaty, Kazakhstan;
[3] Mukhametzhan Tynyshbayev ALT University, Almaty, Kazakhstan.
*E-mail: dinaoral@mail.ru

DEVELOPMENT OF A MULTI-LEVEL MODEL FOR TEXT SUMMARIZATION BASED ON PRETRAINED MODELS

Oralbekova Dina — PhD, associate professor, Institute of information and computational technologies, Almaty, Kazakhstan,
E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Mamyrbayev Orken — PhD, professor, Institute of information and computational technologies, Almaty, Kazakhstan,
E-mail: morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;
Akhmediyarova Ainur — PhD, professor, Satbayev University, Almaty, Kazakhstan,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Kassymova Dinara — PhD, associate professor, Mukhametzhan Tynyshbayev ALT University, Almaty, Kazakhstan,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317;
Alibiyeva Zhibek — PhD, associate professor, Satbayev University, Almaty, Kazakhstan,
E-mail: zh.alibiyeva@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Abstract. This paper explores the application of modern transformer models to the task of abstract text summarization in Kazakh, a resource-poor language characterized by an agglutinative structure and complex morphology. These language features complicate the application of traditional approaches to text processing, necessitating the use of more adapted models. The proposed summarization system utilizes multi-level text processing, including symbolic, subword, word, and contextual representation levels. This structure allows for consideration of both the morphological and semantic properties of the Kazakh language. The base models used are the multilingual transformers mBART, mT5, and XLM-RoBERTa, which were adapted and further trained for the task of abstract text summarization. A specialized corpus of 1,000 Kazakh-language news articles with manually compiled annotations was compiled for training and evaluating

the quality of the models. Preprocessing utilized character representations, SentencePiece word-by-word tokenization, FastText word vectors, and contextual transformer embeddings. The quality of the generated summaries was assessed using a set of automated metrics, including ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR, and BERTScore F1, allowing for the analysis of both superficial matches and semantic correspondence with reference annotations. Experimental results showed that the mBART model demonstrated the best performance across most metrics, while the combination of XLM-RoBERTa and BART also yielded stable and competitive results. The experiments demonstrated that a multi-level text processing scheme contributes to improved abstract summaries. Modern transformer models also demonstrate good results when working with Kazakh-language texts.

Keywords: automatic summarization, multi-level language modeling, Kazakh language, transformer architectures, hybrid approach

For citations: Oralbekova D., Mamyrbayev O., Akhmediyarova A., Kassymova D., Alibiyeva Z. Development of a multi-level model for text summarization based on pretrained models. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 316–336. DOI: https://doi.org/10.32014/2026.2518-1726.415

Оралбекова Д.[1*], Мамырбаев О.[1], Ахмедиярова А.[2], Қасымова Д.[3], Алибиева Ж.[2], 2026.

[1] Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;

[2] Satbayev университеті, Алматы, Қазақстан;

[3] Mukhametzhan Tynyshbayev ALT University, Алматы, Қазақстан.

E-mail: dinaoral@mail.ru

АЛДЫН АЛА ОҚЫТЫЛҒАН МОДЕЛЬДЕР НЕГІЗІНДЕ МӘТІНДІ РЕЗЮМЕЛЕУГЕ АРНАЛҒАН КӨПДЕҢГЕЙЛІ МОДЕЛЬДІ ӘЗІРЛЕУ

Оралбекова Дина — PhD, қауымд. проф., Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Мамырбаев Өркен — PhD, профессор, бас директордың орынбасары, Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;
Ахмедиярова Айнұр — PhD, профессор, Satbayev университеті, Алматы, Қазақстан,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Қасымова Динара — PhD, қауымд. проф., Mukhametzhan Tynyshbayev ALT University, Алматы, Қазақстан,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317.
Алибиева Жибек — PhD, қауымд. проф., Satbayev университеті, Алматы, Қазақстан,
E-mail: zh.alibiyeva@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Аннотация. Бұл мақалада агглютинативті құрылымы мен күрделі морфологиясымен сипатталатын ресурстары аз қазақ тіліндегі абстрактілі мәтінді қысқаша баяндау міндетіне заманауи трансформерлік модельдерді қолдану қарастырылады. Бұл тілдік ерекшеліктер мәтінді өңдеуге дәстүрлі тәсілдерді қолдануды қиындатады, ал бұл бейімделген модельдерді пайдалануды қажет етеді. Ұсынылған қысқаша резюмелеу жүйесі символдық, сөздік, және контекстік көрініс деңгейлерін қоса алғанда, көп деңгейлі мәтінді өңдеуді пайдаланады. Бұл құрылым қазақ тілінің морфологиялық және семантикалық қасиеттерін ескеруге мүмкіндік береді. Қолданылатын базалық модельдер абстрактілі мәтінді қысқаша баяндау міндетіне бейімделген және одан әрі оқытылған көптілді mBART, mT5 және XLM-RoBERTa трансформер модельдері болып табылады. Оқыту және бағалау үшін қолмен дайындалған эталондық аннотациялары бар қазақ тіліндегі 1 000 жаңалық мақаласынан тұратын арнайы корпус құрастырылды. Алдын ала өңдеу кезеңінде таңбалық деңгейдегі ұсынылымдар, SentencePiece негізіндегі субсөздік токенизация, FastText сөздік ендірулері және трансформер модельдерінен алынған контекстік ендірулер пайдаланылды. Жасалған қысқаша резюмелеу сапасы ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR және BERTScore F1 сияқты автоматтандырылған көрсеткіштер жиынтығын пайдаланып бағаланды, бұл үстірт сәйкестіктерді де, сілтеме аңдатпаларымен семантикалық сәйкестікті де талдауға мүмкіндік берді. Тәжірибелік нәтижелер mBART моделі көптеген көрсеткіштер бойынша ең жақсы өнімділікті көрсеткенін, ал XLM-RoBERTa және BART комбинациясы тұрақты және бәсекеге қабілетті нәтижелер бергенін көрсетті. Тәжірибелер көп деңгейлі мәтінді өңдеу схемасы абстрактілі қорытындыларды жақсартуға ықпал ететінін көрсетті. Қазіргі заманғы трансформаторлық модельдер қазақ тіліндегі мәтіндермен жұмыс істеген кезде де жақсы нәтижелер көрсетеді.

Түйін сөздер: автоматты резюмелеу, көп деңгейлі тілдік модельдеу, қазақ тілі, трансформерлік архитектуралар, гибридті тәсіл

Оралбекова Д.[1*], Мамырбаев О.[1], Ахмедиярова А.[2], Касымова Д.[3], Алибиева Ж.[2], 2026.

[1] Институт информационных и вычислительных технологий, Алматы, Казахстан;

[2] Satbayev University, Алматы, Казахстан;

[3] ALT университет имени М. Тынышпаева, Алматы, Казахстан.
E-mail: dinaoral@mail.ru

РАЗРАБОТКА МНОГОУРОВНЕВОЙ МОДЕЛИ В ЗАДАЧАХ РЕЗЮМИРОВАНИЯ ТЕКСТА НА ОСНОВЕ ПРЕДВАРИТЕЛЬНО ОБУЧЕННЫХ МОДЕЛЕЙ

Оралбекова Дина — PhD, ассоц. проф., Институт информационных и вычислительных технологий, Алматы, Казахстан,

E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Мамырбаев Оркен — PhD, профессор, заместитель генерального директора, Институт информационных и вычислительных технологий. Алматы, Казахстан,
E-mail: morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;
Ахмедиярова Айнур — PhD, профессор, Satbayev Университет, Алматы, Казахстан,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Касымова Динара — PhD ассоц. проф., ALT университет имени М. Тынышпаева, Алматы, Казахстан,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317;
Алибиева Жибек — PhD, ассоц. проф., Satbayev Университет, Алматы, Казахстан,
E-mail: zh.alibiyeva@satbayev.university, https://orcid.org/0000-0001-9565-5621.

Аннотация. В работе исследуется применение современных трансформерных моделей для задачи абстрактивного резюмирования текстов на казахском языке, относящемся к малоресурсным языкам и характеризующемся агглютинативной структурой и сложной морфологией. Эти особенности затрудняют использование традиционных методов обработки текста и требуют применения более адаптированных подходов. Предложенная система основана на многоуровневой обработке текста, включающей символьный, подсловный, словный и контекстный уровни представления. Такой подход позволяет учитывать как морфологические, так и семантические характеристики казахского языка. В качестве базовых моделей использованы многоязычные трансформеры mBART, mT5 и XLM-RoBERTa, которые были адаптированы и дообучены для задачи абстрактивного резюмирования. Для обучения и оценки качества моделей сформирован специализированный корпус, включающий 1000 новостных статей на казахском языке с вручную составленными аннотациями. В процессе предобработки применялись символьные представления, подсловная токенизация SentencePiece, словные векторные представления FastText, а также контекстные эмбеддинги трансформерных моделей. Оценка качества сгенерированных резюме проводилась с использованием набора автоматических метрик, включая ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR и BERTScore (F1), что позволило учесть как лексическое совпадение, так и семантическую близость к эталонным аннотациям. Экспериментальные результаты показали, что модель mBART демонстрирует наилучшие показатели по большинству метрик, тогда как комбинация XLM-RoBERTa и BART обеспечивает стабильные и конкурентоспособные результаты. Проведённые эксперименты подтверждают, что многоуровневая схема обработки текста способствует повышению качества абстрактивного резюмирования. Использование современных трансформерных моделей также демонстрирует высокую эффективность при работе с казахскоязычными текстами.

Ключевые слова: автоматическое резюмирование, многоуровневое языковое моделирование, казахский язык, трансформерные архитектуры, гибридный подход

Финансирование. *Данное исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант BR24993166).*

Введение. Языковое моделирование (ЯМ) представляет собой одну из центральных задач в области обработки естественного языка (NLP), которая включает в себя создание и развитие алгоритмов, способных обрабатывать, понимать и генерировать текстовые данные на человеческом языке. В последние годы с развитием алгоритмов глубокого обучения языковое моделирование стало одним из популярных направлений в NLP. Появление архитектур на основе Трансформера (Vaswani et al., 2017; Rahali et al., 2023) внесли существенный вклад в понимании контекста данных, структуру предложения и семантические взаимосвязи слов. На сегодняшний день такие модели используются при решении различных задач NLP, включая машинный перевод, автоматическое составление кратких аннотаций, системы не основе вопросов и ответом, а также в анализе тональности текстов (Narejo et al., 2024; Oralbekova et al., 2024).

Среди задач языкового моделирования значительное внимание уделяется автоматическому резюмированию текста. Существующие методы можно разделить на два типа. К первому можно отнести extractive summarization, который извлекает ключевые предложения из исходного текста. Вторым методом является abstractive summarization, при котором идет генерация нового текста на основе исходного контента (Giarelis et al., 2023). В последние годы модели трансформеров, такие как BERT (Kalyan et al., 2021), GPT (Liu et al., 2024), T5 (Raffel et al., 2020) и BART (Lewis et al., 2020), показали значительные успехи в задачах генерации текста, включая резюмирование. Для абстрактного резюмирования часто используются модели BART и T5. Их архитектура на основе кодер-декодер обеспечивают преобразования исходного текста в краткую и связную форму. Кроме традиционных методов резюмирования, есть еще гибридных подход, который сочетает характеристики обоих типов аннотирования (Kirmani et al., 2019). Гибридное резюмирование включает в себя извлечение ключевых фрагментов текста с последующей генерацией резюме на основе этих фрагментов, что позволяет сочетать точность extractive и гибкость генерации. Гибридные методы резюмирования объединяют свойства извлекательного и abstractive подходов. С их помощью можно сохранить ключевые элементы исходного текста и в тоже время сформировать компактное и связное изложение содеражания. Такой подход особенно актуален для казахского языка, потому что текстовые корпуса и цифровые контенты на этом языке остаются ограниченными.

Развитие многоязычных архитектур также сыграло важную роль в этой области. Модели mBART и mT5 обучаются на текстах разных языков. Они способны работать даже с языками, для которых объем обучающих данных относительно невелик. Именно это особенность позволяет применять общие

алгоритмические подходы при обработке многоязычных текстов. mBART (Multilingual BART) и mT5 (Multilingual T5) продемонстрировали высокие результаты в многозадачных сценариях, таких как перевод, анализ текста и генерация резюме. Модель XLM-RoBERTa (Conneau et al., 2020) представляет собой многоязычное расширение архитектуры RoBERTa (Liu et al., 2019). Данная модель показала высокую эффективность при решении различных задач NLP, включая классификацию текста и аннотирование информации.

Современные языковые модели достигли заметных успехов при обработке распространённых языков, включая английский, французский и китайский. Однако для малоресурсных языков подобные результаты пока достигнуты не в полной мере. К таким языкам можно отнести семейство агглютинативных и тюркоязычных языков, куда и входит казахский язык, что требуют использования специальных методов обработки текста. Кроме этого, объём доступных текстовых корпусов и количество предобученных моделей для казахского языка остаются ограниченными. Поэтому автоматическое резюмирование казахских текстов недостаточно исследованы, а известные подходы показывают умеренные показатели качества (Zulkhazhav et al., 2019; Kynabay et al., 2021; Zhabayev et al., 2021).

В настоящей работе предложена многоуровневая ЯМ, разработанная специально для казахского языка. Мы представляем модель, которая обрабатывает текст на четырех уровнях: символьном, подсловном, словном и контекстном. Применение такой структуры позволяет более полно учитывать морфологические характеристики языка и текста. Кроме того, данный подход способствует повышению точности и информативности автоматически формируемых аннотаций. В данном исследовании предлагается использование моделей mT5, mBART и XLM-RoBERTa, которые были адаптированы и дообучены для казахского языка.

В разделе 2 представлена обзорная литература по задаче резюмирования текста с акцентом на существующие методы и модели, включая как экстрактивные, так и абстрактивные подходы, а также гибридные методы. В разделе 3 подробно описаны используемые модели и их архитектуры, включая многоязычные трансформеры, такие как mT5, mBART и XLM-RoBERTa, а также принципы их адаптации для задачи резюмирования на казахском языке. В разделе 4 изложены методы проведения экспериментов и полученные результаты, с анализом их значимости и обсуждением достигнутых показателей. Раздел 5 содержит заключение, в котором подведены итоги работы, обсуждаются направления для будущих исследований в области NLP.

Материалы и методы исследования. В последние годы значительные усилия были направлены на разработку гибридных моделей резюмирования, ориентированных как на ресурсообеспеченные, так и на низкоресурсные языки. В последние годы опубликован ряд работ, посвящённых разработке моделей автоматического резюмирования для малоресурсных языков.

В работе (Winarko et al., 2025) предложили модель абстрактного резюмирования для индонезийского языка. В данной архитектуре в качестве энкодера используется стек эмбеддингов, а декодирующая часть построена на Трансформере. В качестве эмбеддингов использовались BERT, Byte Pair Encoding (BPE), Character Embedding (CE) и FastText. Авторы провели исследование влияния выбора слоёв BERT и конфигурации эмбеддингов на качество резюме. Модель обучалась на подкорпусах Liputan6 (50K и 75K статей). Наиболее высокие значения ROUGE-1 (37.18), ROUGE-2 (18.19) и ROUGE-L (34.28) были достигнуты при использовании всех слоёв BERT. Предложенная архитектура показала высокие результаты даже при ограниченном обучающем объёме данных.

В исследовании (Raza et al., 2024) описана гибридная система резюмирования текстов на языке урду. Авторы объединили два традиционных метода резюмирования. На первом этапе применялись классические подходы, основанные на TF-IDF, весах предложений и частоте употребления слов Полученный фрагмент затем использовался в качестве входных данных для абстрактной модели, построенной на архитектуре BERT. Качество сформированных резюме оценивалось специалистами, владеющими языком урду. Для обучения и тестирования был сформирован корпус из материалов изданий Express, BBC Urdu, Dawn и других источников на различные темы. Данное исследование относится к числу ранних работ по абстрактному резюмированию текстов на урду и учитывает морфологическую сложность и лексическое разнообразие языка.

В работе (Challagundla et al., 2024) была разработана комплексная структура модели для абстрактного резюмирования. Предложенный подход объединил методы seq2seq с механизмом внимания, семантическое обобщение и разрешении омонимии. Векторизация текста выполнялась с помощью Word2Vec, архитектура включала двунаправленные LSTM, кастомный attention и TimeDistributed слой. Для обучения и тестирования модели использовался корпус из материалов Gigaword, DUC-2004 и CNN/DailyMail. Использование WSD и постобработки улучшило согласованность и охват редких слов, повысив качество резюме по сравнению с базовыми моделями seq2seq без семантических дополнений.

Для обработки патентных документов был предложил гибридный подход к резюмированию (Jayatilleke et al. 2025). В нём сочетается экстрактивный алгоритм LexRank и модель BART, дополнительно адаптированный с применением метода Low-Rank Adaptation (LoRA). Модель дополнительно обобщалась на новые патентные области с помощью методов метаобучения. Архитектура была протестирована на длинных патентных текстах, отличающихся сложной юридической терминологией. Благодаря LoRA удалось сократить вычислительные ресурсы при сохранении качества. Такой подход оказался эффективным для создания абстрактных резюме в области интеллектуальной собственности.

В работе (Do et al., 2021) предложили Discrete Diffusion Language Model (DDLM) с семантически управляемым шумом (Semantic-Aware Noising) и архитектурой CrossMamba — адаптацией модели Mamba для задач резюмирования длинных текстов. Их модель показала превосходство над ранее предложенными дискретными диффузионными моделями и даже опередила автопорождённые (autoregressive) модели по скорости генерации на наборах данных CNN/DailyMail, Arxiv и Gigaword. В данной работе предлагается заменить случайный шум семантически ориентированным. Это позволяет сохранять важные токены и контролировать структуру формируемого текста.

В (Gogireddy et al., 2024) рассматривается система резюмирования, состоящая из двух этапов. На первом этапе используется Graph Neural Networks для выбора наиболее значимых дискурсивных единиц. Затем полученные данные передаются в модель BART для формирования окончательного варианта аннотации. Эксперименты проводились на CNN/DailyMail, и предложенный подход показал более высокие значения ROUGE по сравнению с базовыми трансформерами. Метод демонстрирует эффективность объединения графовых структур с генеративными трансформерами.

Исследование (Faizal et al., 2024) посвящено задаче абстрактного резюмирования финансовых и бизнес-отчётов. В предложенном подходе объединяется анализ текстовых и табличных данных. Модель построена на основе архитектуры Transformer и дополнена элементами Switch Transformer для эффективной обработки многоформатного финансового контента. Особенностью работы является адаптация модели к структурам таблиц, характерным для финансовых документов. Эксперименты проводились на бизнес-датасетах Reuters Financial и Bloomberg. Качество модели оценивалось с помощью метрик ROUGE и F-measure, показавших улучшенные результаты по сравнению с классическим Transformer-ом.

В (Rounak Chakraborti et al., 2025) провели сравнительный анализ различных моделей резюмирования, включая RNN, LSTM, Seq2Seq, BART, T5 и Pegasus. Исследование сосредоточено на задачах новостного резюмирования, где анализировалась точность, связность и общая пригодность моделей. Было использовано несколько корпусов, включая CNN/DailyMail и BBC News. В работе отмечены преимущества моделей Pegasus и BART, особенно в части генерации контекстуально осмысленных резюме. Также обсуждаются ограничения метрик ROUGE и необходимость более сложных оценок (например, по связности и достоверности).

В (Rao et al., 2025) представили новую метрику оценки для абстрактного резюмирования, основанную на семантическом сходстве между референсным и сгенерированным резюме. В качестве базовой модели использовался Universal Sentence Encoder (USE), позволяющий сравнивать смыслы на уровне предложений. Было показано, что предложенная метрика

лучше коррелирует с оценками людей, чем ROUGE, особенно при низком n-граммовом совпадении. Эксперименты проводились на CNN/DailyMail и других наборах данных. В работе также рассмотрены другие альтернативные метрики: BERTScore, MoverScore и ROUGE-G.

В исследовательской работе (Zangooei et al., 2025) разработали модель ARLED для абстрактного резюмирования длинных текстов на персидском языке. Архитектура объединила модели ARMAN (на основе Longformer) и LED (Longformer Encoder-Decoder), что позволило эффективно работать с входами до 8192 токенов. Был создан новый корпус из 49,457 персидских научных статей с Ensani.ir, прошедший глубокую предобработку. Токенизация данных выполнялась с помощью HuggingFace AutoTokenizer. Кроме того, в процессе обучения были реализованы стратегии семантической переупорядоченности предложений. Подобный подход ранее редко применялся при работе с персидским языком.

В (Langston et al., 2024) рассматривается автоматизированный подход к резюмированию аннотаций и заголовков нескольких документов с использованием крупных языковых моделей (LLMs). Предложенная система использует комбинированную стратегию. На первом этапе из текста извлекаются ключевые фрагменты, после чего выполняется их абстрактное обобщение при помощи ЯМ. Эксперименты показали, что система демонстрирует особенно высокую эффективность при работе с техническими и медицинскими текстами. В работе также отмечается важность адаптации моделей к конкретной предметной области и рассматриваются перспективы использования методов обучения с подкреплением (RL) и специализированных механизмов внимания.

Авторы (Barta et al., 2024) представили открытый венгерский корпус HunSum-2, сформированный на основе новостных текстов из Common Crawl и предназначенный для задач как экстрактивного, так и абстрактивного резюмирования. Для предварительной обработки данных применялись FastText (для определения языка) и quntoken (для токенизации и разбиения на предложения). В абстрактивных экспериментах использовались модели на базе mT5 и архитектура Bert2Bert с ограничением на длину входной последовательности до 512 токенов. В экстрактивном подходе были реализованы модели BertSum с использованием huBERT и простого классификатора, где метки предложений генерировались с опорой на косинусное расстояние между их эмбеддингами. Анализ экспериментальных данных показал, что по метрикам ROUGE и BERTScore более высокие значения были получены для экстрактивных моделей. Среди абстрактивных систем наиболее конкурентоспособной оказалась модель mT5, однако её показатели всё же уступали экстрактивным методам. При этом абстрактивные модели чаще допускали несоответствия и отклонения от исходного содержания.

Алгоритм многоуровневого моделирования

Многоуровневое моделирование основывается на идее о том, что для

успешной обработки текста, особенно для языков с богатой морфологией и сложными структурами, необходимо учитывать различные уровни текстовых единиц. В разработанной архитектуре предусмотрено несколько уровней обработки данных (алгоритм 1):

- Символьный уровень предназначен для анализа структуры слова. Для казахского языка такой подход особенно важен, поскольку изменения в морфемах могут существенно менять смысл слова. В модели применяются представления character-level embeddings для обеспечения обработки текста на уровне отдельных символов (Jebbara et al., 2017).

- При работе с агглютинативными языками, такими как казахский, важным этапом является подсловная обработка текста. В таких языках одно слово может включать несколько морфем и иметь сложную структуру. Использование подхода вроде Sentence Piece позволяет модель обрабатывать текст на уровне частей слов (Kudo et al., 2018).

- Словный уровень помогает выявить семантическую составляющую текста. Применение векторных представлений слов с помощью FastText позволяет модели понимать семантическое значение отдельных слов (Stein et al., 2019).

- На контекстном уровне модель анализирует взаимосвязи между словами и предложениями. Архитектуры Transformer дают возможность учитывать контекст их употребления, и грамматически корректно генерировать краткое содержание текста.

Алгоритм 1: Гибридный подход к резюмированию текста на основе предобученных моделей

Цель: Алгоритм предназначен для генерации кратких резюме казахских текстов. Предлагаемый подход основан на многоуровневой обработке данных, включающей символьный, подсловный, словный и контекстный уровни представления текста.

Входные данные: Корпус казахских текстов с соответствующими резюме.

Выходные данные: Сгенерированные резюме для наиболее представительных казахских текстов.

– Производится загрузка датасета, включающего казахские тексты и подготовленные к ним резюме.

– Осуществляется предобработка входных данных.

– К входным данным добавляется специальный префикс «summarize:»

– Токенизация текста на символьном уровне с использованием моделей, таких как Character-level embeddings, чтобы захватить структуру слов, особенно для казахского языка, где морфологические изменения критичны.

– Обработка текста включает этап подсловной токенизации, для которого применяются алгоритмы Byte Pair Encoding (BPE) или SentencePiece.

– Аналогичная процедура токенизации применяется и к соответствующим резюме.

1. Загрузка и инициализация предобученной модели, применяемой для задач типа seq2seq.

2. Дообучение модели на подготовленном датасете (казахские тексты с резюме) с использованием соответствующих параметров обучения, таких как скорость обучения, размер батча и количество эпох.

3. В процессе обучения выполняется вычисление оценочных метрик, которые отслеживают динамику качества работы модели.

4. Для каждого входного текста обученная модель формирует краткое резюме. При этом учитываются контекстные зависимости между словами на уровне всего текста.

5. На завершающем этапе выводятся сгенерированные резюме, после чего проводится оценка качества модели на основе полученных значений метрик.

Модель mT5

mT5 представляет собой многоязычную модификацию модели T5, обученную на корпусах более чем 100 языков, включая казахский. Архитектура модели построена по схеме кодер-декодер и ориентирована на решение задач обработки естественного языка в формате «от текста к тексту». Благодаря этому модель может применяться для различных задач, включая машинный перевод, генерацию текста и автоматическое резюмирование.

Архитектурно mT5 основана на трансформере, где кодер обрабатывает входную последовательность, а декодер формирует выходной текст. Такая структура делает модель удобной для задач seq2seq, где требуется не только извлечение ключевых фрагментов, но и формирование нового текста из исходного документа. Модель mT5 использует SentencePiece токенизацию для преобразования текста в последовательности токенов. В архитектуре в Encoder входной текст разбивается на токены и передается через несколько слоев self-attention механизма для вычисления скрытых состояний. Каждый слой состоит из multi-head attention и feed-forward neural network. После прохождения всех слоев, выходные представления становятся контекстуальными эмбеддингами, которые затем передаются в декодер. В архитектуре модели декодер получает скрытые состояния, сформированные кодером и использует их для последовательной генерации текста. Механизм внимания позволяет учитывать ранее сгенерированные токены. В результате предсказание нового токена выполняется на основе уже сформированной части текста и внутренних представлений модели.

Модель mBART

mBART является многоязычной версией архитектуры BART. Данная модель применяется в задачах генерации текста, включая машинный перевод, суммирование и другие задачи, связанные с обработкой текстовых последовательностей. Модель mBART также как и mT5 использует архитектуру encoder-decoder (рис. 1).

mBART обучена на нескольких языках с использованием многоязычного корпуса. Это позволяет ей работать с текстами на различных языках, включая казахский. mBART использует SentencePiece для токенизации. Модель

продемонстрировала отличные результаты в задачах перевода, генерации и резюмирования текста, особенно в случае, когда нужно работать с текстами на малоресурсных языках. Архитектура mBART объединяет свойства моделей BERT и GPT. Двунаправленный механизм обработки контекста реализован в энкодере, тогда как декодер работает в авторегрессионном режиме.

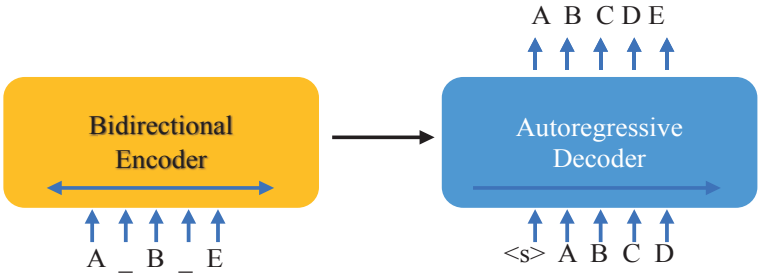


Рисунок 1 – BART architecture

Модель XLM-RoBERTa

XLM-RoBERTa (Cross-lingual RoBERTa) является улучшенной версией модели RoBERTa, разработанной для работы с многоязычными текстами. XLM-RoBERTa построена на основе архитектуры Transformer и обучена на масштабном многоязычном корпусе, содержащем тексты более чем на 100 языках, включая казахский. Архитектура модели предполагает использование двунаправленного контекста, что позволяет учитывать слова, расположенные по обе стороны от текущего токена. Такой механизм способствует более точному моделированию зависимостей в тексте. Благодаря обучению на многоязычных данных модель может эффективно применяться для малоресурсных языков. Она формирует контекстные представления слов и выявляет связи между ними даже в случае редких или морфологически сложных форм. Предобучение выполнялось с использованием стратегии маскированного предсказания токенов, что улучшает качество языковых представлений. В архитектуре модели кодер отвечает за обработку входной последовательности. В ходе кодирования токены преобразуются в контекстные эмбеддинги, отражающие их синтаксические и семантические связи. Применение механизма внимания позволяет учитывать окружение каждого токена с обеих сторон. После обработки формируется набор скрытых представлений, описывающих текст на более абстрактном уровне. Эти представления используются для извлечения признаков на словном и контекстном уровнях и служат основой для дальнейшего анализа и генерации текста.

Токенизация в модели XLM-RoBERTa выполняется с использованием алгоритма BPE. Этот метод разделяет слова на подсловные единицы и облегчает обработку редких или сложных слов и повышает устойчивость модели при работе с новыми языковыми данными.

Применение многоуровневой языковой модели в задаче автоматического резюмирования предполагает анализ текста на нескольких уровнях представления. Такой подход особенно актуален для языков с развитой морфологией. Каждый уровень архитектуры участвует в извлечении и обработке различных типов информации из текста. В рамках исследования используется модель, включающая символьный, подсловный, словный и контекстный уровни обработки.

Обработка текста на символьном уровне реализована с использованием character-level embeddings для учитывания внутреннюю структуру слов. Такой способ представления данных полезен для казахского языка, где словоформы могут значительно изменяться в зависимости от грамматических категорий, а одна и та же морфема иногда имеет несколько вариантов записи. Учет символьной информации облегчает обработку морфологических изменений и разнообразных словоформ.

На подсловном уровне применяется токенизация с использованием алгоритма SentencePiece для улавливания особенности аффиксации, характерные для казахского языка. Это позволяет обрабатывать составные слова и аффиксальные формы, которые часто встречаются в тексте на казахском языке. SentencePiece разбивает слова на более мелкие единицы (подслова). Это дает модели эффективно справляться с неизвестными словами и морфемами.

На словном уровне используется FastText для векторизации слов. Применение FastText позволяет учитывать семантические связи между словами и формировать контекстно-зависимые векторные представления. Одним из преимуществ данной модели является возможность построения представлений даже для слов, которые отсутствовали в обучающем корпусе. Для казахского языка это особенно актуально, поскольку многие слова отличаются сложной морфологической структурой или встречаются редко.

Для формирования контекстных признаков на данном уровне используется модель XLM-RoBERTa или другие трансформерные архитектуры. Их применение позволяет анализировать зависимости между словами внутри предложения и между предложениями в тексте. Модель XLM-RoBERTa способна учитывать как локальные, так и долгосрочные контекстные зависимости. Благодаря этому формируется более точное представление смысла предложения и условий употребления каждого слова. Полученные представления используются при генерации резюме, где важно не только выделить основную информацию, но и передать её в краткой форме.

После завершения многоуровневой обработки текста генерация резюме выполняется моделью BART. BART будет использовать контекстные признаки, полученные на основе представлений XLM-RoBERTa. BART использует архитектуру encoder-decoder, где энкодер извлекает информацию из текста, а декодер создает новое резюме, перефразируя основные идеи исходного текста (рис. 2). Это позволяет получать высококачественные

резюме, которые сохраняют суть исходного материала, но представляют его в более краткой и сжато изложенной форме.

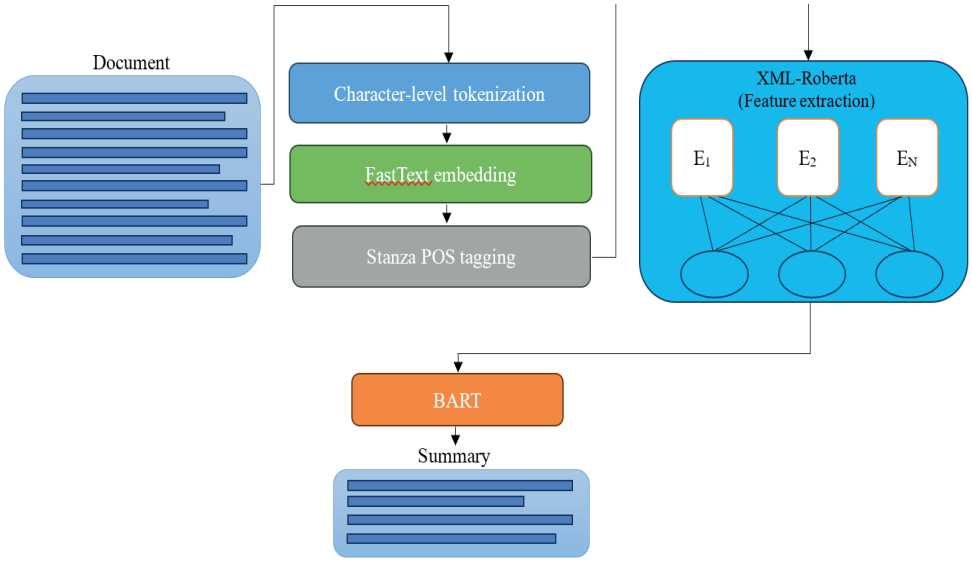


Рисунок 2 – Структура предложенного гибридного подхода

Для моделей mBART и mT5 дополнительное использование BART не требуется. Оба эти подхода уже имеют встроенную архитектуру encoder-decoder. Таким образом, модель mBART и mT5 могут непосредственно генерировать резюме. Эти архитектуры изначально построены по схеме seq2seq и предназначены для задач генерации текста.

Модель XLM-RoBERTa, напротив, реализует только encoder-часть трансформера. Она предназначена для извлечения признаков и не выполняет генерацию текста. По этой причине для формирования резюме используется BART, который выполняет функцию декодера и генерирует итоговое текстовое представление.

Результаты. Для обучения и тонкой настройки моделей был подготовлен корпус новостных материалов. Датасет включает новостные сообщения и аналитические статьи. Каждому тексту сопоставлена краткая аннотация, отражающая его основное содержание.Аннотации формировались вручную специалистом-лингвистом. Всего было собрано 1000 статей для обучения модели.

Для повышения согласованности и стабильности обучения, тексты были приведены к унифицированному формату: подавляющее большинство статей содержит не более 250 слов, а соответствующие аннотации — в среднем около 40 слов. Подобная организация данных позволила сбалансировать корпус по длине текстов и избежать существенных различий между короткими и длинными материалами.

Средняя длина статьи составила 112 слов, а медианная длина — 98 слов; длина резюме варьировалась в пределах 5–50 слов, со средним значением около 21 слова.

Предобработка данных

Перед обучением модели данные были подготовлены с использованием нескольких этапов предобработки. На этапе нормализации текст переводился в нижний регистр, из него удалялись лишние пробелы, а слова приводились к стандартной форме, например к их начальной форме. Это помогло улучшить качество данных и повысить эффективность модели. Из текста были удалены неинформативные символы, такие как специальные знаки, излишние пробелы, а также другие элементы, которые не вносят смысла в задачу резюмирования, такие как ссылки, HTML-теги.

Обработка текста на символьном уровне реализована с применением RNN-модели, которая позволяет учитывать информацию на уровне отдельных символов. Векторные представления слов формировались с помощью FastText. При обучении модели использовались следующие настройки: размер векторного пространства 300, окно контекста 5, а число эпох обучения составляло 10. После этапа токенизации выполнялось POS-размечивание с использованием библиотеки Stanza, и только затем данные передавались в модель для последующего извлечения признаков.

На этапе подготовки данных применялись разные стратегии токенизации в зависимости от архитектуры модели. Для mT5 и mBART использовался токенизатор SentencePiece, который формирует подсловные представления и облегчает обработку редких или составных слов. Для XLM-RoBERTa использовался стандартный токенизатор данной модели, основанный на алгоритме Byte Pair Encoding.

Размер словаря для mT5 и mBART был установлен на уровне 32 000 токенов, тогда как для XLM-RoBERTa он составлял 250 000 подслов. Модель mT5 была настроена с использованием 12 слоев и 768 размером скрытого состояния. Модель mBART использовала 12 слоев и 1024 размер скрытого состояния. Для XLM-RoBERTa и BART использовались 12 слоев с размером скрытого состояния 1024. Для обучения использовались 10 эпох с размером батча 16.

Для обучения использовалось 90% от всего корпуса, оставшиеся 10% были выделены для валидации. Все этапы реализации выполнялись с использованием библиотеки HuggingFace Transformers, включая модули datasets, tokenizers и Trainer, а также gensim для FastText и stanza для POS-аннотации.

При тонкой настройке модели применялись стандартные рекомендации для обучения трансформерных архитектур в задачах генерации текста. В качестве оптимизатора использовался AdamW со скоростью обучения 3e-5. Регуляризация обеспечивалась за счёт Dropout и weight decay с коэффициентом 0.01. Для стабилизации процесса обучения первые 10% шагов

от общего числа использовались в качестве этапа warmup. Чтобы увеличить эффективный размер батча, использовалось накопление градиентов. Норма градиента ограничивалась значением 1.0, а скорость обучения изменялась по линейному расписанию после этапа разогрева. Кроме того, применялся механизм early stopping. Обучение завершалось в случае отсутствия улучшения метрики eval_loss на валидационной выборке в течение трёх последовательных эпох.

Качество работы моделей резюмирования оценивалось с использованием нескольких метрик:

1) Метрики семейства ROUGE используются для сравнения сгенерированного текста с эталонным резюме. Оценка основана на совпадениях n-грамм, подстрок и последовательностей слов. Метрика ROUGE-1 фиксирует совпадения отдельных слов, ROUGE-2 учитывает совпадения пар слов, а ROUGE-L рассчитывается на основе наибольшей общей подпоследовательности. Каждая из метрик ROUGE вычисляется по формуле (1):

$$ROUGE - N = \frac{\sum_{overlap\ of\ N-grams} matches}{\sum_{N-grams\ in\ reference}} \quad (1)$$

где N - размер n-грамм, а «matches» - количество совпавших фрагментов между сгенерированным и эталонным резюме.

2) BLEU (Bilingual Evaluation Understudy) оценивает точность (precision) между сгенерированным и референсным текстом, сравнивая совпадения по n-граммам до 4-го порядка (2):

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} \omega_n \log\ p_n) \quad (2)$$

где - доля совпадающих n-грамм, ω_n - вес (обычно равный), а BP - penalty за слишком короткие предложения.

3) METEOR (Metric for Evaluation of Translation with Explicit ORdering) учитывает совпадения по словам, синонимам, стеммам и парафразам. По сравнению с другими метриками оценки качества резюмирования показатель METEOR более чувствителен к семантическому сходству. Формально (3):

$$METEOR = F_{mean} \cdot (1 - Penalty) \quad (3)$$

где F_{mean} - гармоническое среднее precision и recall, а Penalty зависит от количества перестановок.

4) BERTScore это метрика, основанная на контекстных эмбеддингах. Вместо подсчета поверхностных совпадений между токенами, BERTScore измеряет семантическое сходство между векторными представлениями слов,

извлечённых из предобученной модели. Значение метрики определяется как среднее максимальное косинусное сходство между токенами (4):

$$BERTScore\ F1\ = \frac{1}{N}\sum_{i=1}^{N}\max_{j}\ cos(\mathrm{v}_i, \mathrm{v}_j) \quad (4)$$

здесь N обозначает число токенов в сгенерированном тексте, v_i соответствует векторному представлению i-го токена сгенерированного текста, v_j – вектору j-го токена референсного текста. Функция $cos(\mathrm{v}_i, \mathrm{v}_j)$ определяет сходство между соответствующими векторами.

Данная метрика подходит для оценки качества резюме в случаях, когда сгенерированный текст передает смысл исходного, но не воспроизводит его дословно.

Проведение экспериментов и полученные результаты

Для сравнение полученных результатов были рассмотрены несколько работ для задачи резюмирования на казахских текстах. В настоящее время количество исследований, посвящённых автоматическому резюмированию казахских текстов, остаётся ограниченным. Большинство существующих подходов опираются на экстрактивные методы, такие как TF-IDF и логика нечёткого вывода. Работы по гибридному резюмированию появляются сравнительно недавно, в основном с применением трансформерных архитектур и стратегий трансферного обучения (табл. 1).

В работе (Zulkhazhav et al., 2019) использовалась экстрактивная модель на основе нечёткой логики с предварительным разрешением местоимений и морфологическим анализом. В ходе экспериментов для рассматриваемого метода были получены значения ROUGE-1 F1 = 0.44 и ROUGE-2 = 0.34. Для классического подхода без генеративной архитектуры такие результаты можно считать достаточно высокими.

Экстрактивный метод на основе TF-IDF также представлен в исследовании (Kynabay et al., 2021). В качестве источника текстов использовались новости портала inform.kz. Дополнительное использование специально сформированного списка стоп-слов для казахского языка позволило достичь ROUGE-L F1 = 0.35 без применения абстрактивных моделей.

В работе (Zhabayev et al., 2021) используется подход, при котором предложения сначала извлекаются на основе TF-IDF, после чего упрощаются с помощью модели Seq2Seq. Для обучения модели применялся подход трансферного обучения на корпусе Simple English Wikipedia. По результатам экспериментов значение метрики BLEU достигло 7.0, а показатели ROUGE не приведены.

Таблица 1. Результаты метрик для разных моделей

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore F1	Corpus Size
Zulkhazhav et al. (Fuzzy Logic, with pronoun resolution)	0,44	0,34	0,4	-	-	-	100 новостных текстов
Kynabay et al. (Extractive TF-IDF)	0,4	0,31	0,35	-	-	-	1422 новостных текстов
Zhabayev et al. (TF-IDF + Seq2Seq via Transfer Learning)	-	-	-	7	-	-	Корпус Kaz-skaz
mBART с многоуровневой архитектурой	0,59	0,41	0,51	0,49	0,52	91,40	1000 новостных текстов
mT5-small с многоуровневой архитектурой	0,52	0,37	0,48	0,32	0,37	91,20	1000 новостных текстов
XLM-R + BART с многоуровневой архитектурой	0,57	0,34	0,47	0,33	0,38	90,50	1000 новостных текстов

Обсуждение. На основе проведённых экспериментов и полученных метрик можно выделить несколько ключевых наблюдений, которые показывают различия в производительности моделей mBART, mT5 и XLM-RoBERTa. mBART продемонстрировала значительные улучшения по метрике ROUGE-1, достигая 0.5992. Полученные результаты свидетельствуют о высокой эффективности модели при формировании гибридных резюме. Дополнительным подтверждением служит постепенное увеличение значений метрик в ходе обучения.

Модель mT5 также демонстрирует положительную динамику по метрикам ROUGE-1 и ROUGE-L, достигая максимальных значений 0.5220 и 0.4840 соответственно. Однако её показатели остаются ниже результатов mBART, особенно по метрике ROUGE-2, где значение составляет 0.3760. Эти результаты, хотя и достойны внимания, всё же уступают показателям mBART и XLM-R + BART. Это свидетельствует о меньшей способности mT5 генерировать качественные резюме по сравнению другими моделями.

XLM-R + BART показала уверенное улучшение по метрике ROUGE-2, с максимальным значением 0.3440, а также по метрике ROUGE-L, с результатом 0.4790. Хотя модель XLM-R + BART продемонстрировала заметное улучшение показателей на последних этапах обучения, её значения ROUGE-1 и ROUGE-2 остаются ниже результатов других моделей. Тем не менее система сохраняет способность эффективно учитывать контекст и формировать достаточно точные резюме.

По значениям метрики METEOR модель mBART также демонстрирует наиболее высокие показатели при генерации гибридных резюме. В

экспериментах значение данной метрики достигало 0.5280. Это указывает на способность модели учитывать не только прямые совпадения слов, но и семантические соответствия, включая синонимы, стеммы и парафразированные выражения. Для mT5 и XLM-R + BART результаты по METEOR были ниже, но все равно показывали улучшения по сравнению с предыдущими исследованиями: mT5 достигла 0.3700, а XLM-R + BART — 0.3750. Полученные результаты подтвердили, что данные модели способны генерировать тексты с высоким уровнем семантического сходства.

Оценка по метрике BERTScore F1, предназначенной для измерения семантического сходства между сгенерированными и эталонными текстами, показывает, что модель mBART достигает наиболее высоких значений (0.9140). Такой результат указывает на способность модели сохранять смысловое содержание текста и корректно учитывать контекст при формировании резюме. Это особенно актуально для гибридных методов резюмирования, где ключевым требованием является точная передача смысла исходного документа.

Сравнительный анализ показал, что рассмотренные модели (mBART, mT5, XLM-R + BART) превосходят результаты, представленные в работах [14–16]. В указанных исследованиях значения метрики ROUGE-1 не превышают 0.44 (Zulkhazhav et al.) и 0.35 для ROUGE-L (для Kynabay et al.). Эти данные подтверждают, что современные трансформерные модели, такие как mBART, mT5 и XLM-R + BART, значительно улучшили производительность по сравнению с классическими экстрактивными методами, используемыми в предыдущих исследованиях.

Заключение. В данной работе была разработана и протестирована многоуровневая модель для гибридного резюмирования текстов на казахском языке с использованием современных трансформерных архитектур, таких как mBART, mT5 и XLM-RoBERTa. В представленном исследовании применяется многоуровневый подход к обработке текста. Использование нескольких уровней представления позволяет более полно учитывать особенности казахского языка, включая его морфологическую структуру и семантические связи. Именно такой гибридный подход способствует повышению качества создаваемых резюме. Результаты экспериментов показали, что модели mBART и XLM-RoBERTa, с использованием многоуровневой архитектуры, достигли наилучших результатов по метрикам ROUGE-1, ROUGE-2, ROUGE-L и BERTScore F1.

Дальнейшие исследования могут быть направлены на оптимизацию архитектуры модели, расширение корпуса обучающих данных и адаптацию системы к текстам разных жанров и стилевых типов.

References

Barta B., Lakatos D., Nagy A., Nyist M. K., Ács J. (2024) From News to Summaries: Building a Hungarian Corpus for Extractive and Abstractive Summarization. In Proceedings of the 2024

Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — P. 7503–7509. Torino, Italia. ELRA and ICCL (in Eng.)

Chakraborti R., Banerjee R., Das S. (2025) Evaluating the Efficacy of Text Summarization Models: A Comparison of NLP Algorithms. 2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2025. — P. 1-5, doi: 10.1109/IEMENTech65115.2025.10959463 (in Eng.)

Challagundla B. C., Peddavenkatagari C. (2024). Neural Sequence-to-Sequence Modeling with Attention by Leveraging Deep Learning Architectures for Enhanced Contextual Understanding in Abstractive Text Summarization. International Journal of Machine Learning and Cybernetics. 2. — P. 21-29 (in Eng.)

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. (2020) Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — P. 8440–8451 (in Eng.)

Do H. D., Duc A. D., Anh T. L., Wray B. (2025) Discrete Diffusion Language Model for Efficient Text Summarization. In Findings of the Association for Computational Linguistics: NAACL 2025. — P. 6278–6290 Albuquerque, New Mexico. Association for Computational Linguistics (in Eng.)

Faizal B., Abraham S., Thomas, S. (2024) Automated Business Report Summarization Using Transformer Model. 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), 1. — P. 254-258 (in Eng.)

Giarelis N.; Mastrokostas C.; Karacapilidis N. (2023) Abstractive vs. Extractive Summarization: An Experimental Review. Appl. Sci. 2023, 13. — P. 7620. https://doi.org/10.3390/app13137620 (in Eng.)

Gogireddy Y. R., Bandaru A. N., Sumanth V. (2024) Synergy of Graph-Based Sentence Selection and Transformer Fusion Techniques for Enhanced Text Summarization Performance. Journal of Computer Engineering and Technology (JCET) 7(1). 2024. — P. 33-41 (in Eng.)

Jayatilleke N., Weerasinghe Ruvan. (2025) A Hybrid Architecture with Efficient Fine Tuning for Abstractive Patent Document Summarization. 10.48550/arXiv.2503.10354 (in Eng.)

Jebbara S., Cimiano P. (2017) Improving Opinion-Target Extraction with Character-Level Word Embeddings. In Proceedings of the First Workshop on Subword and Character Level Models in NLP. — P. 159–167 (in Eng.)

Kalyan K.S., Rajasekharan A., Sangeetha S. (2021). AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing. ArXiv, abs/2108.05542 (in Eng.)

Kirmani M., Hakak N., Mohd, M., Mohd, M. (2019) Hybrid Text Summarization: A Survey: Proceedings of SoCTA 2017. 10.1007/978-981-13-0589-4_7 (in Eng.)

Kudo T., Richardson J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — P. 66–71, Brussels, Belgium (in Eng.)

Kynabay B., Aldabergen A., Zhamanov A. (2021) Automatic Summarizing the News from Inform.kz by Using Natural Language Processing Tools. 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 2021. — P. 1-4, doi: 10.1109/SIST50301.2021.9465885 (in Eng.)

Langston O., Ashford B. (2024) Automated Summarization of Multiple Document Abstracts and Contents Using Large Language Models. TechRxiv. August 02, 2024. DOI: 10.36227/techrxiv.172262754.45577350/v1 (in Eng.)

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. (2020) BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — P. 7871–7880 (in Eng.)

Liu X., Zheng Y., Du Z., Ding M., Qian Y., Yang Z., Tang J. GPT understands, too. AI Open. Vol. 5, 2024. — P. 208-215. ISSN 2666-6510, https://doi.org/10.1016/j.aiopen.2023.08.012 (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, abs/1907.11692 (in Eng.)

Oralbekova, D., Mamyrbayev, O., Zhumagulova, S., Zhumazhan, N. (2024) A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach. In: Agarwal, N., Sakalauskas, L., Tukeyev, U. (eds) Modeling and Simulation of Social-Behavioral Phenomena in Creative Societies. MSBC 2024. Communications in Computer and Information Science, vol 2211. Springer, Cham. https://doi.org/10.1007/978-3-031-72260-8_10 (in Eng.)

Qin L., Chen Q., Zhou Y., Chen Z., Li Y., Liao L., Li M., Che W., Yu P. S. (2025) A survey of multilingual large language models. Patterns. — Vol. 6, Issue 1, 2025, 101118, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2024.101118 (in Eng.)

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P.J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research. — Vol. 21, issue 1. — P. 5485-5551 (in Eng.)

Rahali A., Akhloufi M.A. (2023) End-to-End Transformer-Based Models in Textual-Based NLP. AI 2023, 4. — P. 54-110. https://doi.org/10.3390/ai4010004 (in Eng.)

Rani Narejo K., Zan H., Oralbekova D., Parkash Dharmani K., Orken M., Mukhsina K. (2024) "Enhancing Emoji-Based Sentiment Classification in Urdu Tweets: Fusion Strategies With Multilingual BERT and Emoji Embeddings," in IEEE Access, vol. 12. — P. 126587-126600, 2024, doi: 10.1109/ACCESS.2024.3446897 (in Eng.)

Rao A., Aithal S., Singh S. (2025) An Evaluation Metric for Assessing Summary-Level Semantic Similarity in Abstractive Text Summarization. 602-607. 10.1109/AIDE64228.2025.10987460 (in Eng.)

Raza A., Soomro M. H., Salahuddin, Shahzad I., Batool S. (2024) Abstractive Text Summarization for Urdu Language. Journal of Computing & Biomedical Informatics, 7(02). Retrieved from https://jcbi.org/index.php/Main/article/view/596 (in Eng.)

Stein R., Jaques P., Valiati J. (2019) An Analysis of Hierarchical Text Classification Using Word Embeddings. Information Sciences. 471. P. 216-232. 10.1016/j.ins.2018.09.001 (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. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010 (in Eng.)

Winarko E., Tanoto L., Reza M.H. (2025) Indonesian Abstractive Text Summarization Using Stacked Embeddings and Transformer Decoder. IAENG International Journal of Computer Science. Vol. 52, Issue 4, April 2025, pp. 1051-1061 (in Eng.)

Zangooei S., Darmani A., Nezhad H., Mahmoudi L. (2025) ARLED: Leveraging LED-based ARMAN Model for Abstractive Summarization of Persian Long Documents. 10.48550/arXiv.2503.10233 (in Eng.)

Zhabayev T., Tukeyev U. (2021) Development of Technology for Summarization of Kazakh Text. International Journal of Advanced Computer Science and Applications (IJACSA), 12(9), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120914 (in Eng.)

Zulkhazhav A., Kozhirbayev Z., Yessenbayev Z., Sharipbay A. (2019). Kazakh Text Summarization using Fuzzy logic. Computación y Sistemas, 23 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 337–354

https://doi.org/10.32014/2026.2518-1726.416

IRSTI 27.35.33
UDC 665.6: 51–74

© Orazbayev B.B. [1], Zhumadillayeva A.K. [1*], Kurbangalieva N.B. [1], Orazbayeva K.N. [2], 2026.

[1] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2] Esil university, Astana, Kazakhstan.
E-mail: zhumadillayeva_ak@enu.kz

SYNTHESIS OF LINGUISTIC MODELS FOR ASSESSING SULFUR QUALITY AND FUZZY MODELING OF THE SULFUR PRODUCTION PROCESS

Orazbayev Batyr — doctor of technical sciences, professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: batyr_o@mail.ru, https://orcid.org/0000-0003-2109-6999;
Zhumadillayeva Ainur — candidate of Technical Sciences, Associate Professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: zhumadillayeva_ak@enu.kz, https://orcid.org/0000-0003-1042-0415;
Kurbangalieva Nazgul — doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. E-mail: nazgulknb@gmail.com, https://orcid.org/0009-0003-1912-1013;
Orazbayeva Kulman — doctor of technical sciences, professor, Esil University, Astana, Kazakhstan, E-mail: kulman_o@mail.ru, https://orcid.org/0000-0002-1741-7553.

Abstract. Technological objects of various industries belong to complex technological systems, often operating in a fuzzy environment, which significantly complicates the processes of developing their mathematical models. In this regard, currently, solving the problems of synthesizing models of complex, poorly formalized technological production systems in a fuzzy environment is a highly relevant scientific and practical task. In this study, using the example of a sulfur production unit (SPU), the main problems of synthesizing linguistic models of complex technological systems in a fuzzy environment and applying them for fuzzy modeling and optimization of their operating modes are addressed. The main results obtained as a result of the conducted research: a new method for synthesizing effective linguistic models of complex technological systems with fuzzy input and output parameters operating in a fuzzy environment; synthesized linguistic models of the SPU, allowing for modeling and optimization of their operating modes; created rule bases forming linguistic models of the SPU, visualization of the results of fuzzy modeling of the SPU and their interpretation. The study's results were obtained using expert assessment methods and fuzzy

logic rules for conditional inference in fuzzy set theories. Fuzzy Logic Toolbox was used for fuzzy modeling of the sulfur production process. The scientific novelty of the proposed method for synthesizing linguistic models of complex systems in a fuzzy environment lies in the development of methods for modeling poorly formalized complex systems, enabling the modeling and optimization of such systems. The practical significance of the obtained results lies in their ability to effectively model and optimize the operating modes of complex technological systems in various industries in a fuzzy environment.

Keywords: linguistic models; fuzzy modeling; sulfur production unit: decision maker; sulfur production process

Financing. *This research was supported the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. fund this research AP23490206 Development of a decision support system for managing the sulfur production process).*

For citations: Orazbayev B.B., Zhumadillayeva A.K., Kurbangalieva N.B., Orazbayeva K.N. Synthesis of linguistic models for assessing sulfur quality and fuzzy modeling of the sulfur production process. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 337–354. DOI: https://doi.org/10.32014/2026.2518-1726.416

©Оразбаев Б.Б.[1], Жумадиллаева А.К.[1*], Курбангалиева Н.Б.[1], Оразбаева К.Н.[2], 2026.

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Esil university, Астана, Қазақстан.
E-mail: zhumadillayeva_ak@enu.kz

КҮКІРТ САПАСЫН БАҒАЛАУДЫҢ ЛИНГВИСТИКАЛЫҚ МОДЕЛЬДЕРІН СИНТЕЗДЕУ ЖӘНЕ КҮКІРТТІ ӨНДІРУ ПРОЦЕСІН БҰЛЫҢҒЫР МОДЕЛЬДЕУ

Оразбаев Батыр — техника ғылымдарының докторы, профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: batyr_o@mail.ru, https://orcid.org/0000-0003-2109-6999;
Жумадиллаева Айнур — техника ғылымдарының кандидаты, қауымдастырылған профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: zhumadillayeva_ak@enu.kz, https://orcid.org/0000-0003-1042-0415;
Курбангалиева Назгуль — докторант, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: nazgulknb@gmail.com, https://orcid.org/0009-0003-1912-1013;
Оразбаева Кульман — техника ғылымдарының докторы, профессор, Esil university, Астана, Қазақстан,
E-mail: kulman_o@mail.ru, https://orcid.org/0000-0002-1741-7553.

Аннотация. Көптеген өндіріс салаларының технологиялық нысандары әдетте айқын емес ортада жұмысм жасмайтын күрделі технологиялық жүйелерге жататындықтан олардың математикалық модельдерін құру процестері айтарлықтай қиындықтармен сипатталады. Сондықтан қазіргі уақытта кіріс және шығыс параметрлері айқын емес күрделі, нашар формализаланатын технологиялық жүйелерінің модельдерін синтездеу мәселелерін шешу өте өзекті ғылыми және практикалық міндет болып табылады. Бұл зерттеуде күкірт өндіру блогынцың (КӨБ) сияқты күрделі технологиялық жүйелердің лингвистикалық модельдерін айқын емес ортада синтездеу және олардың жұмыс режимдерін айқынсықта модельдеу және оптимизациялау үшін қолданудың негізгі мәселелер зерттеліп, оларды шешу тәсілдемелір ұсынылған. Жүргізілген зерттеулердің негізгі нәтижелері: айқынсыздықпен сипатталатын, кіріс және шығыс параметрлері айқын емес күрделі технологиялық жүйелердің тиімді лингвистикалық модельдерін синтездеудің жаңа әдісі; КӨБ жұмыс режимдерін айқынсыздықта тиімді модельдеуге және оптимизациялауға м,мкіндік беретін оның синтезделген лингвистикалық модельдері; КӨБ лингвистикалық модельдерін қалыптастыратын айқын емес ережелер базасы; КӨБ айқынсыздықта модельдеу нәтижелерін визуализациялануы және олардың интерпретацияланує. Зерттеу нәтижелері эксперттік бағалау тәсілдерін және айқын емес жиындар теорияларын логикалық қорытындылау ережелерін қолдану арқылы алынған, ал күкірт өндіру процесін айқынсыздықта модельдеу үшін MATLAB жүйесінің Fuzzy Logic Toolbox қосымшасы пайдаланылды. Күрделі технологиялық объектілердің лингвистикалық модельдерін айқынсыздықта синтездеудің ұсынылған тәсілінің ғылыми жаңашылдығы мен маңыздылығы айқын емес кіріс және шығыс параметрлері бар нашар формализанған күрделі жүйелерді айқынсыздықта тиімді модельдеуге және оптимизациялауға мүмкіндік беретін тиімді тәсілілін дапмытуда. Алынған нәтижелердің практикалық маңыздылығы – айқын емеса ортада түрлі өндіріс салаларының күрделі технологиялық жүйелері жұмыс режимдерін айқынсыздықта иімді модельдеуге және оптимизациялауға мүмкіндік беретінде.

Түйін сөздер: лингвистикалық модельдер; айқын емес модельдеу; күкірт өндіру блогыі: шешім қабылдаушы тұлға; күкірт өндіру процесі

© Оразбаев Б.Б.[1], Жумадиллаева А.К.[1*], Курбангалиева Н.Б.[1], Оразбаева К.Н.[2], 2026.
[1]Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан;
[2]Esil university, Астана, Казахстан.
E-mail: zhumadillayeva_ak@enu.kz

СИНТЕЗ ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ ОЦЕНКИ КАЧЕСТВА СЕРЫ И НЕЧЕТКОЕ МОДЕЛИРОВАНИЕ ПРОЦЕССА ПРОИЗВОДСТВА СЕРЫ

Оразбаев Батыр — доктор технических наук, профессор, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: batyr_o@mail.ru, https://orcid.org/0000-0003-2109-6999;
Жумадиллаева Айнур — кандидат технических наук, ассоциированный профессор, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: zhumadillayeva_ak@enu.kz, https://orcid.org/0000-0003-1042-0415;
Курбангалиеваазгуль — докторант, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: nazgulknb@gmail.com, https://orcid.org/0009-0003-1912-1013;
Оразбаева Кульман — доктор технических наук, профессор, Университет «Есиль», Астана, Казахстан,
E-mail: kulman_o@mail.ru, https://orcid.org/0000-0002-1741-7553.

Аннотация. Технологические объекты промышленных производств представляют собой сложные системы, функционирующие в условиях неопределённости и нечёткости, что существенно усложняет процессы их математического моделирования. В этой связи задача синтеза моделей сложных, слабо формализуемых технологических систем в нечёткой среде является актуальной научно-практической проблемой. В данном исследовании на примере блока производства серы (БПС) рассматриваются подходы к синтезу лингвистических моделей сложных технологических систем и их применению для нечёткого моделирования и оптимизации режимов работы. В результате исследования предложен новый метод синтеза эффективных лингвистических моделей, предназначенных для описания систем с нечёткими входными и выходными параметрами. Разработаны лингвистические модели БПС, позволяющие осуществлять моделирование и оптимизацию технологических режимов. Сформированы базы правил, лежащие в основе данных моделей, а также реализована визуализация результатов нечёткого моделирования и их интерпретация. Методическая основа исследования включает методы экспертной оценки, аппарат нечётких множеств и нечёткие правила вывода. Для реализации моделей использовано программное средство Fuzzy Logic Toolbox в среде MATLAB. Научная новизна исследования заключается в развитии методов моделирования сложных технологических объектов, функционирующих в условиях

неопределённости, что позволяет повысить точность и эффективность управления такими системами. Практическая значимость работы состоит в возможности применения предложенных моделей для оптимизации режимов работы технологических процессов в различных отраслях промышленности.

Ключевые слова: лингвистические модели; нечеткое моделирование; блок производства серы: лицо, принимающее решение; процесс производства серы

Введение. Качественная сера, получаемая из вредных серосодержащих газов, выделяемых в процессах переработки нефти, и создающие экологические проблемы для нефтеперерабатывающих заводов (НПЗ) является важнейшей продукцией этих НПЗ, позволяющей повысить их экономическую эффективность и решить экологические проблемы. Качественная сера используются для производства медицинских препаратов, порох, взрывчатых веществ, удобрений, резины, красок и других продуктов химического и других производства (Petrova et al. 2019; Kleinjan et al. 2022). В связи с этим эффективное ведение процесса производства качественной серы, получаемая в блоке получения серы (БПС) установки производства серы (УПС) является приоритетным направлением развития нефтеперерабатывающих заводов (НПЗ) (Naumova, 2022). В настоящее время самым эффективным подходом к повышению эффективности процессов производства качественной серы из сероводородов является научно обоснованные методы на основе методов системного анализа, математического моделирования с применением компьютерных технологий (Orazbayev et. al 2020; Medhat et al. 2024).

БПС установки производства серы, предназначенный для получения качественной серы из вредных кислых газов, выделяемых в процессах нефтепереработки, относятся к сложным технологическим системам, в которых протекающие в них процессы характеризуются множеством взаимосвязанных параметров (Edmonson et al. 2019; Ostrovsky et al. 2020). Кроме того, многие сложные технологические системы, как БПС, образующиеся из взаимосвязанных агрегатов, и как правило характеризуются неопределенностью вероятностного и/или нечеткого характера значения параметрах и состояниях. В связи с этим намного усложняются процессы разработки математических моделей таких сложных технологических систем из-за их сложности, неопределенности и нечеткости необходимой информации для эффективного моделирования и оптимизации режимов их работы. Нечеткость доступной информации для оценки качественных показателей серы для синтеза моделей и моделирования процесса производства качественной серы, протекающего в БПС, мотивировали проведения данного исследования, посвященного для синтеза лингвистических моделей для оценки качества серы. Такие лингвистические модели, синтезированные на основе нечетко выраженных опыта, знания и соображения лица, принимающего решения (ЛПР), принимающие решения

по управлению процессом производства серы и экспертов позволяют более адекватно и эффективно моделировать процесса производсва качественноой серы.

Литературный обзор. В последнее время в связи с развитием и достижениями методов теорий нечетких множеств и экспертной оценки, относящихся к методам искусственного интеллекта, задачи и методы моделирования, опьтимизации и управления на основе использования нечеткой информации активно исследуются. Приведем основные результаты обзора и анализа литератур, посвященных к исследованию и решению проблем моделирования сложных технологических систем, в том числе систем производства серы.

Проблемы моделирования сложных трудно формализуемых технологических систем, состоящие из множества компонентов и работающие в условиях различных видов неопределенностей исследованы авторами работ (Protalinskii et al. 2019; Makowski, 2021; Boccara, 2020; Pesci et al. 2023). В перечисленных работах Protalinskii соавторами проанализировали проболемы моделирования сложных многокомпонентных технических, технологических систем с разными типами неопределенностейна основе компонентного подхода и предложили подходы к их решению. Makowski в своей работе привел результаты исследовании проблем структурированного моделирования, позволяющего преодолеть некоторых неопределенностей при решении сложных проблем. В последних двух анализированных работах авторы исследовали основных проблем моделирования сложных систем и оценки неопределенностей, возникающих при их моделирования для решения задач прогнозирования.

Вопросы применения методов искусственного для моделирования различных сложных систем исследованы в следующих (Zhao, 2024; Szczerbicka, 2020; Minullin et al. 2024; Sarker, 2022) и в других работах. Например, в работе Zhao исследованы подходы к решению проблем сложных систем в условиях неопределенности на основе методов искусственного интеллекта. Роль концепций искусственного интеллекта в системном моделировании изучены в работе Szczerbicka и других. Minullin совместно с соавтором проанализировали моделей машинного обучения на основе методов искусственного интеллекта и применение их в задачах образовательной аналитики. Вопросы применения методов искусственного интеллекта для моделирования задач автоматизации и интеллектуальных систем исследованы в работе Sarker.

В анализированных выше и других работах по теме данного исследования в основном исследованы проблемы неопределенности стохастического характера и частичной нечеткости доступной информации. При этом в системах с частичными нечеткими параметрами обычно нечеткими являются только выходные параметры систем, а входные их параметры - четкими. В известных работах для решения проблем неопределенности стохастического

характера исходной информации предлагаются использовать методов теорий вероятностей и математической статистики, а также стохастического программирования. А для решения проблем моделирования систем с нечеткими выходными параметрами используются методы экспертной оценки и нечетких множеств.

Предлагаемые подходы к разработке моделей систем с нечеткими выходными параметрами основаны на использование *α* срез (множества уровня *α)* выходные параметры преобразуются в набор четких параметров. Но в процессе преобразования немалая часть собранной нечеткой информации между *α* срезами, представляющей собой опыт, знания и эрудицию ЛПР, специалистов-экспертов, теряется, что снижает адекватность разрабатываемых моделей. Попытка повышения адекватности разрабатываемых моделей за счет увеличения числа *α* срезов приведет к резкому увеличению размерности задач по разработке моделей и затрат ресурсов, а также значительно снижает эффективность моделирования.

На практике сложные технологические системы типа БПС могут характеризоваться нечеткостью и выходных, режимных параметров, и выходных параметров, т. е. они могут функционировать в условиях полной нечеткости, в нечеткой среде. В анализированных и других известных исследованиях, посвященных к моделированию сложных систем с нечеткими параметрами, не охвачены проблемы разработки моделей и моделирования систем с нечеткими входными и выходными параметрами. В этой связи определена цель исследования данной работы, которая заключается в разработке метода синтеза лингвистических моделей сложных технологических систем с нечеткими входными и входными параметрами, т. е. в нечеткой среде. А также разрабатываемый метод будет реализован на примере синтеза лингвистических для оценки качества серы и нечеткого моделирования процесса производства качественной серы в БПС.

Материалы и методы исследования. В качестве конкретной технологической системы в данной работе исследуется блок производства качественной серы Атырауского НПЗ, который характеризуется нечеткими входными, выходными параметрами и является объектом исследования. Материалами исследования работы являются доступные статистические данные об объекте исследования, от режимных листов его работы и нечеткая информация, получаемая от ЛПР, специалистов-экспертов о нечетких значениях входных параметрах БПС, влияющие и нечетко оцениваемых качественных показателей вырабатываемой серы. В процессе синтеза лингвистических моделей для определения нечетких качественных показателей серы в качестве материалов исследования используются:

- нечеткая информация от ЛПР, специалистов-экспертов, выражающая их знания, опыт и суждения, которая получена на основе методов экспертных оценок (Gutsykova, 2021; Shved, 2020);

- база правил, описывающая различные условия и соответствующие

выводы, которые определяют качественные показатели серы в зависимости от нечетких значений входных6 режимных параметров БПС. База правил создается на основе логических правил условного вывода теорий нечетких множеств (Jørgensen, 2020; Zimmermann, 2023).

На основе нечетких правил условного вывода и различных операций над нечеткими множествами и процедур нечетких множеств предлагается новый метод синтеза лингвистических моделей, позволяющий синтезировать лингвистические модели сложных технологических систем в нечеткой среде. Блок-схема предлагаемого метода синтезе лингвистических моделей сложных технологических систем с нечеткими входными и выходными параметрами представлена ниже на рисунке 1.

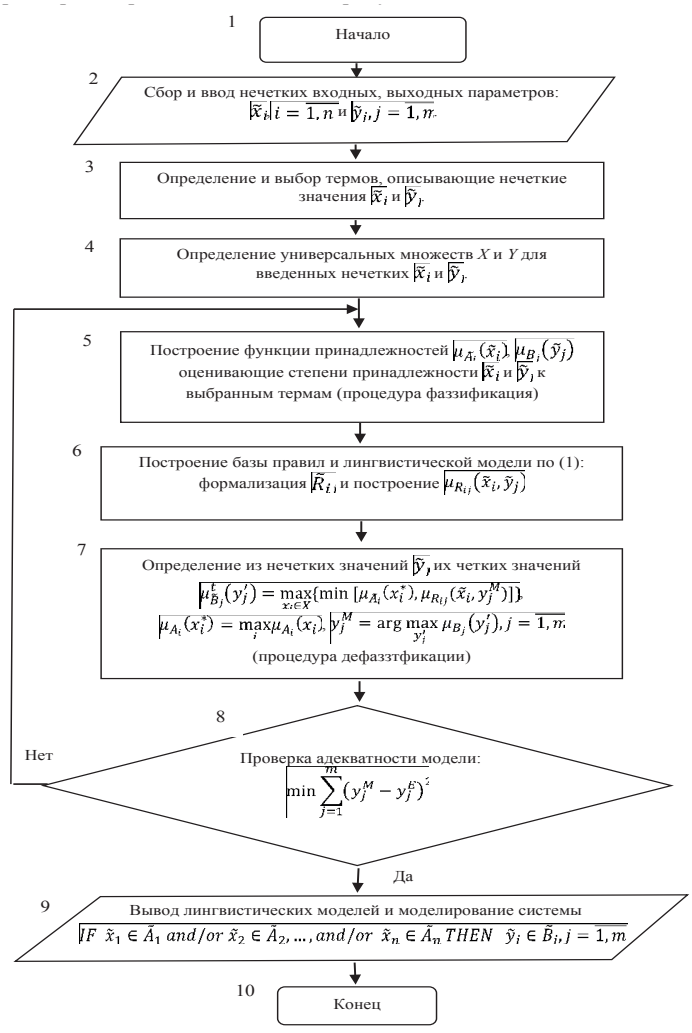


Рисунок 1 - Блок-схема предлагаемого метода синтезе лингвистических моделей сложных технологических систем с нечеткими входными и выходными параметрами

Приведем описание основных блоков разработанного метода синтеза лингвистических моделей сложных систем. В блоке 2 на основе проведенных исследований и экспертной оценки выбираются нечеткие входные, режимные $\tilde{x}_i,\ i=\overline{1,n}$ и выходные $\tilde{y}_j, j=\overline{1,m}$ параметры технологических систем. В этом блоке выбираются те $\tilde{x}_i,\ i=\overline{1,n}$, которые влияют на значения $\tilde{y}_j, j=\overline{1,m}$ технологических систем.

В следующих 3-м и 4-м блоках с участием ЛПР и экспертов определяются терм-множества термы, которые нечетко описывают $\tilde{x}_i$ и $\tilde{y}_j$. Также определяются универсальные множества *X* и *Y*, задающие интервалы изменения входных и выходных параметров.

В 5-м блоке строятся функции принадлежности, которые описывают степени принадлежности термов к нечетким множествам, нечетко описывающих значения $\tilde{x}_i$ и $\tilde{y}_j$, т. е. в этом блоке реализуется процедура фаззификации.

В следующем 6-м блоке метода формализуется база правил, на основе которой синтезируются лингвистические модели, состоящей из логических правил условного вывода, имеющие в общем виде следующую структуру:

$$\overline{IF\ \ \tilde{x}_1 \in \tilde{A}_1\ and/or\ \tilde{x}_2 \in \tilde{A}_2, \dots, and/or\ \ \tilde{x}_n \in \tilde{A}_n}$$

$$\overline{THEN\ \ \tilde{y}_j \in \tilde{B}_j, j=\overline{1,m}} \qquad (1)$$

Таким образом, лингвистические модели нечетко описывающие влияния нечетких входных параметров $\tilde{x}_i, i=\overline{1,n}$ на нечеткие выходные параметры $\tilde{y}_j, j=\overline{1,m}$ синтезируется на основе методов экспертной оценки и нечетких правил условного вывода теорий нечетких множеств. Затем на основе базы правил формализуются нечеткие отображения $\tilde{R}_{ij}$, описывающие нечеткие связи между входными $\tilde{x}_i, i=\overline{1,n}$ и выходными $\tilde{y}_j, j=\overline{1,m}$ параметрами технологических систем. Для удобства нечеткие отображения $\tilde{R}_{ij}$ формализуются через их функции принадлежности $\mu_{\tilde{R}_{ij}}(\tilde{x}_i, \tilde{y}_j)$. Затем на основе композиционного правила вывода $\tilde{B}_j = \tilde{A}_i \circ \tilde{R}_{ij}$, где $\tilde{A}_i \subset X,\ \tilde{B}_j \subset Y$, а *X, Y* - универсальные множества, определенные в блоке 4. Тогда для каждого терма *t* из терм-множеств нечеткое отображение можно определить декартовым произведением соответствующих универсальных множеств: $\tilde{R}_{ij}^t = \tilde{A}_i^t \times \tilde{B}_j^t$.

В блоке 7 для проведения расчетов необходимо на основе функций принадлежностей нечеткого отображения $\tilde{R}_{ij}$ построить матрицы нечетких отношений $\mu_{\tilde{R}_{ij}}(\tilde{x}_i, \tilde{y}_j)$, которую в общем виде на основе операции пересечения можно записать как:

$$\mu_{\tilde{R}_{ij}}^t(\tilde{x}_i, \tilde{y}_j) = min\left[\mu_{\tilde{A}_i}^t(x_i), \mu_{\tilde{B}_j}^t(y_j)\right], i=\overline{1,n}$$

На основе данного правила можно определить значения функции принадлежностей выходных параметров по формуле:

$$\mu_{\tilde{B}_j}^{t}(y_j') = \max_{x_i \in X}\{\min\left[\mu_{\tilde{A}_i}(x_i^*), \mu_{R_{ij}}(\tilde{x}_i, y_j^M)\right]\}. \quad (2)$$

В формуле (2) через $\overline{x_i^*}$ - обозначены оцененные экспертами нечеткие значения режимных параметров. Тогда искомое множество, в которому принадлежат текущие значения режимных параметров, определяется по формуле $\mu_{A_i}(x_i^*) = \max_i \mu_{A_i}(x_i)$, т. е. как множество, в котором значения режимных параметров имеют максимального значения функции принадлежности.

Для дефаззификации результатов, т. е. определения числовых значений выходных параметров $\overline{y_i^M}$ из множества нечетких решений можно применить следующего выражения: $\overline{y_j^M = \arg\max_{y_j'} \mu_{B_j}(y_j'), j = \overline{1,m}}$. Таким образом числовые значения выходных параметров выбираются как аргумент максимального значения функции принадлежности выходных параметров.

В 8-м блоке проверяется условие адекватности полученной модели:

$$\min \sum_{j=1}^{m} \left(y_j^M - y_j^E\right)^2 \le S_R,$$

где y_i^M - дефазифицированные значения выходных параметров технологической системы; y_i^E - реальные значения выходных параметров, полученные экспериментальным путем на исследуемой технологической системы; S_R - заданный уровень адекватности. Значения $\overline{y_i^M}$ и y_i^E должны быть получены при одинаковых значениях $x_i^*, i = \overline{1,n}$.

При выполнении условия блока 8 в блоке 10 выводится синтезированные лингвистические модели с необходимой адекватностью. Если условие не выполняются, то начиная с 5-го блока начинается новая итерация синтеза более адекватных моделей.

Результаты исследования. На основе предложенного в предыдущем разделе метода синтеза лингвистических моделей построены лингвистические модели, оценивающие основного качественного показателя серы - массовую долю серы в полученной на выходе БПС, позволяющие нечетко моделировать в приложения Fuzzy Logic Toolbox (Servin, 2023).

Согласно предложенного метода с использованием результатов исследования и экспертной оценки основными входными, режимными параметрами, наиболее влияющие на $\tilde{y}_1$ - массовую долю серы в получаемой на выходе БПС сере определены: $\tilde{x}_1, \tilde{x}_2$ – соответственно температура и давление в реакторах Cold Bed Absorption (CBA) БПС, характеризуемые нечеткостью.

Определены следующие терм-множества, нечетко описывающие входные и выходные параметры (реализация блока 3):

$\tilde{x}_1$={«низкая», «ниже средней», «средняя», «выше средней», «высокая»};

$\tilde{x}_2$={«низкое», «ниже нормы», «нормальное», «выше нормы», «высокое»};

$\tilde{y}_1$={«очень низкая», «низкая», «средняя», «высокая», «очень высокая»} (блоки 3.4).

Далее построения функции принадлежностей нечетких параметров $\tilde{x}_1, \tilde{x}_2$ и $\tilde{y}_1$ введены сокращенные обозначения термов, которые их описывают (таблица 1)

Таблица 1 - Термы лингвистических переменных и их сокращенные обозначения для построения их функции принадлежностей и формирования базы правил

Термы нечетких параметров	Обозначение
Низкий (Low)	*LW*
Ниже среднего (Below average)	*BA*
Средний (Average)	*AR*
Выше среднего (Above average)	*AA*
Высокий (High)	*HG*
Ниже нормы (Below normal)	*BN*
Нормальный (Normal)	*NR*
Выше нормы (Above normal)	*AN*
Очень низкий (Very low)	*V LW*
Очень высокий (Very high)	*VHG*

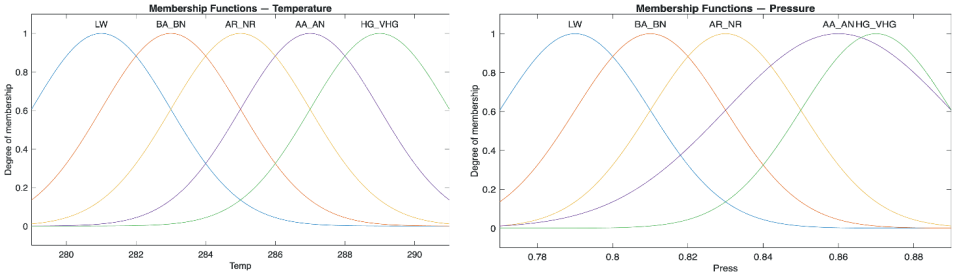
Универсальные множества переменных $\tilde{x}_1, \tilde{x}_2$ и $\tilde{y}_1$ необходимые для построения функции принадлежности (блок 4) приведены в таблице 2.

Таблица 2 - Универсальные множества для $\tilde{x}_1$, $\tilde{x}_2$ и $\tilde{y}_1$.

Нечеткие входные параметры (лингвистические переменные)	Значения нечетких входных параметров				
	LW	*BA, BN*	*AR, NR*	*AA, AN*	*HG*
$\tilde{x}_1$ - температура в реакторах СВА	279–283	281–285	283–287	285–289	287–291
$\tilde{x}_2$ - давление в реакторах СВА	0,77–0,81	0,79–0,83	0,81–0,85	0,83–0,89	0,85–0,89
Нечеткие выходные параметры	Значения нечетких входных переменных				
	V LW	*LW*	*AR*	*HG*	*VHG*
$\tilde{y}_1$ - массовая доля серы	99.80–99.85	99.83–99.88	99.86–99.94	99.92-99.97	99.94–99.99

Далее приведем результаты реализации процедуры фаззификации и других процедур нечеткого вывода с использованием инструментарий Fuzzy Logic Toolbox системы MATLAB.

Ниже на рисунке 2 a), b) приведены функции принадлежности гауссовского типа для нечетких входных параметров $\tilde{x}_1$, и $\tilde{x}_2$, а на рисунке 3 - для выходного параметра $\tilde{y}_1$, построенные с применением Fuzzy Logic Toolbox.



a) b)

Рисунок 2 - Функции принадлежности нечетких входных параметров a) $\tilde{x}_2$ - температура в реакторах СВА; b) $\tilde{x}_2$ - давление в реакторах СВА

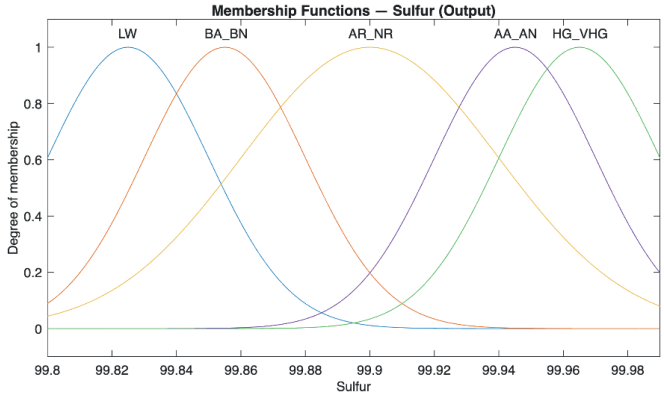


Рисунок 3 - Функции принадлежности выходного параметра $\tilde{y}_1$- массовая доля серы производимой серы

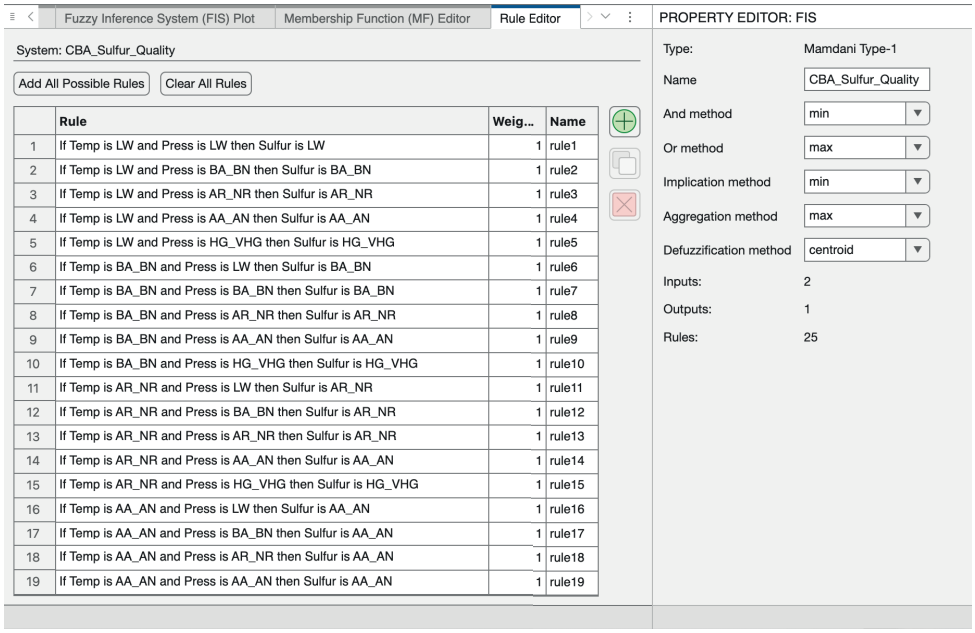
Созданная база правил для системы нечеткого вывода, позволяющая оценить качества серы от нечетких входных параметров представлены ниже в виде правил нечетких продукций.

Rule 1: *IF* «$\tilde{x}_1$ *is LW*» *and* «$\tilde{x}_2$ *is LW*» *THEN* «$\tilde{y}_1$ *is VLW*» F_1;
Rule 2: *IF* «$\tilde{x}_1$ *is LW*» *and* «$\tilde{x}_2$ *is BN*» *THEN* «$\tilde{y}_1$ *is VLW*» F_2;
Rule 3: *IF* «$\tilde{x}_1$ *is BA*» *and* «$\tilde{x}_2$ *is LW*» *THEN* «$\tilde{y}_1$ *is LW*» F_3;
Rule 4: *IF* «$\tilde{x}_1$ *is BA*» *and* «$\tilde{x}_2$ *is BN*» *THEN* «$\tilde{y}_1$ *is VLW*» F_4;
Rule 5: *IF* «$\tilde{x}_1$ *is LW*» *and* «$\tilde{x}_2$ *is NR*» *THEN* «$\tilde{y}_1$ *is VLW*» F_5;
Rule 6: *IF* «$\tilde{x}_1$ *is AR*» *and* «$\tilde{x}_2$ *is LW*» *THEN* «$\tilde{y}_1$ *is LW*» F_6;
Rule 7: *IF* «$\tilde{x}_1$ *is BA*» *and* «$\tilde{x}_2$ *is NR*» *THEN* «$\tilde{y}_3$ *is LW*» F_7;
Rule 8: *IF* «$\tilde{x}_1$ *is AR*» *and* «$\tilde{x}_2$ *is BN*» *THEN* «$\tilde{y}_1$ *is LW*» F_8;
Rule 9: *IF* «$\tilde{x}_1$ *is AR*» *and* «$\tilde{x}_2$ *is NR*» *THEN* «$\tilde{y}_1$ *is LW*» F_9;
Rule 10: *IF* «$\tilde{x}_1$ *is LW*» *and* «$\tilde{x}_2$ *is AA*» *THEN* «$\tilde{y}_1$ *is LW*» F_{10};
Rule 11: *IF* «$\tilde{x}_1$ *is AA*» *and* «$\tilde{x}_2$ *is LW*» *THEN* «$\tilde{y}_1$ *is AR*» F_{11};
Rule 12: *IF* «$\tilde{x}_1$ *is BA*» *and* «$\tilde{x}_2$ *is AA*» *THEN* «$\tilde{y}_1$ *is AR*» F_{12};
Rule 13: *IF* «$\tilde{x}_1$ *is AA*» *and* «$\tilde{x}_2$ *is BN*» *THEN* «$\tilde{y}_1$ *is AR*» F_{13};
Rule 14: *IF* «$\tilde{x}_1$ *is AR*» *and* «$\tilde{x}_2$ *is AA*» *THEN* «$\tilde{y}_1$ *is AR*» F_{14};
Rule 15: *IF* «$\tilde{x}_1$ *is AA*» *and* «$\tilde{x}_2$ *is NR*» *THEN* «$\tilde{y}_1$ *is AR*» F_{15};
Rule 16: *IF* «$\tilde{x}_1$ *is AA*» *and* «$\tilde{x}_2$ *is AA*» *THEN* «$\tilde{y}_1$ *is HG*» F_{16};

Rule 17: *IF* «$\tilde{x}_1$ *is LW*» *and* «$\tilde{x}_2$ *is HG*» *THEN* «$\tilde{y}_1$ *is HG*» F_{17};
Rule 18: *IF* «$\tilde{x}_1$ *is HG*» *and* «$\tilde{x}_2$ *is LW*» *THEN* «$\tilde{y}_1$ *is HG*» F_{18};
Rule 19: *IF* «$\tilde{x}_1$ *is BA*» *and* «$\tilde{x}_1$ *is HG*» *THEN* «$\tilde{y}_1$ *is HG*» F_{19};
Rule 20: *IF* «$\tilde{x}_1$ *is HG*» *and* «$\tilde{x}_2$ *is BN*» *THEN* «$\tilde{y}_3$ *is HG*» F_{20};
Rule 21: *IF* «$\tilde{x}_1$ *is AR*» *and* «$\tilde{x}_2$ *is HG*» *THEN* «$\tilde{y}_1$ *is VHG*» F_{21};
Rule 22: *IF* «$\tilde{x}_1$ *is HG*» *and* «$\tilde{x}_2$ *is NR*» *THEN* «$\tilde{y}_1$ *is VHG*» F_{22};
Rule 23: *IF* «$\tilde{x}_1$ *is AA*» *and* «$\tilde{x}_2$ *is HG*» *THEN* «$\tilde{y}_1$ *is VHG*» F_{23};
Rule 24: *IF* «$\tilde{x}_1$ *is HG*» *and* «$\tilde{x}_2$ *is AA*» *THEN* «$\tilde{y}_1$ *is VHG*» F_{24};
Rule 25: *IF* «$\tilde{x}_1$ *is HG*» *and* «$\tilde{x}_2$ *is HG*» *THEN* «$\tilde{y}_1$ *is VHG*» F_{25};

В приведенных правилах F_1,...,F_{25} - весовые коэффициенты, отражающие степень уверенности в истинности подзаключений, которые принимают значения на интервале от 0 до 1. В нашей задаче все они равны 1.

Приведенная нечеткая база знаний реализована с помощью Fuzzy Logic Toolbox и представлена на рисунке 4.



	Rule	Weig...	Name
1	If Temp is LW and Press is LW then Sulfur is LW	1	rule1
2	If Temp is LW and Press is BA_BN then Sulfur is BA_BN	1	rule2
3	If Temp is LW and Press is AR_NR then Sulfur is AR_NR	1	rule3
4	If Temp is LW and Press is AA_AN then Sulfur is AA_AN	1	rule4
5	If Temp is LW and Press is HG_VHG then Sulfur is HG_VHG	1	rule5
6	If Temp is BA_BN and Press is LW then Sulfur is BA_BN	1	rule6
7	If Temp is BA_BN and Press is BA_BN then Sulfur is BA_BN	1	rule7
8	If Temp is BA_BN and Press is AR_NR then Sulfur is AR_NR	1	rule8
9	If Temp is BA_BN and Press is AA_AN then Sulfur is AA_AN	1	rule9
10	If Temp is BA_BN and Press is HG_VHG then Sulfur is HG_VHG	1	rule10
11	If Temp is AR_NR and Press is LW then Sulfur is AR_NR	1	rule11
12	If Temp is AR_NR and Press is BA_BN then Sulfur is AR_NR	1	rule12
13	If Temp is AR_NR and Press is AR_NR then Sulfur is AR_NR	1	rule13
14	If Temp is AR_NR and Press is AA_AN then Sulfur is AA_AN	1	rule14
15	If Temp is AR_NR and Press is HG_VHG then Sulfur is HG_VHG	1	rule15
16	If Temp is AA_AN and Press is LW then Sulfur is AA_AN	1	rule16
17	If Temp is AA_AN and Press is BA_BN then Sulfur is AA_AN	1	rule17
18	If Temp is AA_AN and Press is AR_NR then Sulfur is AA_AN	1	rule18
19	If Temp is AA_AN and Press is AA_AN then Sulfur is AA_AN	1	rule19

Рисунок 4 - Нечеткая база правил оценки качества серы

Результаты визуализации нечеткого логического вывода в RuleViewer приведены на рисунке 5. В поле Input указаны значения входных переменных, для которых выполняется логический вывод.

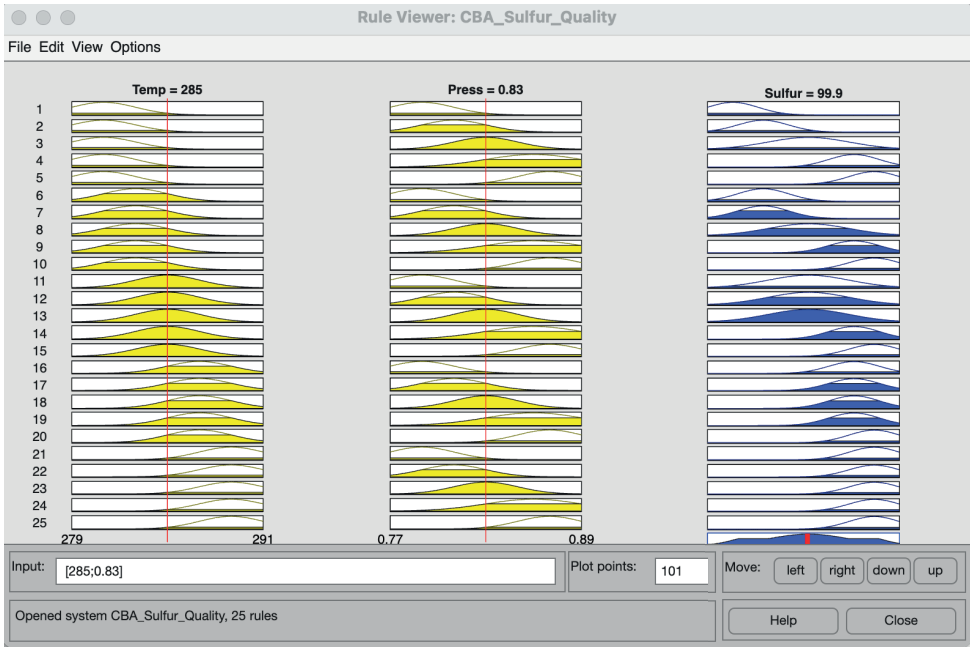


Рисунок 5 - Визуализация нечеткого логического вывода в RuleViewer

Поверхность “входы-выход”, которая соответствует полученной нечеткой системе, которая позволяет визуально просмотреть результаты нечеткого моделирования приведена на рисунке 6.

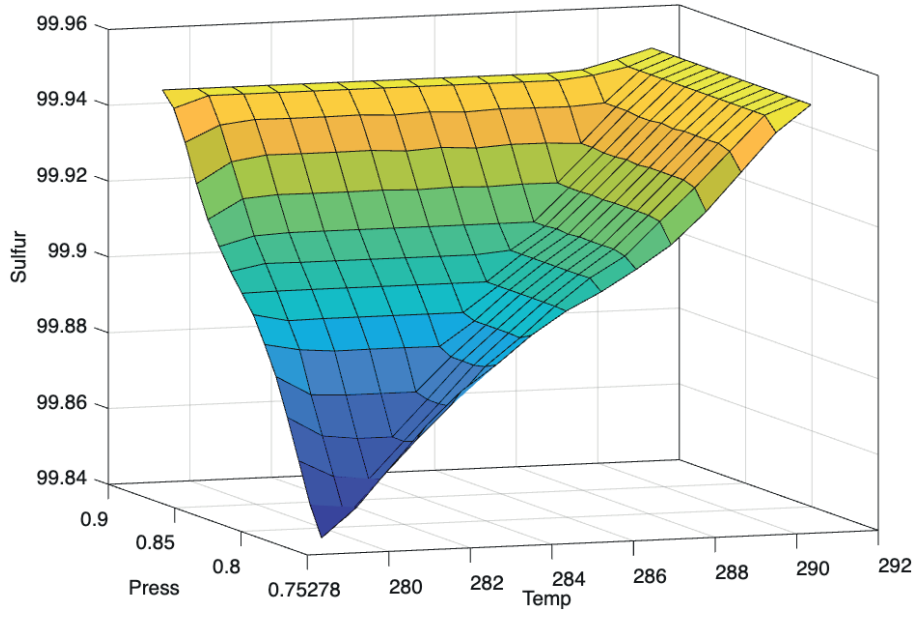


Рисунок 6 - Поверхность “входы-выход” в окне SurfaceViwer

Обсуждение результатов. Предложенный метод синтеза лингвистических моделей сложных технологических систем с нечеткими входными и выходными параметрами основан на использование нечеткой информации и нечетких правил условного вывода и различных операций над нечеткими множествами и процедур нечетких множеств. Разработанный метод синтеза лингвистических моделей на основе экспертной оценки и методов теорий нечетких множеств позволяет синтезировать лингвистических моделей сложных систем и эффективно моделировать их в нечеткой среде.

Лингвистические модели, оценивающие массовую долю серы с выхода реакторов СВА, синтезированные с помощью предложенного метод синтеза лингвистических моделей, позволяют нечетко моделировать процесса производства серы в Fuzzy Logic Toolbox и выбрать оптимальный режим работы реакторов СВА. Как видно из результатов нечеткого моделирования, из поверхности «входы-выход» в окне SurfaceViwer (рисунок 6): чем меньше температура и давление реакторов, тем больше доля серы в выработанной серы, т. е. качество серы улучшается и наоборот. Это соответствует разработанной выше базе правил и означает, что необходимо выбрать компромиссное решение при выборе значений входных параметров реакторов СВА, так как они с одной стороны позволяет уменьшает объема серы, с другой стороны улучшает ее качество. Для этого придется решать задачу принятия решений по выбору наилучшего решение, в зависимости от спроса на объем серы и на ее качество, т. е. компромиссного решения.

Результаты визуализации нечеткого вывода в Rule Viewer (рисунок 5) позволяет провести обзор правил для нечеткого вывода каждого правила, результирующего нечеткого множества и реализации процедуры дефаззификации. На рисунке 5 приведен результаты моделирования в случае, когда введены входные параметры: значения температуры реакторов СВА 287^{o}C и давления 0,85 кг/см^{2}. Тогда в результате нечеткого моделирования массовая доля серы в выработанной сере на входе БПС, т. е. ее основная качественная показатель равна 99.97. Это означает, что качество серы очень высокое, так как если массовая доля серы в полученной сере выше 99.94 ее качества по стандарту считается очень высоким. Таким образом с помощью данного окна можно, меняя нечетких значений входных параметров реакторов СВА, т. е. моделируя их режимы работы, оценить и выбрать наилучшее значения качества серы на выходе БПС.

Заключение. В статье разработан новый метод синтез лингвистических моделей сложных технологических систем, позволяющие эффективно моделировать их в нечеткой среде. Разработанный метод реализован для синтеза лингвистических моделей оценки качества серы и нечеткого моделирования процесса производства качественной серы в блоке получения серы Атырауского НПЗ.

В качестве основных результатов исследования можно выделить:

- разработанный метод синтез лингвистических моделей сложных, плохо

формализуемых технологических систем в нечеткой среде для нечеткого моделирования и опьтимизации режимов их работы;

- синтезированные лингвистические модели на основе предложеенного метода, позволяющие оценить качества серы и эффективно моделировать, а также оптимизировать процесса производства качественной серы в БПС в нечеткой среде;

- созданные базы правил на основе экспертной оценки и реализованные в приложении Fuzzy Logic Toolbox процедуры фаззификации, дефаззификации, нечеткое моделирование и визуализация их результатов.

Новизна результатов исследования заключается в создании нового метода синтеза лингвистических моделей сложных технологических систем с нечеткими входными и выходными параметрами, позволяющего эффективно моделировать и оптимизировать режимов их работы в нечеткой среде.

Дальнейшим направлением развития данного исследования является разработка эвристических алгоритмов принятия решений, которые на основе созданных лингвистических моделей БПС позволяют автоматизированном режиме принимать решение по эффективному управлению процессами производства качественной серы.

References

Boccara N. (2020) Modeling Complex Systems. Springer. -New York Dordrecht Heidelberg London. — 157 p. https://doi.org/10.1007/978-1-4419-6562-2 (in Eng.).

Edmonson N. (2020) Industrial sulfur demand: Analysis and prospect. Nat Resour Res. –Vol. 3. — P.205–215 https://doi.org/10.1007/BF02259046 (in Eng.).

Jørgensen B. (2020) The rules of conditional inference: Is there a universal definition of nonformation? Journal It. Statist. Soc. — Vol. 3. — P. 355–384. https://doi.org/10.1007/BF02589024 (in Eng.).

Гуцыкова С. (2021) Метод экспертных оценок: Теория и практика: –М.: Когито-Центр. —147 с. (in Eng.).

Kleinjan W.E., Keizer A., Janssen, A.J. (2022) Biologically Produced Sulfur. In: Steudel, R. (eds) Elemental Sulfur and Sulfur-Rich Compounds I. Topics in Current Chemistry, Springer, Berlin, Heidelberg. — Vol. 230. — P.167-188. https://doi.org/10.1007/b12114 (in Eng.).

Makowski M. (2021) Structured Modeling for Coping with Uncertainty in Complex Problems. In: Coping with Uncertainty. Lecture Notes in Economics and Mathematical Systems. Springer, Berlin, Heidelberg. — Vol. 581. — P.47-64. https://doi.org/10.1007/3-540-35262-7_3 (in Eng.).

Medhat A., Shehata W., Gad F. et al. (2024) Process simulation, optimization, and cost analysis of a proposed sulfur recovery unit by applying modified Claus technology. J. Eng. Appl. Sci. — Vol. 71, No 109. — P.25–37. https://doi.org/10.1186/s44147-024-00437-3 (in Eng.).

Minullin D.A., Gafarov F.M. (2024) Analyzing Machine Learning Models Based on Explainable Artificial Intelligence Methods in Educational Analytics. Autom. Doc. Math. Linguist. — Vol. 58. — P.115–122. https://doi.org/10.3103/S0005105525700189 (in Eng.).

Naumova V.V. (2022) Sulfur production from sulfur-containing gases. International Journal of Applied Sciences and Technologies. — No 4. — P.285-297. (in Eng.).

Orazbayev B.B., Kenzhebaeva T.S., Abitova G.A., Orazbayeva K.N., Spichak Y.Y. (2020) Study and Design of Mathematical Models for Chemical-Technological Systems under Conditions of Uncertainty Based on the Systems Analysis. Proceeding of the International Symposium on Industrial Engineering and Operations Management (IEOM). Bristol, UK. July 24–25, 140528. — P.776–790.

Ostrovsky G.M., Lapteva T.V., Ziyatdinov N.N. (2020) Design of Chemical Engineering Systems with Chance Constraints. Theor Found Chem Eng. –Vol. 51. — No 5. — P. 961–971. https://doi.org/10.1134/S0040579517060136. (in Eng.).

Petrova A., Juraeva M. (2023) Primenenie sery pri poluchenii vazhnyh produktov razlichnyh otraslej promyshlennosti [The use of sulfur in the production of important products of various industries]. Bulletin of Tomsk University. — Volume 55. — pp.25-37. (in Russ.).

Pesci M.H., Mouris K., Haun S. (2023) Assessment of uncertainties in a complex modeling chain for predicting reservoir sedimentation under changing climateю Model. Earth Syst. Environ. — Vol. 9. — P. 3777–3793. https://doi.org/10.1007/s40808-023-01705-6 (in Eng.).

Protalinskii O.M. (2019) Analysis and Modelling of Complex Engineering Systems Based on the Component Approach. World Applied Sciences Journal. — Vol. 24. No 2. — P.276–283. (in Eng.).

Sarker I.H. (2022) AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. SN Computer Science. — Vol. 3. — P. 158-170. https://doi.org/10.1007/s42979-022-01043-x (in Eng.).

Servin C., Becker B.A., Eaton E., Kumar A. (2023) Fuzzy Logic: Towards Developing Fuzzy Education Curricula Using ACM/IEEE/AAAI CS2023. In: Cohen, K., Ernest, N., Bede, B., Kreinovich, V. (eds) Fuzzy Information Processing. NAFIPS 2023. Lecture Notes in Networks and Systems, Springer, Cham. — Vol 751. — P.184-193. https://doi.org/10.1007/978-3-031-46778-3_17 (in Eng.).

Shved A., Kovalenko I., Davydenko Y. (2020) Method of Detection the Consistent Subgroups of Expert Assessments in a Group Based on Measures of Dissimilarity in Evidence Theory. In: Shakhovska, N., Medykovskyy, M.O. (eds) Advances in Intelligent Systems and Computing IV. CSIT 2019. Advances in Intelligent Systems and Computing, Springer, Cham. —Vol 1080. — P. 36-53. https://doi.org/10.1007/978-3-030-33695-0_4 (in Eng.).

Szczerbicka H., Uthmann T. (2020) The Role of Artificial Intelligence Concepts in System Modelling and Simulation: An Overview. In: Wolfinger, B. (eds) Innovationen bei Rechen- und Kommunikationssystemen. Informatik aktuell. Springer, Berlin, Heidelberg. — P. 390–397. https://doi.org/10.1007/978-3-642-51136-3_51 (in Eng.).

Zhao T. (2024) Artificial Intelligence in Mathematical Modelling of Complex Systems. EAI Endorsed Transactions on e-Learning. –Vol. 10. No 3. — P. 2-12. (in Eng.).

Zimmermann H.-J. (2023) Fuzzy Set Theory – and Its Applications: Springer Science+Business Media, LLC. Fifth. Edition. — 527 p. https://doi.org/10.1007/978-94-010-0646-0 (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 355–372

https://doi.org/10.32014/2026.2518-1726.417

IRSTI 62.50.43
UDC 004.5

© Sarsenbayeva A.K.[1]*, Rakhimova D.R.[1], Shormakova A.N.[1], Mansurova M.E.[1], Adali E.[2], 2026.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[21]stanbul Technical University, Istanbul, Turkey.
E-mail: as.sarsenbayeva@gmail.com

APPLICATION OF SEMANTIC METHODS IN THE FIELD OF LEGISLATION: AN INTELLECTUAL SYSTEM FOR ANALYSIS OF AGGLUTINATIVE TEXTS

Sarsenbayeva Assiya — PhD student of Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: as.sarsenbayeva@gmail.com, https://orcid.org/0009-0008-0053-1182;
Rakhimova Diana — PhD, Associate Professor of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Shormakova Assem — PhD, acting Associate Professor of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: shormakovaassem@gmail.com, https://orcid.org/0000-0002-1637-4643;
Mansurova Madina — PhD, acting Associate Professor of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: m_mansurova@kaznu.kz, https://orcid.org/0000-0001-6284-8283;
Adali Esref — PhD, Professor of Istanbul Technical University, Istanbul, Turkey,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255.

Abstract. This article presents a semantic method for determining the similarity of terms, phrases, and short expressions in agglutinative languages, particularly Kazakh and Turkish. Semantic similarity plays a key role in many natural language processing tasks, including information retrieval, question–answering systems, text classification, and legal document analysis. However, the development of such systems for agglutinative and low-resource languages remains challenging due to complex morphology, limited linguistic resources, and insufficient annotated corpora. The proposed approach is based on the Ri coefficient, which evaluates semantic similarity using frequency relationships between words within their contextual environments. The method analyzes contextual sets of words that frequently co-occur with target terms and calculates normalized proximity measures between them. By integrating contextual frequency information and

synonym relations, the approach allows for a more accurate evaluation of semantic relationships between legal terms and expressions. To assess the effectiveness of the proposed method, a comparative study was conducted with widely used similarity metrics, including Jaccard similarity, Cosine similarity, and Normalized Google Distance. Experimental evaluation was carried out on a corpus of legal texts in the Kazakh language containing approximately 10,000 legal sentences and a specialized legal dictionary consisting of 8,400 terms and about 10,000 synonyms. Additional frequency data were obtained from open legal resources such as the Adilet legal information system. The findings confirm that the proposed semantic similarity approach is effective for processing legal texts in agglutinative languages and can be integrated into intelligent legal information systems, including question–answering systems and legal document retrieval platforms. Furthermore, the method has the potential to be adapted for other Turkic and low-resource languages with similar morphological characteristics.

Keywords: semantic similarity, synonyms, Kazakh language, Turkish language, legal terms, agglutinative languages

For citations: Sarsenbayeva A.K., Rakhimova D.R., Shormakova A.N., Mansurova M.E., dali E. Application of semantic methods in the field of legislation: an intellectual system for analysis of agglutinative texts. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 355–372. DOI: https://doi.org/10.32014/2026.2518-1726.417

© Сарсенбаева А.К. [1] *, Рахимова Д.Р.[1], Шормакова А.Н. [1], Мансурова М.Е.[1], Адали Э.[2], 2026.

[1] Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2] Стамбул Техникалық университеті, Стамбул, Түркия.
E-mail: as.sarsenbayeva@gmail.com

СЕМАНТИКАЛЫҚ ӘДІСТЕРДІ ЗАҢНАМА САЛАСЫНДА ҚОЛДАНУ: АГГЛЮТИНАТИВТІ МӘТІНДЕРДІ ТАЛДАУҒА АРНАЛҒАН ИНТЕЛЛЕКТУАЛДЫ ЖҮЙЕ

Сарсенбаева Асия — Әл-Фараби атындағы Қазақ ұлттық университетінің PhD докторанты, Алматы, Қазақстан, Email: as.sarsenbayeva@gmail.com, https://orcid.org/0009-0008-0053-1182;
Рахимова Диана — PhD, доцент, Әл-Фараби атындағы Қазақ ұлттық университетінің профессоры, Алматы, Қазақстан,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Шормакова Асем — PhD, Әл-Фараби атындағы Қазақ ұлттық университетінің қауымдастырылған профессор міндетін атқарушы, Алматы, Қазақстан,
Email: shormakovaassem@gmail.com, https://orcid.org/0000-0002-1637-4643;
Мансурова Мадина — PhD, Әл-Фараби атындағы Қазақ ұлттық университетінің профессоры, Алматы, Қазақстан, Email: m_mansurova@kaznu.kz, https://orcid.org/0000-0001-6284-8283;
Адали Эшреф — PhD, Стамбул Техникалық университетінің профессоры, Стамбул, Түркия, Email: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255.

Аннотауция. Бұл мақалада агглютинативті тілдердегі, оның ішінде қазақ және түрік тілдеріндегі терминдер, сөз тіркестері және қысқа сөз тіркестерінің семантикалық ұқсастығын анықтаудың әдісі ұсынылған. Семантикалық ұқсастық ақпаратты іздеу, сұрақ-жауап жүйелері, мәтінді жіктеу және құқықтық құжаттарды талдау сияқты көптеген табиғи тілді өңдеу тапсырмаларында маңызды рөл атқарады. Дегенмен, агглютинативті және ресурстары шектеулі тілдер үшін мұндай жүйелерді әзірлеу олардың күрделі морфологиясына, тілдік ресурстарының шектеулілігіне және аннотацияланған корпустардың аздығына байланысты қиындық тудыруда. Семантикалық ұқсастықты есептеу әдісі Ri коэффициентіне негізделген, ол терминдер мен сөз тіркестерінің контексттеріндегі жиілік қатынастарын ескере отырып есептеледі. Бұл коэффициент екі сөз арасындағы байланысты бағалауға мүмкіндік береді, және ұсынылған тәсіл сұрақ-жауап жүйелеріне интеграцияланған, мұнда бірнеше ұқсас сөз тіркестерінің нұсқалары берілгенде, ең тиісті және дұрыс жауап таңдалуы қажет. NGD, Jaccard және Cosine сияқты түрлі әдістер қарастырылып, олардың салыстырмалы зерттеуі жүргізілген. Алынған эксперименттік нәтижелер көрсеткендей, ұсынылған әдіс терминдерді анықтауда және таңдауда жоғары дәлдік көрсетеді, бұл әсіресе агглютинативті тілдердегі заңдық лексика саласында маңызды болып табылады. Эксперименттік бағалау шамамен 10 000 заң сөйлемінен тұратын қазақ тіліндегі заң мәтіндерінің корпусында және 8 400 термин мен шамамен 10 000 синонимнен тұратын мамандандырылған заң сөздігінде жүргізілді. Нәтижелер семантикалық ұқсастыққа ұсынылған тәсілдің агглютинативті тілдердегі заң мәтіндерін өңдеу үшін тиімді екенін және сұрақ-жауап жүйелері мен заң құжаттарын іздеу платформаларын қоса алғанда, интеллектуалды құқықтық ақпараттық жүйелерге біріктірілуі мүмкін екенін растайды. Сонымен қатар, әдісті ұқсас морфологиялық сипаттамалары бар басқа түркі және аз зерттелген тілдерге бейімдеу мүмкіндігі бар.

Түйін сөздер: семантикалық ұқсастық, синонимдер, қазақ тілі, түрік тілі, заң терминдері, агглютинативті тілдер

Алғыс**.** *Бұл зерттеу Қазақстан Республикасының жоғары білімі және Ғылым министрлігінің қолдауымен BR24993001 жобасымен қаржыландырылды.*

© Сарсенбаева А.К.[1]*, Рахимова Д.Р.[1], Шормакова А.Н.[1], Мансурова М.Е.[1], Адали Э.[2] , 2026.
[1]Казахский Национальный университет имени аль-Фараби, Алматы, Казахстан;
[2]Стамбульский Технический Университет, Стамбул, Турция.
E-mail: as.sarsenbayeva@gmail.com

ПРИМЕНЕНИЕ СЕМАНТИЧЕСКИХ МЕТОДОВ В ОБЛАСТИ ЗАКОНОДАТЕЛЬСТВА: ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ДЛЯ АНАЛИЗА АГГЛЮТИНАТИВНЫХ ТЕКСТОВ

Сарсенбаева Асия — докторант Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: as.sarsenbayeva@gmail.com, https://orcid.org/0009-0008-0053-1182;
Рахимова Диана — PhD, ассоциированный профессор Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Шормакова Асем — PhD, и.о. доцента Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: shormakovaassem@gmail.com, https://orcid.org/0000-0002-1637-4643;
Мансурова Мадина — PhD, профессор Казахского Национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: m_mansurova@kaznu.kz, https://orcid.org/0000-0001-6284-8283;
Адали Эшреф — PhD, профессор в Стамбульском Техническом Университете, Стамбул, Турция,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255.

Аннотация. В статье представлен семантический метод оценки сходства терминов, фраз и коротких выражений в агглютинативных языках, в частности в казахском и турецком. Семантическое сходство играет ключевую роль в задачах обработки естественного языка, включая информационный поиск, системы вопросов и ответов, классификацию текстов и анализ юридических документов. Разработка подобных систем для агглютинативных и малоресурсных языков остаётся сложной задачей из-за их морфологической сложности, ограниченности лингвистических ресурсов и недостаточного объёма аннотированных корпусов. Предложенный подход основан на коэффициенте R_i, который позволяет оценивать семантическое сходство на основе частотных соотношений слов в их контекстном окружении. Метод контекстного анализа формирует множества слов, совместно встречающихся с целевыми терминами, и вычисляет нормализованные меры близости между ними. Интеграция контекстной частотной информации и синонимических связей обеспечивает более точную оценку семантических отношений между юридическими терминами и выражениями. Для оценки эффективности метода проведено сравнительное исследование с использованием распространённых метрик сходства, включая коэффициент Жаккара, косинусное сходство и нормализованное расстояние Google. Экспериментальная проверка

выполнена на корпусе юридических текстов на казахском языке, включающем около 10 000 предложений, а также специализированном юридическом словаре, содержащем 8400 терминов и около 10 000 синонимов. Результаты показывают, что предложенный метод обеспечивает высокую точность оценки семантического сходства и может эффективно применяться для обработки юридических текстов на агглютинативных языках. Разработанный подход может быть интегрирован в интеллектуальные юридические информационные системы, включая системы вопросов и ответов и платформы поиска правовой информации. Кроме того, метод обладает потенциалом адаптации к другим тюркским и малоресурсным языкам с аналогичной морфологической структурой.

Ключевые слова: семантическое сходство, синонимы, казахский язык, турецкий язык, юридические термины, агглютинативные языки

Кіріспе. Сөздер немесе сөйлемдер арасындағы семантикалық ұқсастық олардың семантикалық байланыс дәрежесін сипаттайды. Бұл параметр табиғи тілді өңдеу саласында маңызды болып табылады, себебі ол көптеген жасанды интеллект тапсырмаларының тиімділігін анықтайды: мәтіндерді автоматты түрде санаттау және қорытындылау (Bhattacharya et al. 2020; Mandal et al., 2021), машиналық аударма (Bhattacharya et al., 2022), ақпаратты іздеу және деректерді өңдеу (Voronina, 2020), мәтінді талдау, веб-аналитика және басқа да қолданбалар (Sultanova et al., 2019). Сонымен қатар, когнитивті ғылымда семантикалық ұқсастық әртүрлі аналитикалық өлшеулер жүргізу және мағынаның қалыптасуы мен эволюциясы процестерін зерттеу үшін қолданылады.

Қысқа мәтіндердегі семантикалық ұқсастық мәселесі зерттеушілердің ерекше назарын аударды, бұл сандық ортада қысқа мәтін форматтарының кеңінен қолданылуына байланысты: өнім сипаттамалары, сурет тақырыптары мен тегтері, жаңалықтар тақырыптары және т.б. (Ayazbayev et al., 2023). Семантикалық ұқсастық білім беру технологияларында да маңызды рөл атқарады, мысалы, автоматтандырылған тестілеу мен бағалауда. Біздің зерттеуіміз аясында құқықтық мәтіндердің семантикалық ұқсастығы Қазақстан Республикасының заңнамасы үшін интеллектуалды сұрақ-жауап жүйесін әзірлеу үшін қажет.

Сот құжаттарының ұқсастық дәрежесін анықтау күрделі және маңызды міндет болып табылады, ол сот тәжірибесін талдау, ұқсас істерді анықтау және дәйексөзге сілтеме жасау бойынша ұсыныстарды тұжырымдау үшін өте маңызды. Қазіргі уақытта бұл мәселені шешудің екі негізгі тәсілі бар: желілік әдістер және мәтіндік модельдер. Олардың негізгі мақсаты - құқықтық прецеденттер арасындағы ұқсастық дәрежесін анықтау. Статистикалық ұқсастықты бағалаудың классикалық тәсілі семантикалық кеңістік құруға негізделген, мұнда сөздер мәтін корпусындағы таралуын көрсететін векторлармен көрсетіледі. Семантикалық ұқсастық осы векторлардың

салыстырмалы орналасуымен анықталады. Бұл тәсіл мағыналары ұқсас сөздер ұқсас контексттерде кездеседі деген таралу гипотезасына негізделген. Бұл гипотеза сөздер сөйлемде міндетті түрде бір-бірінің жанында емес, керісінше, олар бірдей контексттік элементтер жиынтығымен бірге кездеседі деп болжайды.

Заңды құжаттарды талдаудың заманауи әдістері әдетте мәтіндік мазмұнға немесе сілтеме желісінің құрылымына негізделген. Пахели Бхаттачарья және т.б. (Bhattacharya et al., 2022) жазған «Заңды іс құжаттарының ұқсастығы: Сізге желі де, мәтін де қажет» атты зерттеу екі тәсілдің де шектеулері бар екенін көрсетеді және заңдық ұқсастықты бағалаудың дәлдігін арттыру үшін мәтін мен желілік ақпаратты біріктіруді ұсынады. Көптеген NLP ресурстары жасалған ағылшын тілінен айырмашылығы, көптеген басқа тілдер корпустардың, сөздіктердің және құралдардың жетіспеушілігіне тап болады. Бұл әсіресе күрделі морфологиясы бар ресурстарға тапшы тілдерге қатысты. Сондықтан, көптеген семантикалық модельдер Ағылшын тілі үшін жасалған ұқсастықтар басқа тілдік жағдайларда қолданылмайды.

Біз жоғары морфологиялық өзгергіштікпен және шектеулі тілдік инфрақұрылыммен сипатталатын қазақ тілі үшін әдіс әзірлеу кезінде дәл осы мәселеге тап болдық. Сондықтан біз семантикалық ұқсастықты есептеуге негізделген тәсілді қабылдадық және тек қазақ тіліне ғана емес, сонымен қатар ұқсас тілдік ерекшеліктері бар басқа да ресурстарға тапшы тілдерге қолданылатын әдістемені әзірледік.

Әдеби шолу. Сөйлемдер арасындағы ұқсастықты анықтау тілді өңдеудің көптеген салаларында, соның ішінде мәтінді қайта пайдалануды анықтау, мәтіннің өзектілігі, парафразаны анықтау, ақпаратты алу, ақпаратты іздеу, қысқа жауаптарды бағалау, ұсыныс және бағалау сияқты қолданбалы маңызды міндет болып табылады. Ғылыми еңбектердегі мәтін ұқсастығын бағалаудың бұрынғы көптеген әдістері лексикалық тәсілдерге сүйенді, мұнда ұқсастық сәйкес келетін белгілер санымен немесе белгілер тізбегімен анықталады. Дегенмен, бұл әдістер мәтіндер синонимдік терминдерді қолданғанда немесе мағыналары ұқсас, бірақ бірдей болмаған кезде тиімсіз болады. Мысалы, «Парламент заңды қабылдады» және «Заң шығарушы жиналыс заң жобасын мақұлдады» немесе мәтіндердің мағыналары ұқсас, бірақ бірдей болмаған кезде, мысалы, «Президент заңға қол қойды» және «Мемлекет басшысы заң шығару бастамасын мақұлдады» сияқты.

Б. Бакиевтің «Қазақ тіліндегі мәтіндік құжаттардың синонимдерін ескере отырып ұқсастығын анықтау әдісі: TF-IDF кеңейтімі» (Bakiyev, 2022) еңбегінде синонимдерді ескеруді қамтитын TF-IDF әдісінің кеңейтімі ұсынылған. Ұсынылған әдістің тиімділігі қазақ тіліндегі мәтіндердің ұқсастығын бағалау үшін Cosine, Dice және Jaccard сияқты функцияларды пайдаланатын тәжірибелермен расталады. Зерттеудің мақсаты - екі мәтіндік құжаттың ұқсастығын салыстырмалы түрде талдау. Құжаттар Q құжаты - сұраныс және D құжаты ретінде белгіленеді. Бірінші кезең барлық сөздерді

ұқсастықты бағалауға дайындауды қамтиды, бұл маңызды емес қысқа сөздерді (мысалы, есімдіктер, сандар, таңбалар және т.б.) алып тастауды, содан кейін қалған сөздердің негізгі формаларын алуды қамтиды. Әдетте, бұл үшін Портер немесе Сноубол алгоритмдері қолданылады, бірақ бұл зерттеуде белгілі бір ережелерге негізделген қазақ тілінің стеминг алгоритмі қолданылды (Sultanova et al., 2019). Содан кейін сөздердің бұл негізгі формалары TF-IDF әдісін пайдаланып векторларға түрлендіріледі, бұл сандық кестелерді жасауға әкеледі. Осыдан кейін, құжаттардың ұқсастығын бағалау үшін косинус метрикасы немесе басқа да тиісті әдістер қолданылады. Ұсынылған әдіс мәтінде лексикалық элементтердің жеткілікті саны болған жағдайда екі қазақ тіліндегі құжаттың ұқсастығын бағалайды, бұл салыстыру процесін жеңілдетеді. Дегенмен, қысқа мәтіндерде семантикалық ұқсастық мәселесі күрделене түседі. Салыстыру үшін лексикалық элементтердің саны шектеулі болғандықтан, сәйкес терминдер арасындағы байланысты анықтау қажет.

«Контекст жиынтығын пайдаланып терминдердің семантикалық ұқсастығын анықтау» мақаласында Д.В. Бондарчук (Bondarchuk, 2020) семантикалық жағынан ұқсас терминдер ұқсас контексттерде қолданылады деген болжамға негізделген терминдердің семантикалық ұқсастығын есептеу әдісін ұсынды. Ұсынылған әдістің нәтижелері Жаккард қашықтығымен және 50 адамнан тұратын топ жүргізген бағалаумен салыстырылады. Нәтижелер жоғары корреляция коэффициентін көрсетті, бұл ұсынылған әдістің Жаккард қашықтығын есептеуге негізделген әдіске қарағанда объективтірек екенін көрсетеді (Bondarchuk, 2020). Дегенмен, бұл тәсілді қолдану жаңа мәселені тудырады: термин идентификаторларын таңдаудың қиындығы. Ресурстары аз тілдер үшін айқын анықталған құрылымы бар веб-ресурстарды табу қиын.

«Таратылған Apache Spark жүйесін пайдаланып, қазақ тіліндегі семантикалық жағынан ұқсас сөздерді анықтау» (Ayazbayev et al., 2023) жұмысының аясында жүргізілген қазақстандық ғалымдардың зерттеуінде семантикалық жағынан ұқсас сөздерді анықтау үшін NLPL репозиторийінен алынған векторлар және Polyglot сөз кірістірулері пайдаланылды. Косинус ұқсастығын есептеудің үлкен көлемін өңдеу үшін тапсырма Apache Spark платформасында параллель орындалды. Әрбір мақсатты сөз және сәйкес вектор үшін 176 643 косинус мәні есептелді, содан кейін олар ең қолайлы он кілт сөзді таңдау үшін сұрыпталды. Нәтижелер Polyglot және NLPL модельдері үшін сәйкесінше 88,61% және 92,2% дәлдікті көрсетті. Бұл семантикалық жағынан ұқсас сөздер Уикипедиядағы сәйкес мақалаларды табуға арналған Komekshy іздеу жүйесінде қолданылды. Дегенмен, бұл жұмыста ұқсастық тек сөздердің түбірлерімен анықталғанын, «қолды қылу» сияқты мүмкін фразеологиялық бірліктерді ескермегенін атап өткен жөн, бұл контексте бір нәрсені рұқсатсыз иеленуді білдіреді.

Кесте 1 – Семантикалық жақындық бойынша ғылыми еңбектер

Автор / Жыл	Атауы	Тіл	Әдіс	Нәтиже
Светлана Билощицкая, Арайм Тілеубаева, Александр Кучанский, Салтанат Шарипова – 2025 (Biloshchytska et al., 2025)	Агглютинативті тілдердегі мәтін ұқсастығын анықтау: гибридті N-грамм және семантикалық модельдерді қолдана отырып, қазақ тілінің жағдайын зерттеу	Қазақша	Гибридті N-Gram + TF-IDF + LSH + LSA + LDA + семантикалық модельдер	F1 ұпайы = 0,84, дәлдік = 1,00, еске түсіру = 0,73
Йесси Асри және т.б. басылымы – 2025 (Asri et al., 2025)	Косинус ұқсастығын пайдаланып контекстік ұқсастық үшін сөзді енгізу	ағылшын	Қарама-қарсы оқыту + аспектіге негізделген сөйлемдерді енгізу + косинус ұқсастығы	Ақпаратты іздеу тапсырмалары бойынша ~+3,97%-ға жақсарды
А.В.К.рюкова және т.б. – 2020 (Kryukova et al., 2020)	DKPro құралын пайдаланып мәтіндердің семантикалық ұқсастығын анықтау Ұқсастық	орыс	Семантикалық ұқсастық үшін жолдық метрикалар + ML модельдері	әртүрлі » мәтіндер үшін жіктеуіш негізіндегі тәсілдермен жақсы нәтижелер
З. Шен және З. Сяо – 2024 (Shen and Xiao, 2024)	Сөйлем деңгейіндегі және сөз тіркесі деңгейіндегі семантиканы біріктіретін қытай тіліндегі қысқа мәтіндік ұқсастық әдісі	Қытай	Сөйлем деңгейіндегі + Фразалық деңгейдегі семантика + сөз/сөз тіркестерінің ендірілуі	Қысқа мәтінді сәйкестендіру тапсырмаларының дәлдігі ≈ 90,16%

Кесте 1-де шетел тілдері мен қазақ тілінің семантикалық ұқсастығын талдау бойынша ақпараттық жүйелер саласындағы соңғы зерттеулер мен ғылыми еңбектер келтірілген. Қолданыстағы еңбектерді талдау көрсеткендей, аз ресурсты тілдерге арналған сөйлемдердегі семантикалық ұқсастықтарды анықтау барлық қолжетімді зерттеулерді шолу арқылы әдіснаманы әзірлеуді қажет ететін өзекті мәселе болып табылады. Біздің зерттеуімізде мақсат - қысқа сөйлемдермен байланысты болуы мүмкін заңдық сұраулардағы синонимдерді анықтау. Қысқа мәтіндерде семантикалық ұқсастықты анықтау міндеті салыстыруға арналған сөздердің шектеулі жиынтығымен күрделене түседі, бұл сәйкес терминдер арасындағы байланыстарды табуды маңызды етеді. Қолданыстағы әдістерді салыстырғаннан кейін, келесі тәсіл таңдалды, себебі ол агглютинативті және аз ресурсты тілдерге жарамды.

Қазақ және қырғыз тілдерінің тілдік жақындығы олардың ортақ шығу тегі мен тарихи өзара ықпалдастығымен түсіндіріледі. Екі тіл де түркі тілдер тобына кіреді және лексика, фонетика мен грамматика тұрғысынан ұқсастықтарға ие (Tüymebayev, 2024). Зерттеулер қазақ және қырғыз этностарының

қалыптасуына, сондай-ақ олардың дүниетанымы мен мәдениетінің дамуына бірдей тайпалар мен халықтар қатысқанын көрсетеді, бұл олардың тілдік туыстығын одан әрі дәлелдейді.

Мақалада (Toleush et al., 2021) көне ұйғыр тіліндегі жұп сөздер қазіргі қазақ тіліндегі баламаларымен салыстырылып, осы лексикалық бірліктердің сақталу деңгейі мен өзгерістері анықталған.

Кесте 2-де түркі тілдер тобына кіретін қазақ, қырғыз және түрік тілдері арасындағы лексикалық және фонетикалық ұқсастықтардың салыстырмалы талдауы ұсынылған. Кестеде күнделікті тұрмыста қолданылатын негізгі зат есімдер (су, бас, күн, үй және т.б.) қарастырылып, ортақ түркі түбірлерінің айқын байқалатыны көрсетілген (Bekmanova et al., 2017). Қазақ және қырғыз тілдеріндегі сөздер ресми кирилл әліпбиінде және латынша транскрипцияда берілген.

Көптеген жағдайларда мағынасы мен фонетикалық тұлғасы жағынан дерлік толық сәйкестік байқалады, мысалы: su – suu – su (су), jol – jol – yol (жол). Сондай-ақ түркі тілдеріне тән тұрақты фонетикалық сәйкестіктер көрінеді. Атап айтқанда:

Қазақ және қырғыз тілдеріндегі j дыбысы көбінесе түрік тіліндегі y дыбысына сәйкес келеді (мысалы, jol → yol, jyl → yıl);

kök – gök сияқты жұптарда k және g дауыссыздарының алмасуы сөздің позициясына тәуелді фонологиялық заңдылықты көрсетеді.

Бұл салыстырулар аталған тілдердің терең тілдік туыстығын айқындайды. Алынған нәтижелер тарихи лингвистика, салыстырмалы грамматика салалары үшін маңызды болып табылады, сондай-ақ табиғи тілдерді өңдеу (NLP) саласында, әсіресе көптілді немесе аралық тілдік мәтіндерді өңдеу модельдерін әзірлеуде практикалық қолданысқа ие (Kairat and Burkitbayeva, 2024).

Кесте 2 – Қазақ, түрік және қырғыз тілдеріндегі сөздердің салыстырмасы

Kazakh	**Turkish**	**Kyrgyz**
су (su)	su	суу (suu)
бас (bas)	baş	баш (bash)
күн (kün)	gün	күн (kün)
жыл (jyl)	yıl	жыл (jyl)
түн (tün)	tun	түн (tün)
ата (ata)	ata	ата (ata)
бала (bala)	bala	бала (bala)
жол (jol)	yol	жол (jol)
көк (kök)	gök	көк (kök)

Eşref Adalı зерттеулерінде (Adali, 2024) қазақ және түрік етістіктерінің морфологиялық ережелерін сипаттайтын онтологиялық модельдер құрастырылған. Бұл модельдер олардың морфологиялық ерекшеліктерін NLP жүйелерінде қолдануға мүмкіндік береді.

Қазақ және басқа агглютинативті тілдер арасындағы құрылымдық және

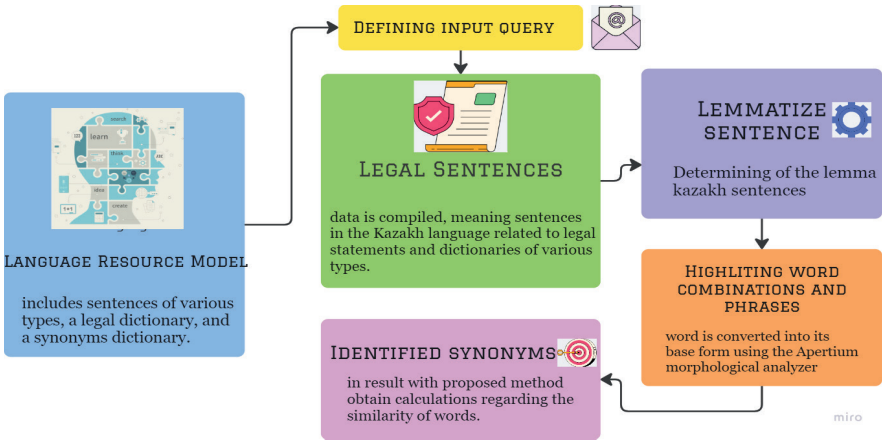
морфологиялық ұқсастықтарды ескере отырып, біз әзірлеген мәтіндерді өңдеу және талдау әдістерінің қырғыз, өзбек немесе түрік тілдеріне бейімделіп, қолданылуы мүмкін деген болжам жасауға болады.

Материалдар мен әдістер. Деректер жиынтығын жасау үшін egov.kz, online.zakon.kz және adilet.zan.kz порталдарын қоса алғанда, құқықтық құжаттарды қамтитын ашық қазақстандық ресурстар пайдаланылды. Бұл платформалар ұқсас құрылымға ие және құқықтық ақпараттың кең ауқымын ұсынады. adilet.zan.kz веб-сайты ең үлкен деректер көлеміне ие, кесте 3-те сайттағы қазақша құжаттар саны көрсетілген, 1947 жылдан 2023 жылға дейінгі құқықтық құжаттарды қамтиды және пайдаланудың қарапайымдылығымен ерекшеленеді. Дегенмен, бұл ресурстар бойынша нақты сұрақтарға жауап іздеу әдетте қолмен орындалады, бұл көп уақытты алады. Нәтижесінде 8,400 заңдық термин мен осы терминдерге арналған 10,000 синонимді қамтитын кешенді қазақ тіліндегі заңдық сөздік жасалды. Сондай-ақ 10,000 заңдық сөйлемнен тұратын корпус дайындалып, ол семантикалық талдау үшін пайдаланылды. Сөздік пен корпус үшін деректер ашық ресурстардан алынды, оның ішінде adilet.kz платформасы да бар, ол Қазақстанның қазіргі заң мәтіндерін ұсынады. Бұл деректер контексттер мен үлгі сөйлемдерді алу үшін негіз болып, заңдық лексика мен контексттердің әртүрлілігін қамтамасыз етіп, қазақ тілінде семантикалық модельдерді әрі қарай талдау және әзірлеу үшін пайдаланылды.

Кесте 3 – Adilet сайтынан алынған мәліметтер

Website	Sentences	Words	Number of documents	Size
https://adilet.zan.kz/kaz/	*506636*	*14656047*	*9575*	*233*

Алдын ала өңдеу бөлімі 1-суретте көрсетілген бірнеше тізбекті кезеңдерді қамтиды:



Сурет 1 – Тіл ресурстары модулі және алдын ала өңдеу құбыры

Тілдік ресурстар модуліне заңдық сөйлемдер, заңдық сөздік және синонимдік сөздік кіреді. 1-суретте көрсетілгендей, енгізу процесі бірнеше кезеңнен тұрады. Алдымен, қазақ тіліндегі заңдық сөйлемдер мен тиісті сөздіктерді қамтитын деректер жиынтығы дайындалды. Деректерді дайындағаннан кейін алынған сөз тіркестері мен сөз тіркестері 1-6 формулаларын пайдаланып есептеулер үшін пайдаланылды.

Нәтижелер. Терминдер мен сөз тіркестерін қамтитын деректерді дұрыс өңдеу үшін лемматизация қажет. Қазақ тілі күрделі морфологияға ие болғандықтан, сөздер мен сөз тіркестері әртүрлі формада болуы мүмкін. Оларды негізгі формаларына түрлендіру үшін Apertium морфологиялық анализаторы қолданылды. Мысалы, «Құқық қорғаушы заңды қорғады» деген сөйлемді Apertium көмегімен талдау келесі лексемаларды анықтайды:

құқық <n> <sg> <nom>
қорғаушы <n> <g> <атауы>
заңды <n> <g> <acc>
қорғадас <v> <tv> <өткен> <sg>

Деректерді дайындағаннан кейін, одан әрі есептеу үшін қазақ сөйлемдері қолданылады.

1. Сөздердің контекстік жиынтығын қалыптастыру.

C1 = {c 11,c 12,...,c1n} және C2 = {c21,c22,...,c2m} сәйкесінше w 1 және w 2 сөздерінің контекстік жиындары болсын. Бұл жиындарда w1 және w2 жиі бір контексте қолданылатын сөздер бар. Содан кейін біз сөздердің ортақ контекстік жиынын құрамыз: $C = C1UC2$

2. w_1 және w_2 сөздерінің әрқайсысы арасындағы нормаланған жақындықтарды есептеу:

$$sim(c_i, w_1) = \frac{f(c_i, w_1)}{\max f\ (w_1)} \qquad (1)$$

$$\text{sim}(c_i, w_2) = \frac{f\ (c_i, w_2)}{\max f(w_2)} \qquad (2)$$

мұндағы жақындық (c_i, w_1) - c_i және w_1 бірге кездесетін құжаттар саны , ал макс.жиілік (w_j) C алынған барлық сөздер үшін максималды жиілік ретінде есептеледі:

$$\max f\ (wj\ ) = max(f(\text{ci}\ , wj\ )), ci \in C \qquad (3)$$

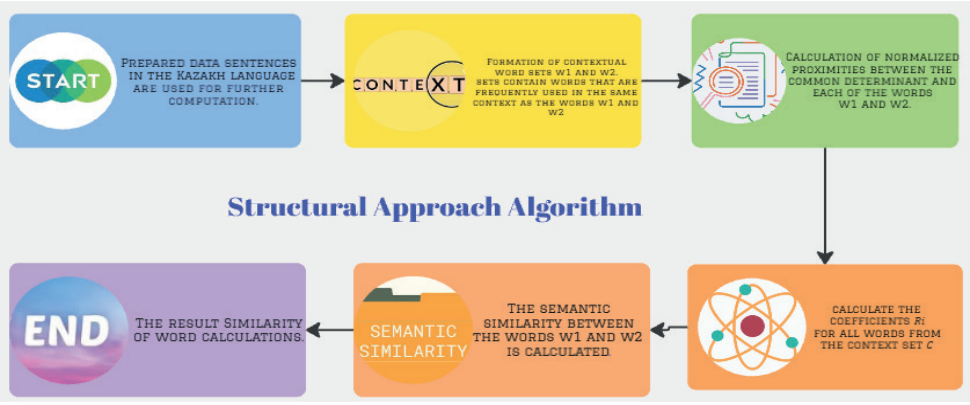
3. C контекст жиынындағы барлық сөздер үшін R_i коэффициенттерін төмендегі формуланы пайдаланып есептейміз:

$$R_i = \frac{min\{sim(c_1, w_1), sim(c_1, w_2)\}}{max\{sim(c_1, w_1), sim(c_1, w_2)\}} \qquad (4)$$

$$\text{SemSim}(w_1, w_2) = \frac{\sum_1^k \left(\frac{p_i R_i}{1+R_i} + s\right)}{1+s} \quad (5)$$

Ri – бүкіл үлгідегі w 1 және w 2 бірлесіп пайда болу коэффициенті болсын , екі сөз де бір құжатта кездескенде 2-ге тең, ал басқа жағдайда 1 болса, s – синонимдік коэффициент, егер w 1 және w 2 сөздері синонимдер болса 1-ге тең, ал басқа жағдайда 0-ге тең.

Анықтық үшін сөйлемнің семантикалық ұқсастығын есептеу алгоритмі ұсынылады. Ұсынылған семантикалық тәсілдің жалпы диаграммасы мен негізгі кезеңдері 2-суретте көрсетілген.



Сурет 2 - Ұсынылған семантикалық тәсілдің алгоритмі

4-кестеде семантикалық ұқсастықты есептеу үшін лексикалық ресурсты қалыптастыруда қолданылатын қазақ тіліндегі семантикалық жағынан жақын заң терминдері мен сөз тіркестерінің мысалдары келтірілген.

Кесте 4 – Қазақ тіліндегі мағыналық жағынан жақын заң терминдерінің мысалдары

Құқықтық ұғым санаты	**Семантикалық жағынан ұқсас терминдер**
Нормативтік құқықтық актілер	заң , қаулы , акт, құжат
Заңды әрекеттер	бас тарту , тоқтату , қысқарту
Қылмысқа қатысу	қылмас жасауға айдап салу, қылмыс жасауға азғыру
Сынақ	сот ісі , сот процесі , соттық қудалау
Заңды рөлдер	адвокат, қорғаушы
Жауапкершілікті тоқтату	ақтау , кешу , жазадан босату

Әдістің тиімділігін «заң» және «қаулы » сөздерінде тексеріп көрейік, оларды қолданыстағы қазақ корпустары мен сөздіктерін пайдаланып оқытайық. Бұл жағдайда w1 «заң», ал w2 «қаулы». Сондай-ақ, осы сөздер үшін контекст жиынтықтарын жасаймыз:

С 1 ={ зан , қабылда, шеш, бекіт, отыр, ен , өзгерт }

С 2 ={ қаулы, жанарт, бекіт, өзгерт, ауыс }

«Заң» және «қаулы» сөздері мағынасы жағынан ұқсас (синонимдер) болғандықтан, егер құжатта «заң» сөзі болса, онда «қаулы» сөзі де бар деп есептейміз. 5-кестеде 1-формуланы пайдаланып есептелген ортақ контекст жиынтығы мен сөздер арасындағы қалыпқа келтірілген жақындықтар көрсетілген.

Кесте 5 – Қалыптастырылған жақындық арасында ортақ контекстік жиынтық және заң және қаулы сөздері

	заң	*қаулы*
қабылда	0,653	0,609
шеш	0,555	0,476
бекіт	0,667	0.5001
ен	0,143	0,0224
өзгер	0,833	0.1875

Мысалы , 5-кестедегі бірінші жол үшін есептеулер келесідей орындалады:

$$sim(c_i, w_1) = \frac{f\,(c_i, w_1)}{\max f\,(w_1)} = 980/1500=0{,}653$$

$$sim(c_i, w_2) = \frac{f\,(c_i, w_2)}{\max f(w_2)} = 672/1102=0.609$$

Ri коэффициенттерін есептейміз; нәтижелері 6-кестеде көрсетілген. Есептеу жалпы контекст жиынынан алынған әрбір сөз үшін жүргізілді. Мысалы, «заң» және «қаулы» сөздері үшін *Ri* коэффициенттері келесі мәндерге ие:

$$R_i = \frac{min\{sim(c_1, w_1), sim(c_1, w_2)\}}{max\{sim(c_1, w_1), sim(c_1, w_2)\}} = 0.609/0.653=0.93$$

Кесте 6 – R_i коэффициенті (*қаулы, заң*)

заң	0,93
қабылда	0,87
шеш	0,63
бекіт	0,74
ен	0,58
өзгерт	0,61
ауыс	0,47
жаңарт	0,78
қаулы	0,93

Осы коэффициенттерді есептегеннен кейін, семантикалық ұқсастықты нақты есептеуге көшуге болады. (1) формуланы пайдаланып, « қаулы » және

«заң» сөздерінің ұқсастығын есептеп, 0,93 аламыз. Сондай-ақ, семантикалық ұқсастықты Жаккард қашықтығын (Cahyani et al., 2025) пайдаланып есептеп, 0,38 аламыз. Қосымша есептеулерсіз нәтижелердің бір-бірінен айтарлықтай ерекшеленетіні анық.

Семантикалық ұқсастықты есептеудің ұсынылған әдісінің тиімділігін бағалау үшін кеңінен қолданылатын семантикалық ұқсастық көрсеткіштерімен, атап айтқанда косинуспен салыстырмалы талдау жүргізілді. Ұқсастық, Жаккард Ұқсастық және қалыпқа келтірілген Google қашықтығы (NGD).

Косинус ұқсастығы — салыстырылатын мәтіндердің векторлық көріністері жасалады және осы векторлар арасындағы бұрыштың косинусы есептеледі (Kanishkaa and Santhi, 2025).

$$cos\theta = C(\vec{X},\vec{Y}) = C(\overrightarrow{Y},\vec{X}) = \frac{\vec{X}\times\vec{Y}}{\overrightarrow{X\vee\times\vee\vec{Y}\vee}=\frac{\Sigma_{i=1}^{n} x_i y_i}{\sqrt{\Sigma_{i=1}^{n} x_i^2}\sqrt{\Sigma_{i=1}^{n} y_i^2}}} \qquad (6)$$

Жаккард өлшемі — құжаттар n-граммға бөлінеді, ал Жаккард коэффициенті жалпы n-грамм санының n-граммның жалпы санына қатынасы ретінде есептеледі.

$$J(\vec{X},\vec{Y}) = \frac{\vec{X}\times\vec{Y}}{|\vec{X}|^2\times|\vec{Y}|^2-\vec{X}\times\vec{Y}} \qquad (7)$$

мұндағы x_i – X y_i векторының компоненті, - Y векторының компоненті, n - ұзындық X және Y векторлары.

NGD (Қалыптанған Google қашықтығын) келесі формуланы пайдаланып есептеуге болады (Ihnaini, 2024).:

$$NGD(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logN - min\{logf(x), logf(y)\}} \qquad (8)$$

мұндағы N - веб-беттердің жалпы саны, Google іздеу жүйесімен өңделген,

f(t1) және f(t2) - t1 және t2 терминдері бөлек кездесетін беттер саны,

f (t1, t2)- t1 және t2 терминдері бірге орналасқан беттер саны.

Бұл көрсеткіштер бастапқы салыстыру әдістері ретінде пайдаланылды, бұл бізге ұсынылған тәсілдің қазақ тіліндегі заң терминдері мен сөз тіркестерін өңдеу кезіндегі артықшылықтарын сандық бағалауға мүмкіндік берді. Салыстырмалы эксперимент нәтижелері 7-кестеде келтірілген.

Кесте 7 – Сөз жұптары үшін метрикалар мен ұсынылған әдіс арасындағы салыстырмалы талдау

Бірнеше сөз	Jaccard	Cossine	NGD	Proposal method
Заң қабылда	0,282	0,44	0,83	0,96

Қаулы енгіз	0,311	0,561	0,783	0,89
Заң ен	0,572	0,697	0,802	0,901

8-кестеде ұсынылған әдістің тиімділігінің салыстырмалы нәтижесі сөз тіркестері үшін Jaccard қашықтығымен, NGD және Cossine мәндерімен салыстырғанда көрсетілген.

Кесте 8 – Салыстырмалы талдау метрика және сөз тіркестері үшін ұсынылған әдіс

Сөз тіркестері	Jaccard	Cossine	NGD	Proposal method
қамаудағы жазбаша түсінік беру	0,0482	0,066	0,545	0,632
кісі өлтіруде айыпталған тұлға	0,058	0,068	0,601	0,728
кісі өлтіруде сезікті адам	0,012	0,034	0,045	0,0521

Жүргізілген семантикалық ұқсастықты есептеу нәтижелеріне сүйене отырып, бірнеше негізгі қорытынды жасауға болады. Біріншіден, ұсынылған әдіс, контекстік талдау мен Ri коэффициентін пайдалану арқылы, Жаккард, Косинус және NGD сияқты классикалық метрикалардан жоғары дәлдік көрсетті. Мысалы, «Заң қабылда» сөз тіркесі үшін ұсынылған әдіс коэффициенті 0,96 болды, ал Жаккард пен Косинус метрикалары сәйкесінше 0,282 және 0,44 көрсетті, бұл ұсынылған әдістің елеулі артықшылығын дәлелдейді.

Екіншіден, күрделі сөз тіркестері үшін, мысалы, «қамаудағы жазбаша түсінік беру», әдіс те жақсы нәтижелер көрсетті, бірақ бұл фразалар үшін көрсеткіштер төмен болды. Бұл сөз тіркесі үшін ұсынылған әдіс 0,632 нәтижесін көрсетті, ал Жаккард пен Косинус метрикалары 0,0482 және 0,066 көрсетті. Бұл күрделі сөз тіркестері үшін контекстік деректердің аздығына байланысты семантикалық ұқсастықты есептеу қиынырақ екенін көрсетеді.

Кесте 9. Түрік және қазақ тілдеріндегі семантикалық ұқсастықтың әртүрлі әдістері бойынша F1-ұпайлары

Әдіс	F1-score (қазақ тілі)	F1-score (түрік тілі)
NGD	0,28	0,27
Jaccard	0,26	0,25
Cossine	0,24	0,23
Proposal method	0,69	0,48

Кесте 9 нәтижелер бойынша, NGD әдісінің F1-ұпайы үшін түрік тілінде 0,2764, ал қазақ тілінде 0,2824 болды. Бұл көрсеткіштердің жақын болуы, екі тілдің де агглютинативті құрылымға ие болуымен және ұқсас интернет-ресурстардан алынған деректермен қамтамасыз етілуімен түсіндіріледі. Алайда, қазақ тілінің көрсеткіші сәл жоғары, бұл қазақ тіліндегі заңды терминдердің кеңірек корпусы мен нақты терминдер арасындағы байланыстардың күшті екенін көрсетеді.

Jaccard әдісі бойынша, түрік тілінде 0,2587, ал қазақ тілінде 0,2653 F1-ұпайы есептелді. Бұл айырмашылық, қазақ тіліндегі терминдер мен фразалардың көбірек дәл сәйкес келуімен байланысты болуы мүмкін, әсіресе заң мәтіндерінде. Осылайша, қазақ тілінің терминологиялық базасы жоғары сапалы әрі көпшілікке ашық болғандықтан, Jaccard метрикасы сәл жоғары нәтиже көрсетті.

Cosine әдісінде түрік тілінде 0,2419, қазақ тілінде 0,2421 F1-ұпайы болды. Бұл әдіс бойынша нәтижелер бір-біріне өте жақын, бұл көрсеткіштердің контекстуалды ұқсастықтарды бағалаудағы тиімділігін білдіреді, бірақ толық емес контекстер мен морфологиялық ерекшеліктерді ескермеуі мүмкін.

Ұсынылған әдіс екі тіл үшін де едәуір жоғары нәтижелер көрсетті: қазақ тілінде F1-ұпайы 0,69, ал түрік тілінде 0,48 болды. Мұндай айырмашылық ұсынылған тәсілдің контекстік жиындарды және Ri коэффициентін пайдалану арқылы семантикалық байланыстарды тереңірек модельдейтінімен түсіндіріледі. Қазақ тіліндегі жоғары көрсеткіш корпус деректерінің құрылымдық бірізділігі мен контексттердің тығыздығымен байланысты болса, түрік тіліндегі салыстырмалы түрде төмен нәтиже корпус көлемі мен лексикалық таралу ерекшеліктерінің әсерінен болуы мүмкін. Жалпы алғанда, алынған нәтижелер ұсынылған әдістің агглютинативті тілдердегі семантикалық ұқсастықты бағалауда тиімді тәсіл екенін дәлелдейді.

Эксперименттік есептеулер үшін 10000 заңдық сөйлемнен тұратын монолингвальды корпус қолданылды, сондай-ақ деректер Adilet.kz платформасынан алынған. Бұл платформа Қазақстанның заң мәтіндерінің соңғы деректерін ұсынады. Осы платформадағы корпус 506636 сөйлемнен тұрады, ал жалпы көлемі 14656047 сөз болды, бұл контексттер мен сөз тіркестерін талдау кезінде жоғары дәлдікті қамтамасыз етеді. Зерттеу жұмыстары түрік тіліндегі семантикалық ұқсастықты анықтау бойынша жүргізілді, автор ұсынған әдіс жиілік талдауы мен Ri коэффициентіне негізделген. Түрік тілі үшін жиілік деректері Түрік Ұлттық Корпусының (TNC) мәтіндерінен алынды. Корпустың жалпы көлемі шамамен 10 миллион сөз тіркесінен тұрады.

Талдаудың дәлдігін қамтамасыз ету үшін барлық сөздер Zemberek ашық бастапқы морфологиялық анализаторы арқылы лемматизацияланды. Контекстуалды етістіктер (мысалы, kabul et – қабылдау, onayla – мақұлдау, değiştir – өзгерту) қолмен таңдалды, олар мақсатты есімдермен (мысалы, iş – жұмыс, görev – тапсырма) жиіліктік қатынастарына негізделіп алынды. Осы жиіліктер кейін семантикалық ұқсастықты есептеу үшін ұсынылған Ri коэффициентімен қолданылды.

Алынған деректер көрсеткендей, ұсынылған әдіс универсалдылыққа ие және қазақ тілі мен түрік тілі сияқты басқа агглютинативті тілдерге де бейімделуі мүмкін.

Осылайша, эксперименттік нәтижелер ұсынылған әдістің жоғары тиімділігін және оның қазақ тіліндегі заң мәтіндерін өңдеуде қолдануға

жарамдылығын растайды. Әдіс семантикалық ұқсастықты дәл және сенімді бағалауды қамтамасыз етеді, агглютинативті тілдер үшін бұл тәсілдің болашағы зор екені көрсетілді.

Талқылау. Эксперименттік зерттеу семантикалық ұқсастықты есептеудің ұсынылған әдісін Косинус, Жаккард және NGD сияқты классикалық семантикалық ұқсастық көрсеткіштерімен салыстырды. Эксперименттер қазақ тіліндегі заң мәтіндерінің корпусында жүргізілді және жеке терминдерді де, орфографиялық сөз тіркестерін де, қысқа сөз тіркестерін де қамтыды.

Сөз жұптарының салыстырмалы талдау нәтижелері 5-кестеде келтірілген. Алынған мәндер ұсынылған әдістің бастапқы көрсеткіштермен салыстырғанда семантикалық ұқсастықтың жоғары дәрежесін көрсететінін көрсетеді. Бұл контекст жиынтықтарын пайдалану және терминдер арасындағы байланыстарды симметриялы бағалаумен байланысты, бұл құқықтық саладағы семантикалық байланыстарды дәлірек есепке алуға мүмкіндік береді.

6-кестеде көп сөзден тұратын сөз тіркестерінің эксперименттік нәтижелері көрсетілген. Ұсынылған деректерден көрініп тұрғандай, тұтастай алғанда сөз тіркестерінің семантикалық ұқсастық мәндері жеке сөздерге қарағанда төмен. Бұл көп сөзден тұратын құрылымдарды өңдеудің күрделілігінің жоғарылығымен, сондай-ақ қазақ тілінің морфологиялық ерекшеліктерінің және үш немесе одан да көп сөздердің бірге қолданылу жиілігінің таралуының әсерімен түсіндіріледі.

Әртүрлі метрикаларды пайдаланып алынған нәтижелерді салыстыру косинус ұқсастығы және Жаккард ұқсастығы сияқты классикалық тәсілдердің, әсіресе шектеулі деректермен, заң терминдері арасындағы семантикалық қатынастарды анықтауға онша қабілетті емес екенін көрсетеді. NGD метрикасы тұрақты мәндерді көрсетеді, бірақ жиілік таралуына және деректер көздеріне сезімтал болып қалады. Дегенмен, ұсынылған әдіс контекст пен синонимдік қатынастарды ескере отырып, семантикалық ұқсастықты сенімдірек бағалауды қамтамасыз етеді.

Жалпы алғанда, эксперименттік нәтижелер қазақ тіліндегі заң терминдері мен сөз тіркестерінің семантикалық ұқсастығын талдау үшін ұсынылған тәсілдің тиімділігін растайды. Әдіс әртүрлі мәтін түрлерінде тұрақты жұмыс істеуді көрсетті және заң құжаттарын, әсіресе ресурстары аз тілдерде, ақылды іздеу және талдау үшін пайдаланылуы мүмкін.

Қорытынды. Бұл мақалада ресурстары аз тілдің нақты терминологиясы мен сипаттамаларын ескере отырып, қазақ тіліндегі заң мәтіндерінің семантикалық ұқсастығын анықтауға контекстке негізделген тәсіл ұсынылады. Жалпы мақсаттағы статистикалық өлшемдерден айырмашылығы, әзірленген әдіс мамандандырылған заң саласындағы семантикалық қатынастарды талдауға бағытталған және үлкен, белгіленген деректер корпустарын пайдалануды талап етпейді.

Жүргізілген семантикалық ұқсастықты есептеу нәтижелері бойынша ұсынылған әдіс классикалық метрикалар, атап айтқанда Жаккард және NGD

көрсеткіштерінен айтарлықтай жоғары нәтижелер көрсетті. Контекстік талдау мен Ri коэффициентін қолданатын ұсынылған әдіс, әсіресе заң мәтіндерінде, семантикалық ұқсастықты анықтауда жоғары дәлдік көрсетті. Мысалы, «Заң қабылда» сөз тіркесі үшін Ri коэффициенті 0,96 болды, бұл Жаккард пен Косинус нәтижелерімен салыстырғанда 0,282 және 0,44 айтарлықтай жоғары. Бұл ұсынылған әдістің агглютинативті тілдер, мысалы, қазақ тілінде терминдердің семантикалық ұқсастығын дәл бағалайтынын көрсетеді.

Түрік тілінде де ұсынылған әдіс жоғары дәлдікті көрсетті, бірақ жалпы F1-ұпайлары төмен болды (0,546), бұл екі тілдің агглютинативті құрылымдарының ұқсастығы мен заңдық лексикада жалпы семантикалық байланыстардың бар екенін дәлелдейді. Алайда, күрделі сөз тіркестері, мысалы, «қамаудағы жазбаша түсінік беру», нәтижелерінің дәлдігі төмендеді, өйткені бұл тіркестер корпуста жиі кездеспейді. Алайда, ұсынылған әдіс әлі де жоғары дәлдікті көрсетті, бұл Жаккард пен Косинус метрикаларына қарағанда контекст пен синонимдер арасындағы байланыстарды дәл есептеуге мүмкіндік береді. Бұл зерттеу қазақ және түрік тілдеріндегі заң мәтіндерін өңдеуге арналған ұсынылған әдістің тиімділігін дәлелдейді және оны басқа агглютинативті тілдерге бейімдеуге мүмкіндік береді.

Алайда, күрделі сөз тіркестері, мысалы, «қамаудағы жазбаша түсінік беру», нәтижелерінің дәлдігі төмендеді, өйткені бұл тіркестер корпуста жиі кездеспейді. Алайда, ұсынылған әдіс әлі де жоғары дәлдікті көрсетті, бұл Жаккард пен Косинус метрикаларына қарағанда контекст пен синонимдер арасындағы байланыстарды дәл есептеуге мүмкіндік береді. Бұл зерттеу қазақ тіліндегі заң мәтіндерін өңдеуге арналған ұсынылған әдістің тиімділігін дәлелдейді және оны басқа агглютинативті тілдерге бейімдеуге мүмкіндік береді. Ұсынылған тәсілді құқықтық құжаттарды іздеу, салыстыру және түсіндіру жүйелерін қоса алғанда, интеллектуалды құқықтық ақпаратты өңдеу жүйелерінде, сондай-ақ сұрақ-жауап жүйелерінде қолдануға болады. Оны қолдану ұқсас терминдер мен сөз тіркестерін түсіндірудің дәлдігін жақсартады, бұл құқықтық мәтіндер үшін өте маңызды. Әрі қарайғы зерттеу бағыттарына құқықтық деректер корпусын кеңейту, лексикалық ресурстарды толықтыру және әдісті ұзын мәтін фрагменттерімен және басқа да аз ресурстарды қажет ететін тілдермен жұмыс істеуге бейімдеу кіреді.

References

Adali E. (2024) The logic of Turkish language for NLP. Journal of Problems in Computer Science and Information Technologies, 2:34–46 (in Eng.).

Asri Y., Kuswardani D., Sari A.A., Ansyari A.R. (2025) Word embedding for contextual similarity using cosine similarity. Indonesian Journal of Electrical Engineering and Computer Science, 38:1170–1180. DOI: 10.11591/ijeecs.v38.i2.pp1170-1180 (in Eng.).

Ayazbayev D., Bogdanchikov A., Orynbekova K., Varlamis I. (2023) Defining semantically close words of Kazakh language with distributed system Apache Spark. Informatics, 7(4):160. DOI: 10.3390/informatics7040160 (in Eng.).

Bakiyev B. (2022) Method for determining the similarity of text documents for the Kazakh language taking into account synonyms: Extension to TF-IDF (in Eng.).

Bekmanova G., Sharipbay A., Altenbek G., Adali E., Zhetkenbay L., Kamanur U., Zulkhazhav A. (2017) A uniform morphological analyzer for the Kazakh and Turkish languages. Proceedings of AIST (Supplement):20–30 (in Eng.).

Bhattacharya P., Ghosh K., Pal A., Ghosh S. (2020) Hier-SPCNet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval:1657–1660. DOI: 10.1145/3397271.3401191 (in Eng.).

Bhattacharya P., Ghosh K., Pal A., Ghosh S. (2022) Legal case document similarity: You need both network and text. Information Processing & Management, 59(6):103069. DOI: 10.1016/j.ipm.2022.103069 (in Eng.).

Biloshchytska S., Tleubayeva A., Kuchanskyi O., Sharipova S. (2025) Text similarity detection in agglutinative languages: A case study of Kazakh using hybrid n-gram and semantic models. Applied Sciences, 15:6707. DOI: 10.3390/app15126707 (in Eng.).

Bondarchuk D.V. (2020) Opredelenie semanticheskoi blizosti termov s pomoshchyu kontekstnogo mnozhestva [Determination of semantic similarity of terms using a contextual set] (in Russ.).

Cahyani A.D., Fathoni M.W., Rachman F.H., et al. (2025) Automatic essay scoring: Leveraging Jaccard coefficient and cosine similarity with n-gram variation. arXiv preprint. DOI: 10.48550/arXiv.2501.01234 (in Eng.).

Ihnaini B. (2024) Semantic similarity on multimodal data: A comprehensive survey. Journal of Information and Data Management, 15(2):1–20. DOI: 10.5753/jidm.2024.3210 (in Eng.).

Kairat A.A., Burkitbayeva Sh.D. (2024) Reduplicated words in Old Uyghur and modern Kazakh languages. Tiltanym, 2:133–139. DOI: 10.55491/2411-6076-2024-2-133-139 (in Eng.).

Kanishkaa S., Santhi K. (2025) A comparative analysis of Jaccard and cosine similarity for plagiarism detection. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 13(3). DOI: 10.17148/IJIREEICE.2025.13305 (in Eng.).

Kryukova A.V. et al. (2020) Opredelenie semanticheskih svyazei s pomoshchyu DKPro Similarity [Determination of semantic relations using DKPro Similarity] (in Russ.).

Mandal A., Ghosh K., Ghosh S., Mandal S. (2021) Unsupervised approaches for measuring textual similarity between legal court case reports. Artificial Intelligence and Law, 29:1–25. DOI: 10.1007/s10506-020-09260-2 (in Eng.).

Shen Z., Xiao Z. (2024) A Chinese short text similarity method integrating sentence-level and phrase-level semantics. Electronics, 13:4868. DOI: 10.3390/electronics13244868 (in Eng.).

Sultanova N., Kozhakhmet K., Jantayev R., Botbayeva A. (2019) Stemming algorithm for Kazakh language using rule-based approach. Proceedings of the 15th International Conference on Electronics, Computer and Computation (ICECCO):1–4. DOI: 10.1109/ICECCO48375.2019.9043249 (in Eng.).

Toleush A., Israilova N., Tukeyev U. (2021) Development of morphological segmentation for the Kyrgyz language on complete set of endings. Proceedings of the 13th Asian Conference on Intelligent Information and Database Systems (ACIIDS). DOI: 10.1007/978-3-030-73100-7_46 (in Eng.).

Tuymebayev J. (2024) Comparative grammar of the Kazakh and Kyrgyz languages (in Eng.).

Voronina L.V. (2020) Voprosy relevantnosti referenta i referentnogo vyrazheniya antecedentno-anaforicheskogo kompleksa, realizuyushchego semantiku tseli v politicheskom diskurse [Issues of relevance of the referent and referential expression of the antecedent-anaphoric complex implementing the semantics of purpose in political discourse]. Otechestvennaya filologiya, 5:16–25. DOI: https://doi.org/10.18384/2310-7278-2020-5-16-25 (in Russ.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 373–392

https://doi.org/10.32014/2026.2518-1726.418

IRSTI 28.23.25
UDC 004.855.5

© Serek A.[1], Shoiynbek A.[2]*, Sharipov K.[2], Kuanyshbay D.[2], Mukhametzhanov A.[3], 2026.
[1]Astana IT University, Astana, Kazakhstan;
[2]Narxoz University, Almaty, Kazakhstan;
[3]SDU University, Kaskelen, Kazakhstan.
E-mail: aisultan.shoiynbek@gmail.com

ANALYSIS AND CLASSIFICATION OF TELEPHONE FRAUD BASED ON LEXICAL FEATURES OF SPEECH TRANSCRIPTIONS

Serek Azamat — PhD, Associate Professor, Astana IT University, Astana, Kazakhstan,
E-mail: azamat.serek@astanait.edu.kz, https://orcid.org/0000-0001-7096-6765;
Shoiynbek Aisultan — PhD, Professor, Narxoz University, Almaty, Kazakhstan,
E-mail: aisultan.shoiynbek@gmail.com, https://orcid.org/0000-0002-9328-8300;
Sharipov Karim — Master's student, Narxoz University, Almaty, Kazakhstan,
E-mail: karim.sharipov@narxoz.kz, https://orcid.org/0009-0003-2452-8803;
Kuanyshbay Darkhan — PhD, Assistant-Professor, Narxoz University, Almaty, Kazakhstan,
E-mail: darkhan.kuanyshbay @narxoz.kz, https://orcid.org/0000-0001-5952-8609;
Mukhametzhanov Assylbek — Master's student, SDU University, Kaskelen, Kazakhstan,
E-mail: 221107046@stu.sdu.edu.kz, https://orcid.org/0009-0009-8528-9985.

Abstract. Telephone fraud (vishing) is one of the most common forms of social engineering, causing significant financial and psychological damage. In conditions of constant number changes and fraud scenarios, traditional protection methods are not effective enough, which makes it necessary to automatically analyze the contents of telephone conversations. This paper presents an experimental study of the use of machine learning methods for the automatic detection of fraudulent phone calls based on text transcriptions of speech. To conduct the experiment, a balanced Russian-language corpus of 1,400 telephone conversations was formed, including fraudulent and legitimate calls received from open sources. The audio recordings were automatically transcribed using the Whisper neural network speech recognition model, after which the texts were normalized and lemmatized. TF–IDF unigrams and bigrams were used as the feature representation. Based on the data obtained, several classical machine learning models were trained and compared, including Logistic Regression, Linear SVM, Multinomial Naive Bayes, Random Forest, and XGBoost. Experimental results showed that all the considered

models achieve high classification accuracy, while the best performance was demonstrated by linear models and the Multinomial Naive Bayes classifier with minimal smoothing (accuracy up to 94%, ROC–AUC up to 0.99). The analysis of lexical features made it possible to identify stable verbal markers of fraudulent speech, characteristic of typical scenarios of social engineering. The stability and generalizing ability of the models were confirmed using k-fold cross-validation and ROC–AUC analysis. The results obtained indicate the practical applicability of the proposed approach for the automatic detection of telephone fraud in real conditions.

Keywords: telephone fraud; vishing; machine learning; natural language processing; speech recognition; TF–IDF; text classification

Financing. *This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510301 "Development of technology for recognizing fraudulent actions during a telephone conversation and/or text message exchange in messengers based on artificial intelligence algorithms").*

For citations: Serek A., Shoiynbek A., Sharipov K., Kuanyshbay D., Mukhametzhanov A. Analysis and classification of telephone fraud based on lexical features of speech transcriptions. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 373–392. DOI: https://doi.org/10.32014/2026.2518-1726.418

© Серек А.[1], Шойынбек А.[2]*, Шарипов К.[2], Қуанышбай Д.[2], Мухаметжанов А.[3], 2026.

[1]Астана IT Университеті, Астана, Қазақстан;
[2] Нархоз университеті, Алматы, Қазақстан;
[3] SDU Университеті, Қаскелең, Қазақстан.
E-mail: aisultan.shoiynbek@gmail.com

СӨЙЛЕУ ТРАНСКРИПЦИЯЛАРЫНЫҢ ЛЕКСИКАЛЫҚ БЕЛГІЛЕРІНЕ НЕГІЗДЕЛГЕН ТЕЛЕФОН АЛАЯҚТЫҚТАРЫН ТАЛДАУ ЖӘНЕ ЖІКТЕУ

Серек Азамат — PhD, Қауымдастырылған профессор, Астана IT Университеті, Астана, Қазақстан,
E-mail: azamat.serek@astanait.edu.kz, https://orcid.org/0000-0001-7096-6765;
Шойынбек Айсұлтан — PhD, профессор, Нархоз университеті, Алматы, Қазақстан,
E-mail: aisultan.shoiynbek@gmail.com, https://orcid.org/0000-0002-9328-8300;
Шарипов Карим — магистрант, Нархоз университеті, Алматы, Қазақстан,
E-mail: karim.sharipov@narxoz.kz, https://orcid.org/0009-0003-2452-8803;
Қуанышбай Дархан — PhD, ассистент-профессор, Нархоз университеті, Алматы, Қазақстан,
E-mail: darkhan.kuanyshbay@sdu.edu.kz, https://orcid.org/0000-0001-5952-8609;
Мухаметжанов Асылбек — магистрант, SDU Университеті, Қаскелең, Қазақстан,
E-mail: 221107046@stu.sdu.edu.kz, https://orcid.org/0009-0009-8528-9985.

Аннотация. Телефондық алаяқтық (вишинг) - бұл қаржылық және психологиялық тұрғыдан айтарлықтай зиян келтіретін әлеуметтік инженерияның кең таралған түрлерінің бірі. Алаяқтардың нөмірлері мен сценарийлерінің үнемі өзгеруі жағдайында дәстүрлі қорғаныс әдістері тиімсіз болып шығады, бұл телефон қоңырауларының мазмұнын автоматты түрде талдауды қажет етеді. Бұл жұмыста мәтіндік сөйлеу транскрипцияларына негізделген алаяқтық телефон қоңырауларын автоматты түрде анықтау үшін машиналық оқыту әдістерін қолдану бойынша эксперименттік зерттеу ұсынылған. Эксперимент жүргізу үшін ашық көздерден алынған алаяқтық және заңды қоңырауларды қамтитын 1400 телефон қоңырауларынан тұратын теңдестірілген орыс тілді корпус құрылды. Дыбыстық жазбалар Whisper-дің сөйлеуді танудың нейрондық моделі арқылы автоматты түрде транскрипцияланды, содан кейін мәтіндер қалыпқа келтіріліп, лемматизацияланды. TF–IDF униграммалары мен биграммалары қолтаңба көрінісі ретінде пайдаланылды. Алынған мәліметтер Logistic Regression, Linear SVM, Multinomial Naive Bayes, Random Forest және XGBoost сияқты бірнеше классикалық Машиналық оқыту модельдерін оқыды және салыстырды. Эксперименттік нәтижелер барлық қарастырылған модельдер жоғары жіктеу дәлдігіне қол жеткізетінін көрсетті, ең жақсы көрсеткіштер сызықтық модельдер мен Multinomial Naive Bayes классификаторын минималды Тегістеу арқылы көрсетті (accuracy 94% дейін, ROC–AUC 0.99 дейін). Лексикалық белгілерді талдау әлеуметтік инженерияның типтік сценарийлеріне тән алаяқтық сөйлеудің тұрақты ауызша белгілерін анықтауға мүмкіндік берді. Модельдердің тұрақтылығы мен жалпылау қабілеті ROC-AUC кросс-валидациясы мен талдауының k–еселігі арқылы расталды. Нәтижелер нақты жағдайларда телефон алаяқтықтарын автоматты түрде анықтау үшін ұсынылған тәсілдің практикалық қолданылуын көрсетеді.

Түйін сөздер: телефон алаяқтық; вишинг; Машиналық оқыту; табиғи тілді өңдеу; сөйлеуді тану; TF-IDF; мәтіндерді жіктеу

© Серек А.[1], Шойынбек А.[2]*, Шарипов К.[2], Куанышбай Д.[2], Мухаметжанов А.[3], 2026.

[1] Астана IT Университет, Астана, Казахстан;

[2] Университет Нархоз, Алматы, Казахстан;

[3] SDU Университет, Каскелен, Казахстан.

E-mail: aisultan.shoiynbek@gmail.com

АНАЛИЗ И КЛАССИФИКАЦИЯ ТЕЛЕФОННОГО МОШЕННИЧЕСТВА НА ОСНОВЕ ЛЕКСИЧЕСКИХ ПРИЗНАКОВ РЕЧЕВЫХ ТРАНСКРИПЦИЙ

Серек Азамат — PhD, ассоциированный профессор, Астана IT Университет, Астана, Казахстан,

E-mail: azamat.serek@astanait.edu.kz, https://orcid.org/0000-0001-7096-6765;

Шойынбек Айсултан — PhD, профессор, Университет Нархоз, Алматы, Казахстан,
E-mail: aisultan.shoiynbek@gmail.com, https://orcid.org/0000-0002-9328-8300;
Шарипов Карим — магистрант, Университет Нархоз, Алматы, Казахстан,
E-mail: karim.sharipov@narxoz.kz, https://orcid.org/0009-0003-2452-8803;
Куанышбай Дархан — PhD, ассистент-профессор, Университет Нархоз, Алматы, Казахстан,
E-mail: darkhan.kuanyshbay@sdu.edu.kz, https://orcid.org/0000-0001-5952-8609;
Мухаметжанов Асылбек — магистрант, SDU Университет, Каскелен, Казахстан,
E-mail: 221107046@stu.sdu.edu.kz, https://orcid.org/0009-0009-8528-9985.

Аннотация: Телефонное мошенничество (вишинг) представляет собой одну из наиболее распространённых форм социального инжиниринга, наносящую значительный финансовый и психологический ущерб. В условиях постоянной смены номеров и сценариев мошенников традиционные методы защиты оказываются недостаточно эффективными, что обусловливает необходимость автоматизированного анализа содержимого телефонных разговоров. В данной работе представлено экспериментальное исследование применения методов машинного обучения для автоматического выявления мошеннических телефонных звонков на основе текстовых транскрипций речи. Для проведения исследования сформирован сбалансированный русскоязычный корпус, включающий 1400 телефонных разговоров (мошеннические и легитимные звонки), собранных из открытых источников. Аудиозаписи были автоматически транскрибированы с использованием нейросетевой модели распознавания речи Whisper, после чего тексты подвергались нормализации и лемматизации. В качестве признакового представления использованы TF–IDF-признаки (униграммы и биграммы). На основе подготовленных данных были обучены и сравнены несколько моделей машинного обучения, включая Logistic Regression, Linear SVM, Multinomial Naive Bayes, Random Forest и XGBoost. Результаты экспериментов показали, что все модели демонстрируют высокую точность классификации, при этом наилучшие показатели достигнуты линейными моделями и классификатором Multinomial Naive Bayes с минимальным сглаживанием (accuracy до 94%, ROC–AUC до 0,99). Анализ лексических признаков позволил выявить устойчивые языковые маркеры мошеннической речи, характерные для типичных сценариев социального инжиниринга. Надёжность и обобщающая способность моделей подтверждены с использованием k-кратной кросс-валидации и анализа ROC–AUC. Полученные результаты свидетельствуют о высокой практической применимости предложенного подхода для автоматического выявления телефонного мошенничества в реальных условиях.

Ключевые слова: телефонное мошенничество; вишинг; машинное обучение; обработка естественного языка; распознавание речи; TF–IDF; классификация текстов

Введение. Телефонное мошенничество («вишинг», от voice phishing) стало серьезной угрозой, ежегодно приносящей жертвам значительные финансовые потери (Jones et al., 2021). Злоумышленники звонят, выдавая себя за банковских

сотрудников, коммунальные службы или другие организации, чтобы обманом получить конфиденциальные данные или деньги (Jones et al., 2021). Традиционные меры защиты – например, чёрные списки номеров – недостаточно эффективны, ведь мошенники часто меняют номера (Triantafyllopoulos et al., 2025). Поэтому все больше внимания уделяется автоматическому распознаванию мошеннических звонков по содержанию разговора (Triantafyllopoulos et al., 2025).

Современные технологии искусственного интеллекта (ИИ) и обработки речи позволяют анализировать сам текст разговора и манеру общения в реальном времени (Triantafyllopoulos et al., 2025; Kim et al., 2021; Lee et al., 2023). Уже появляются решения, способные во время звонка распознать характерные признаки «скрипта» мошенников и предупредить потенциальную жертву о угрозе. Такие системы внедряются сотовыми операторами – например, «СберМобайл», «Тинькофф Мобайл» и др. – и показывают высокую точность выявления злоумышленников, по данным открытых материалов операторов связи (Sberbank, 2026; T-Bank, 2026). Автоматизация распознавания мошенничества с помощью ИИ имеет значительный потенциал в повышении эффективности защиты, так как позволяет определять именно мошеннические паттерны речи, а не просто блокировать заранее известные номера. В данной работе представлено экспериментальное исследование применения методов машинного обучения для автоматического обнаружения мошеннических телефонных разговоров по их расшифровкам. Нами был собран корпус звонков, выполнена их транскрипция с помощью нейросетевой модели распознавания речи и обучение нескольких моделей бинарной классификации по тексту разговора.

Вклад настоящей работы заключается в следующем. Во-первых, был сформирован и подготовлен специализированный русскоязычный корпус телефонных разговоров, включающий как мошеннические, так и легитимные звонки, с последующей автоматической транскрипцией и текстовой обработкой. Во-вторых, проведено систематическое экспериментальное сравнение классических методов машинного обучения (Logistic Regression, Linear SVM, Multinomial Naive Bayes, Random Forest, XGBoost) для задачи обнаружения телефонного мошенничества на основе текстовых транскрипций. В-третьих, выполнен детальный анализ лексических паттернов мошеннической речи, что позволило выявить устойчивые словесные маркеры, характерные для сценариев социального инжиниринга. В-четвёртых, эмпирически показано, что относительно простые и интерпретируемые модели (в частности, Multinomial Naive Bayes и линейные классификаторы) способны достигать качества, сопоставимого или превосходящего более сложные ансамблевые методы. Наконец, устойчивость и обобщающая способность моделей были подтверждены с использованием k-кратной кросс-валидации и анализа ROC–AUC, что повышает надёжность полученных выводов и их практическую применимость.

Литературный обзор

Проблема телефонного мошенничества и социальной инженерии активно исследуется в последние годы в связи с глобальным ростом киберпреступности. Исследователи выделяют несколько ключевых направлений в разработке систем противодействия вишингу: анализ сетевых характеристик звонка, биометрический анализ голоса и семантический анализ содержания разговора.

Традиционные методы защиты, такие как создание «черных списков» номеров, демонстрируют снижающуюся эффективность из-за массового использования злоумышленниками технологий подмены номера (Caller ID spoofing) и IP-телефонии. Как отмечают Triantafyllopoulos и соавторы (2025), динамическая природа мошеннических сценариев требует перехода от статической фильтрации к интеллектуальному анализу контента в реальном времени.

Важным этапом в автоматизации процесса обнаружения мошенничества стало развитие технологий распознавания речи (ASR). Использование современных нейросетевых моделей, таких как Whisper от OpenAI, позволяет получать высокоточные текстовые транскрипции даже в условиях зашумленного аудиоканала телефонной связи. Исследования Radford et al. (2023) подтверждают, что масштабное обучение на слабо контролируемых данных делает такие модели устойчивыми к различным акцентам и диалектам, что критически важно для анализа русскоязычного сегмента звонков.

В области классификации текстов для выявления мошенничества (fraud detection) исследователи часто сравнивают эффективность классических алгоритмов машинного обучения и глубоких нейронных сетей. Несмотря на развитие трансформеров, работы Kowsari et al. (2019) и Ferhati et al. (2025) показывают, что методы на основе TF–IDF векторизации в сочетании с линейными моделями (SVM, Logistic Regression) или наивным байесовским классификатором (Multinomial Naive Bayes) остаются крайне эффективными для задач с высокой размерностью признаков и ограниченным объемом данных.

Особое внимание в литературе уделяется лексическим маркерам и сценариям общения. Исследования Jones et al. (2021) подчеркивают, что мошеннические звонки характеризуются использованием специфических «скриптов», направленных на создание чувства срочности или авторитетности. Ключевые слова, связанные с банковской безопасностью, подтверждением транзакций и персональными кодами, служат устойчивыми предикторами класса «мошенничество».

Таким образом, современные системы обнаружения вишинга эволюционируют в сторону комплексных решений, объединяющих автоматическое распознавание речи и продвинутые методы обработки естественного языка (NLP) для выявления скрытых паттернов в поведении злоумышленников.

Материалы и методы

Создание корпуса.

Для эксперимента сформирован сбалансированный датасет из 1400 записей телефонных разговоров на русском языке. Из них 700 – мошеннические звонки, например распространенные сценарии звонков псевдосотрудников банка, коммунальных служб и т.д., а 700 – обычные телефонные разговоры (Рис. 1).

```
Распределение классов:
label
fraud        700
not_fraud    700
Name: count, dtype: int64
```

Рисунок 1 – Распределение классов

Аудиозаписи были найдены на открытых источниках YouTube – например, на каналах, посвященных разоблачению телефонных мошенников, а также записи бытовых телефонных разговоров. Каждая аудиозапись была снабжена меткой класса: «fraud» (мошенничество) или «not_fraud» (нет мошенничества). Метки присваивались автоматически исходя из источника и контекста (например, файлы в папке с мошенническими звонками помечены как fraud).

Транскрипция аудио.

Все 1400 звонков были автоматически распознаны (транскрибированы в текст) с помощью модели Whisper (OpenAI) – нейросетевого преобразователя речи в текст (Radford, A. et al., 2023). Использовалась модель Whisper Small, обученная для русского языка, что обеспечило достаточно высокое качество распознавания при приемлемом времени обработки. Для каждого звонка получена текстовая расшифровка. Пример фрагмента транскрипта мошеннического звонка: «… Система контроля звонков в онлайне не будет работать в личных беседах клиента…» (звонящий пытается говорить официально, ссылаясь на некую «систему контроля»). Все полученные пары «текст – метка» сохранены в таблицу для дальнейшего использования в обучении моделей.

Предобработка текста.

Для последующего анализа тексты звонков были приведены к стандартизированному виду. Выполнены следующие шаги:

1. приведение к нижнему регистру;
2. удаление посторонних символов и пунктуации, оставлены только слова (включая последовательности букв и цифр);
3. лемматизация – приведение слов к начальной форме. Лемматизация выполнялась с помощью библиотеки Pymorphy3: например, фраза «картой в карте карт» преобразуется в «карта в карта». Это уменьшает вариативность словоформ в русском языке (Kowsari et al., 2019).
4. удаление стоп-слов – наиболее частотных и малоинформативных слов.

Мы использовали кастомный список из ~15 слов, включающий междометия, приветствия и общие слова (например, «здравствуйте», «алло», «да», «нет», «вот», «это», «как», «уже» и т.п.), которые часто встречаются в речи и не помогают отличить мошенников. После очистки и нормализации текстов каждая транскрипция была представлена как набор значимых токенов (слов) в базовой форме.

Признаки и представление данных.

Для преобразования текстов в формат, пригодный для обучения моделей, применен метод «Bag of Words» с взвешиванием TF–IDF (Kowsari et al., 2019). С помощью TfidfVectorizer были извлечены униграммы и биграммы (ngram_range = (1,2)) из текстов, учитывая слова, встречающиеся как минимум в 5 разговорах, и игнорируя слова, присутствующие в большинстве звонков (Kowsari et al., 2019). Таким образом, очень редкие слова и чрезмерно частые не учитывались как признаки.

Отбор терминов осуществлялся на этапе построения словаря признаков: в него включались только те униграммы и биграммы, которые встречались не менее чем в пяти различных разговорах (параметр *min_df = 5*), что позволяло исключить случайные, шумовые токены, не обладающие статистической значимостью для задачи классификации. Одновременно из словаря исключались термины, присутствующие более чем в 85% разговоров (параметр *max_df = 0.85*), поскольку такие слова не несут дискриминативной информации и одинаково характерны как для мошеннических, так и для легитимных звонков (Kowsari et al., 2019).

Данный механизм фильтрации не искажает семантическое представление данных, поскольку модель ориентируется не на отдельные редкие или общеупотребительные слова, а на устойчивые лексические паттерны, характерные для мошеннических сценариев общения. Напротив, исключение нерелевантных признаков снижает размерность пространства признаков, уменьшает уровень шума и способствует повышению обобщающей способности модели, снижая риск переобучения (Kowsari et al., 2019; Ferhati et al., 2025).

После векторизации каждый звонок представлен разреженным вектором большой размерности, где каждому допустимому слову или биграмме соответствует признак – TF–IDF значение (важность этого термина для данного разговора относительно всего корпуса).

Разбиение на выборки.

Исходный набор из 1400 примеров разделен на обучающую и тестовую выборки в отношении 70/30 с сохранением баланса классов. Затем тестовая часть (30% данных) была поровну разделена на Valid и Test (по 15% каждый). Валидационная выборка использовалась для контроля обучения ансамблевых моделей и подбора гиперпараметров, тогда как финальная оценка проводилась на независимой тестовой выборке. В итоге: 980 звонков в обучающей выборке и по 210 звонков в валидационной и тестовой.

Модели машинного обучения.

Для решения задачи бинарной классификации телефонных звонков на мошеннические и легитимные были протестированы несколько классических алгоритмов машинного обучения. Выбор данных методов обусловлен их широкой распространённостью в задачах анализа текстов, устойчивостью к высокой размерности признакового пространства и возможностью интерпретации полученных результатов (Kowsari et al., 2019; Ferhati et al., 2025).

1. Логистическая регрессия

Логистическая регрессия (Logistic Regression) является линейным классификатором, который моделирует вероятность принадлежности объекта к определённому классу как логистическую функцию от взвешенной линейной комбинации входных признаков. Несмотря на относительную простоту, данный метод демонстрирует высокую эффективность при работе с разреженными высокоразмерными признаками, характерными для текстовых данных (Kowsari et al., 2019; Ferhati et al., 2025).

В рамках данного исследования логистическая регрессия применялась с L2-регуляризацией, позволяющей ограничить величину коэффициентов модели и снизить риск переобучения. Интенсивность регуляризации задавалась параметром C = 0.01, что соответствует усиленному штрафу на сложность модели. Дополнительно использовалась балансировка классов, направленная на повышение чувствительности модели к мошенническому классу.

2. Random Forest

Random Forest представляет собой ансамблевый метод, основанный на построении множества решающих деревьев, обучаемых на различных бутстрап-выборках исходных данных (Ferhati et al., 2025). Итоговое решение формируется путём голосования отдельных деревьев.

Поскольку отдельные решающие деревья обладают высокой дисперсией, ансамблирование позволяет существенно повысить устойчивость модели. В данном исследовании глубина деревьев была ограничена значением 5, а количество деревьев установлено равным 1000, что позволило сдерживать переобучение при сохранении достаточной выразительной способности модели.

3. Линейный SVM

Метод опорных векторов (Support Vector Machine, SVM) является мощным алгоритмом обучения с учителем, основная цель которого заключается в нахождении оптимальной разделяющей гиперплоскости между классами. Для задач классификации текстов особенно эффективен линейный SVM, хорошо работающий с разреженными и высокоразмерными признаковыми представлениями.

В рамках эксперимента использовалась линейная версия SVM с коэффициентом регуляризации C = 0.025, обеспечивающим баланс между максимизацией зазора и минимизацией ошибки классификации. Максимальное число итераций оптимизации было увеличено до 3000 для гарантированной сходимости алгоритма.

4. XGBoost

XGBoost является одной из наиболее эффективных реализаций градиентного бустинга на решающих деревьях (Ferhati K. et al., 2025). Метод последовательно строит ансамбль моделей, каждая из которых корректирует ошибки предыдущих, минимизируя регуляризованную функцию потерь.

Для обеспечения устойчивого обучения были выбраны следующие гиперпараметры: 500 деревьев глубиной 3, малый шаг обучения (learning_rate = 0.02), а также бутстрапирование 80% объектов и признаков. Дополнительно применялась L1- и L2-регуляризация ($\alpha = 0.5$, $\lambda = 1.5$). Обучение проводилось с контролем качества на валидационной выборке; при этом признаков существенного переобучения выявлено не было.

5. Классификатор Multinomial Naive Bayes

Классификатор Multinomial Naive Bayes основан на теореме Байеса и предположении условной независимости признаков. Несмотря на данное упрощение, байесовские методы широко применяются в задачах классификации текстов и часто демонстрируют высокую эффективность.

В ходе эксперимента особое внимание уделялось параметру Лапласовского сглаживания (α), регулирующему влияние редко встречающихся терминов. Были протестированы значения $\alpha = 1.0$, $\alpha = 10.0$ и $\alpha = 0.1$. Большие значения α обеспечивают более «гладкие» вероятностные оценки и снижают влияние редких слов, тогда как малые значения позволяют модели учитывать редкие, но потенциально информативные лексические маркеры мошеннической речи.

Обучение и оценка всех моделей проводились в среде Jupyter Notebook с использованием библиотеки scikit-learn, а для градиентного бустинга применялся официальный Python-пакет xgboost. Для каждой модели были получены предсказания на тестовой выборке, рассчитаны метрики качества, построены матрицы ошибок и ROC-кривые.

Результаты

1. Общее качество классификации.

Полученные модели продемонстрировали высокую точность в распознавании мошеннических звонков по тексту. Практически все алгоритмы достигли accuracy выше 0.8, а лучшие – выше 0.9 (Lee et al., 2023; Chichwadia et al., 2024). Наилучший результат показал Naive Bayes с минимальным сглаживанием (α=0.1): 94% точности на тестовой выборке. При этом баланс метрик по классам также выдающийся: precision и recall около 0.90–0.98 для обоих классов, то есть модель почти одинаково хорошо выявляет мошеннические звонки и распознаёт легитимные разговоры. Это подтверждается матрицей ошибок для модели Naive Bayes ($\alpha = 0.1$), представленной на рисунке 2, где число ложных срабатываний и пропусков минимально.

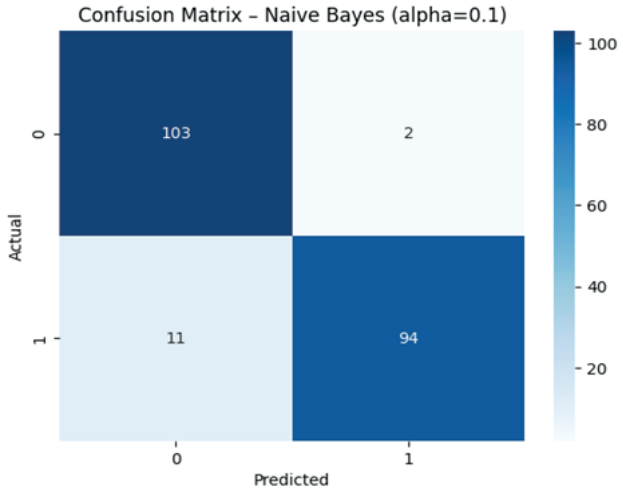


Рисунок 2 – Confusion Matrix - Naive Bayes (alpha = 0.1)

Linear SVM и XGBoost уступили совсем немного – они показали ~92% точности. Логистическая регрессия правильно классифицировала около 90% звонков. Несколько хуже выступил Random Forest (более сложная нелинейная модель) – порядка 84% правильных классификаций. Для наглядности сводка результатов представлена в таблице 1. Помимо accuracy в таблице представлены значения F1-score для класса *fraud*, что позволяет более объективно оценить баланс между точностью и полнотой моделей при обнаружении мошеннических звонков.

Таблица 1 – Метрики качества моделей машинного обучения при классификации телефонных звонков

Модель	Accuracy (точность)	Precision (fraud)	Recall (fraud)	F1-score (fraud)	Precision (not_fraud)	Recall (not_fraud)
Naive Bayes (α=0.1)	0.94	0.90	0.98	0.94	0.98	0.90
Linear SVM	0.92	0.94	0.90	0.92	0.90	0.94
XGBoost	0.92	0.92	0.92	0.92	0.92	0.92
Logistic Regression	0.90	0.92	0.88	0.90	0.88	0.92
Random Forest (деревья=1000)	0.84	0.85	0.83	0.84	0.83	0.86
Naive Bayes (α=1.0)	0.87	0.80	0.98	0.88	0.98	0.76
Naive Bayes (α=10.0)	0.76	0.67	1.00	0.80	1.00	0.51

Как видно, все модели, кроме сильно сглаженного Naive Bayes (α=10), обеспечили высокий recall для мошеннических звонков – 0.88 и выше, что важно с практической точки зрения (не пропускать мошенников)

(Triantafyllopoulos et al., 2025; Chichwadia et al., 2024). Однако при сильном сглаживании (α=10) Naive Bayes фактически склонен всегда относить звонок к классу «мошенничество», отсюда recall(fraud)=1.0 ценой очень низкого recall для нормальных звонков (лишь ~51%). Напротив, без сглаживания (α=0.1) Naive Bayes добивается баланса ~0.9–0.98 на обоих классах, показывая, что учет даже редких слов сыграл роль в идентификации.

Другие алгоритмы также достаточно сбалансированы: например, SVM имеет precision и recall ~0.90–0.94 для обеих категорий, что значит минимум ошибок как первого, так и второго рода. Logistic Regression дает f1-score = 0.90 для каждого класса– близко к SVM, подтверждая, что линейные модели хорошо справляются с разделением данных. Хуже баланс у Random Forest: его precision/recall ~0.84–0.86, то есть он чаще ошибается (в сравнении с линейными методами). Вероятно, ограничение глубины деревьев (до 5) не позволило Random Forest уловить некоторых важных комбинаций признаков, а большее углубление могло вести к переобучению на небольшой выборке.

2. Анализ ключевых признаков.

Модели машинного обучения позволили также выяснить, какие слова или фразы наиболее характерны для мошеннических звонков по сравнению с обычными. Логистическая регрессия и SVM (линейные модели) присвоили наименьшие отрицательные веса следующим словам: «**код**», «**номер**», «**посылка**», «**ключ**», «**письмо**» и т.п. – то есть лексике, связанной с банковскими операциями, безопасностью счетов, заявками на услуги и др. Эти термины часто фигурируют в речевых скриптах мошенников (Jones et al., 2021; Lee et al., 2023). Например, требование сообщить код из СМС, разговоры о блокировке номера телефона, заявке на отключение услуг и т.д. Данный вывод наглядно иллюстрируется графиками лексических признаков класса fraud, представленными на рисунках 3 и 4, где показаны слова с наибольшими по модулю весами в линейных моделях.

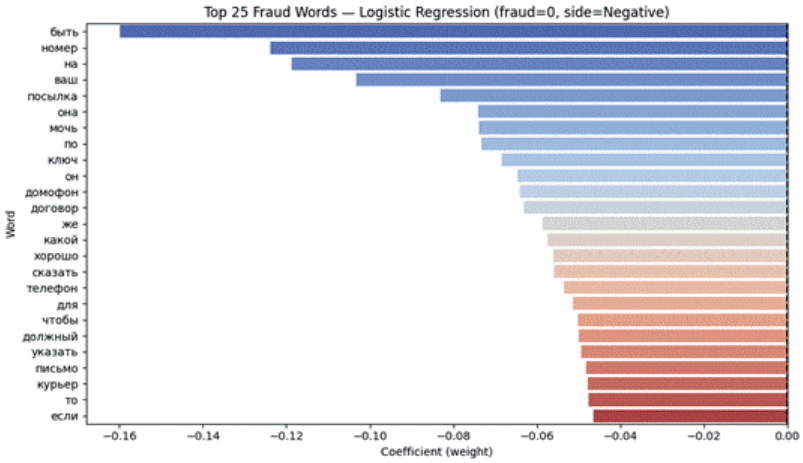


Рисунок 3 – Лексические признаки класса fraud (Logistic Regression)

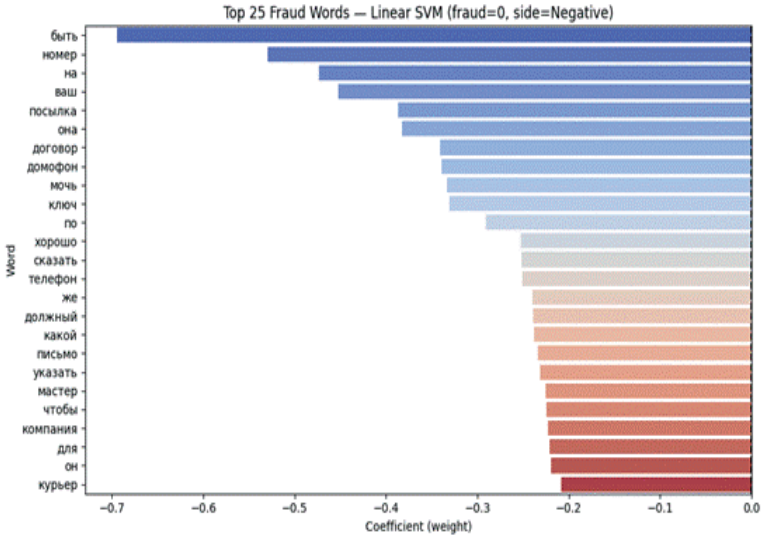


Рисунок 4 – Лексические признаки класса fraud (Linear SVM)

Напротив, слова, получившие положительные веса (ассоциированные с классом обычных звонков) – это, как правило, обращения по имени, разговорная лексика, неформальные выражения. Это наглядно показано на рисунках 5 и 6, где приведены топ-25 лексических признаков класса not_fraud, выявленных линейным SVM и логистической регрессией соответственно. Интересно, что встречаемость имени-отчества собеседника оказалась существенным индикатором: мошенники часто обращаются формально («Галина Петровна», «Роман Шарафудинович»), тогда как в бытовых разговорах такой стиль редок.

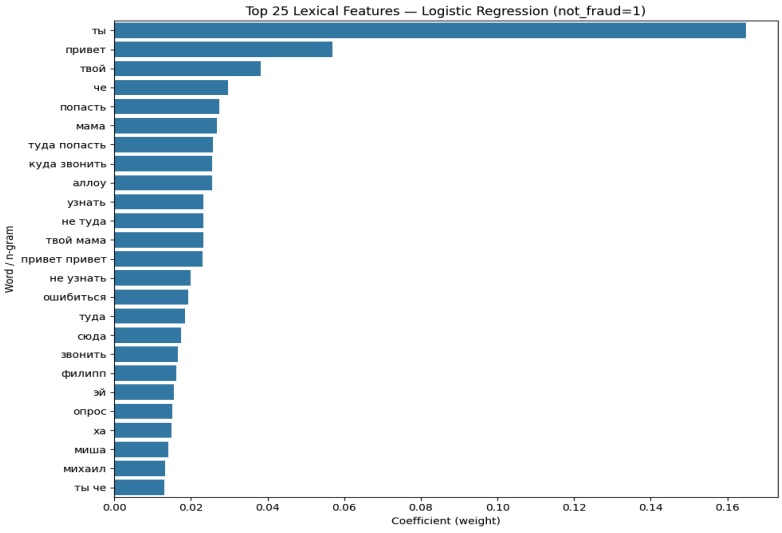


Рисунок 5 - Лексические признаки класса not_fraud (Logistic Regression)

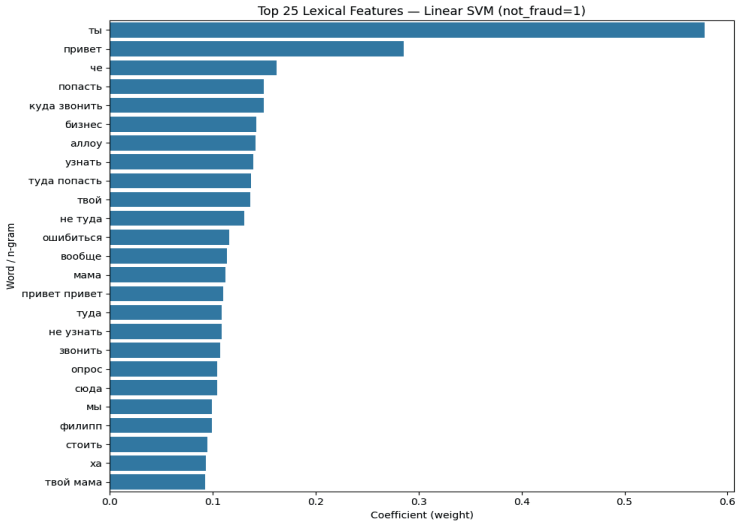


Рисунок 6 – Лексические признаки класса not_fraud (Linear SVM)

Классификатор Naive Bayes подтвердил эти наблюдения: разница лог-вероятностей показала, что слова вроде «ключ», «посылка», «курьер», «письмо» и т.п. намного более вероятны в мошеннических звонках, тогда как слова «привет», «дома», «алло» и т.п чаще встречаются в нормальных разговорах. Это наглядно иллюстрируется на рисунке 7, где представлены наиболее информативные лексические признаки, отсортированные по разнице лог-вероятностей между классами для модели Naive Bayes. Таким образом, модели научились узнавать мошенничество «по словам» – специфическому набору терминов и формулировок.

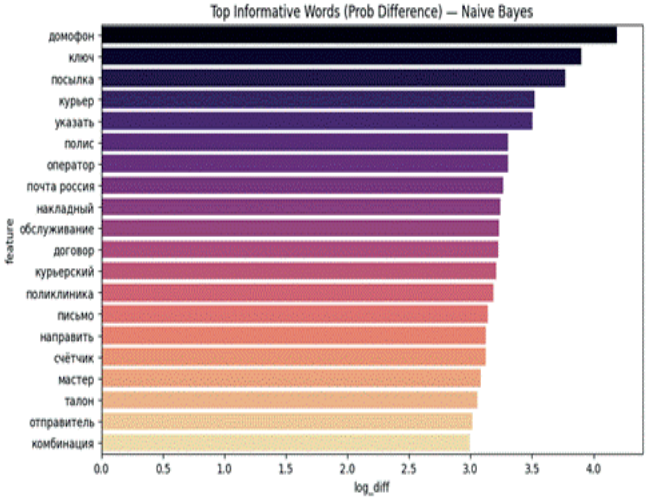


Рисунок 7 - Лексические признаки класса fraud (Naive Bayes)

3. ROC–AUC и устойчивость моделей.

Для оценки устойчивости и обобщающей способности моделей дополнительно была проведена *k-кратная кросс-валидация* на обучающей части датасета. Использовалась стратифицированная схема разбиения, обеспечивающая сохранение баланса классов «fraud / not_fraud» в каждом фолде. Значение параметра *k* было установлено равным 5, что является компромиссом между вычислительной сложностью и надёжностью оценки.

В рамках кросс-валидации для каждой модели вычислялись средние значения accuracy и стандартное отклонение по всем фолдам. В качестве признакового представления использовались те же TF–IDF униграммы и биграммы, а гиперпараметры моделей соответствовали настройкам, описанным в разделе «Методы». Это позволило оценить именно устойчивость моделей, а не эффект подбора параметров под конкретное разбиение данных.

Результаты кросс-валидации показали, что все рассмотренные модели демонстрируют стабильное качество классификации с относительно малым разбросом метрик между фолдами. Наиболее устойчивыми оказались линейные модели и классификатор Multinomial Naive Bayes с минимальным сглаживанием (α=0.1). Так, Linear SVM и Logistic Regression продемонстрировали среднюю accuracy на уровне ~0.90–0.92 при стандартном отклонении менее 0.02, что указывает на хорошую обобщающую способность и низкую чувствительность к выбору обучающей подвыборки.

Классификатор Multinomial Naive Bayes (α = 0.1), показавший наилучший результат на тестовой выборке, также подтвердил свою устойчивость в условиях кросс-валидации: средняя accuracy составила более 0.93 при минимальном разбросе значений. Это свидетельствует о том, что высокая точность данной модели не является следствием удачного разбиения данных, а отражает наличие устойчивых лексических паттернов мошеннической речи в корпусе.

Ансамблевые методы (Random Forest и XGBoost) продемонстрировали несколько больший разброс результатов между фолдами. Особенно это характерно для Random Forest, где вариативность accuracy была выше по сравнению с линейными моделями. Вероятной причиной является чувствительность деревьев решений к конкретному составу обучающих данных при высокой разреженности признакового пространства.

В целом результаты кросс-валидации подтверждают выводы, полученные на отложенной тестовой выборке: модели не переобучаются, демонстрируют устойчивое качество и способны обобщать выявленные закономерности на ранее невидимые данные. Это также подтверждается ROC-кривыми, представленными на рисунке 8, которые демонстрируют высокую площадь под кривой для всех рассмотренных моделей и близость кривых к левому верхнему углу, что указывает на хорошую разделимость классов. Это особенно важно с практической точки зрения, поскольку в реальных условиях

распределение телефонных разговоров может отличаться от обучающего корпуса.

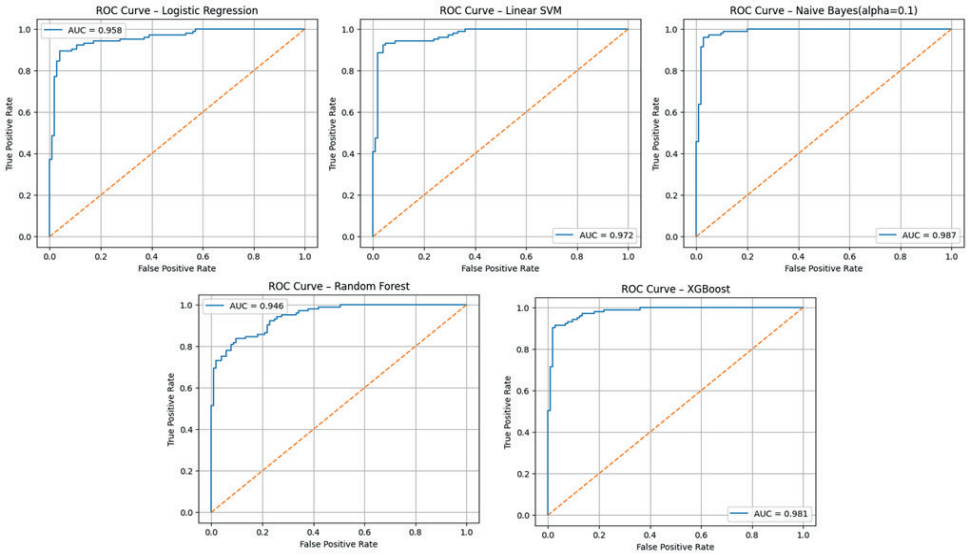


Рисунок 8 - ROC-кривые моделей машинного обучения для классификации телефонных звонков

Для количественной оценки качества помимо Accuracy рассмотрены площади под ROC-кривой. Все модели продемонстрировали очень высокие значения AUC: от ~0.95 у Random Forest до ~0.98–0.99 у лучших моделей. Так, ROC–AUC для SVM составил ~0.972, для XGBoost ~0.981, а наивысший результат дал Naive Bayes (α=0.1) – 0.987. Это указывает на отличную разделимость классов: даже при изменении порога классификации алгоритмы способны достичь высокой чувствительности при приемлемом количестве ложных тревог. В нашем сбалансированном наборе выбор порога 0.5 оказался близким к оптимальному (что отражено в precision/recall), но на практике при сильном дисбалансе (мошенников намного меньше, чем обычных звонков) порог можно было бы настроить для снижения false positives.

Высокий AUC Naive Bayes даже при среднем Accuracy=0.87 (при α=1) намекает, что эта модель ошибалась систематически из-за несбалансированности или выбора порога, хотя ранжировала разговоры довольно хорошо. Правильная калибровка вероятностей или подбор α позволили существенно повысить его качество до уровня лучших моделей.

Обсуждение

Эксперимент показал, что автоматическое распознавание телефонного мошенничества по речевым данным является практически осуществимой задачей и может достигать высоких значений F1-score (0,90–0,94). Использование данной метрики является принципиально важным, поскольку

в задачах обнаружения мошенничества классы, как правило, распределены неравномерно, и метрика accuracy может давать завышенную оценку качества модели за счёт доминирующего класса (Triantafyllopoulos et al., 2025).

Особенно показателен высокий F1-score, достигнутый наивным байесовским классификатором (до 0,94 при оптимальной настройке), несмотря на простоту модели. Это указывает на то, что наличие отдельных ключевых слов и устойчивых лексических маркеров играет решающую роль в определении мошеннического характера звонка (Jones et al., 2021; Lee et al., 2023; Chichwadia et al., 2024). Байесовский подход эффективно использует такие слова-«триггеры» (например, упоминания банковских терминов или призывов к срочным действиям), предполагая условную независимость признаков и практически не учитывая сложные взаимосвязи между ними.

Линейные модели, включая SVM и логистическую регрессию, также продемонстрировали высокие значения F1-score, что подтверждает их способность назначать информативные веса отдельным терминам и тем самым выявлять характерный «лексический почерк» мошеннических сценариев.

В то же время более сложные нелинейные методы, такие как XGBoost, не обеспечили значимого прироста F1-score (Moussavou Boussougou et al., 2023). Это позволяет предположить, что в рассматриваемой задаче дискриминативная информация сосредоточена преимущественно в отдельных словах и простых n-граммах, а не в сложных нелинейных комбинациях признаков. Аналогичный вывод подтверждается и относительно скромными результатами случайного леса: увеличение глубины деревьев не приводит к росту F1-score без существенного расширения обучающей выборки, тогда как неглубокие ансамбли уступают линейным моделям по способности точно балансировать между полнотой и точностью.

1. Практическая значимость.

Полученные результаты вселяют уверенность, что интеграция подобных моделей в системы телефонии сможет повысить защиту абонентов (Kim et al., 2022). Например, модель может в режиме реального времени анализировать транскрибируемый разговор и выводить предупреждение, если обнаружен типичный паттерн мошенничества (как уже реализовано в пилотных проектах операторов 2) (Tlahun et al., 2025). При точности ~90% ложные срабатывания будут случаться, но их доля может быть приемлемой с учетом ценности предотвращенного мошенничества. Важным аспектом здесь является баланс: что хуже – пропустить мошенника или зря потревожить клиента предупреждением? Наши модели можно настроить под нужный компромисс. Например, при чуть пониженном пороге вероятности мошенничества recall возрастет почти до 100%, но и вероятность возникновения ложных тревог станет больше. Напротив, повышение порога обеспечит практически полное отсутствие ложных тревог ценой некоторого снижения чувствительности. В реальной эксплуатации оптимум зависит от терпимости пользователей к предупреждениям и от вреда от пропущенного звонка.

2. Ограничения исследования.

Следует отметить, что наше исследование проводилось на относительно небольшом датасете (Dai et al., 2025). Собранные нами 700 мошеннических разговоров могут не охватывать всего разнообразия сценариев. Возможно, многие из них – вариации популярных схем, таких как «звонок из банка», «служба безопасности карты», «лжесотрудник энергокомпании» и прочие, и модель могла обучиться распознавать именно эти сюжеты. Встреться ей иной тип мошенничества, качество может снизиться (Li et al., 2025; Figueiredo et al., 2024). Аналогично, нормальные разговоры в датасете в основном бытовые и дружеские беседы. Они явно контрастируют с официально вежливым стилем мошенников, что облегчает классификацию. Но в реальной жизни бывают легитимные звонки от банков, магазинов, сервисных служб – по лексике похожие на мошеннические. Наши модели пока не учились отличать настоящего банковского оператора от мошенника, если оба говорят о картах и счетах. Для решения этой более тонкой задачи понадобятся дополнительные признаки (например, акустические: интонация, уверенность голоса; или анализ диалога: мошенники чаще давят на срочность, не дают перезвонить и т.д.) (Kim et al., 2025; Kang et al., 2022; Elizalbe et al., 2021).

Кроме того, качество автоматической транскрипции влияет на результаты. Мы использовали модель Whisper Small, и, хотя она достаточно точна, ошибки распознавания речи всё же присутствуют. В наших экспериментах это не помешало алгоритмам – даже с опечатками и пропусками модель выявляла ключевые слова. Однако в условиях шума или плохой связи распознавание может сильно ухудшиться, что снизит надежность классификации. В будущем стоит протестировать более крупные модели ASR (Whisper Large) либо специальные модели, обученные на телефонных переговорах, чтобы повысить точность транскрипции (Sim et al., 2025).

Заключение

В работе показана возможность эффективного автоматического обнаружения телефонного мошенничества на основе текстовых транскрипций разговоров с использованием методов машинного обучения. На основе сбалансированного русскоязычного корпуса и TF–IDF представления текстов были обучены и сравнены несколько классических моделей, продемонстрировавших высокое качество классификации. Наилучшие результаты показал классификатор Multinomial Naive Bayes с минимальным сглаживанием (accuracy выше 0.9 и ROC–AUC до 0.99), при этом линейные модели также подтвердили свою устойчивость и эффективность.

Анализ лексических признаков выявил устойчивые словесные маркеры, характерные для мошеннических сценариев социального инжиниринга, что подтверждает интерпретируемость и практическую применимость предложенного подхода. Результаты k-кратной кросс-валидации и анализа ROC–AUC показали хорошую обобщающую способность моделей и отсутствие переобучения. В целом полученные результаты свидетельствуют

о возможности использования текстового анализа телефонных разговоров в практических системах выявления мошенничества и создают основу для дальнейшего развития мультимодальных решений.

References

Chichwadia A.E., & Mpekoa N. (2024) Detecting smishing and vishing attacks using machine learning. International Journal of Intelligent Computing Research, 15(1). — P. 1234–1241. https://doi.org/10.20533/ijicr.2042.4655.2024.0151 (in Eng.).

Dai H., Liu Z., Liao W., et al. (2025) AugGPT: Leveraging ChatGPT for text data augmentation. IEEE Transactions on Big Data, 11(5). — P. 907–918 (in Eng.).

Elizalde B., & Emmanouilidou D. (2021) Detection of robocall and spam calls using acoustic features of incoming voicemails. Proceedings of Meetings on Acoustics, 45, 060004 (in Eng.).

Ferhati K., Burlea-Schiopoiu A., & Nascu A.-G. (2025) A text-based project risk classification system using multi-model AI: Comparing SVM, logistic regression, random forests, naive Bayes, and XGBoost. Systems, 13(12). — 1078 p. https://doi.org/10.3390/systems13121078 (in Eng.).

Figueiredo J., Carvalho A., Castro D., Gonçalves D., & Santos N. (2024) On the feasibility of fully AI-automated vishing attacks. arXiv preprint. https://arxiv.org/abs/2409.13793 (in Eng.).

Jones K.S., Armstrong M.E., Tornblad M.K., & Siami Namin A. (2021) How social engineers use persuasion principles during vishing attacks. Information & Computer Security, 29(2). — P. 314–331 (in Eng.).

Kang Y., Kim W., Lim S., Kim H., & Seo H. (2022) DeepDetection: Privacy-enhanced deep voice detection and user authentication for preventing voice phishing. Applied Sciences, 12(22), 11109. https://doi.org/10.3390/app122211109 (in Eng.).

Kim J., Gu S., Kim Y., Lee S., & Kang C. (2025) A multimodal voice phishing detection system integrating text and audio analysis. Applied Sciences, 15(20). — 11170 p. https://doi.org/10.3390/app152011170 (in Eng.).

Kim J., Kim J., Wi S., Kim Y., & Son S. (2022) HearMeOut: Detecting voice phishing activities in Android. In Proceedings of the 20th International Conference on Mobile Systems, Applications and Services (MobiSys). — P. 422–435 (in Eng.).

Kim J.-W., Hong G.-W., & Chang H. (2021) Voice recognition and document classification-based data analysis for voice phishing detection. Human-Centric Computing and Information Sciences, 11, Article 45. https://doi.org/10.1186/s13673-021-00245-9 (in Eng.).

Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., & Brown D. (2019) Text classification algorithms: A survey. Information, 10(4). — 150 p. https://doi.org/10.3390/info10040150 (in Eng.).

Lee M., & Park E. (2023) Real-time Korean voice phishing detection based on machine learning approaches. Journal of Ambient Intelligence and Humanized Computing, 14. — P. 1–13. https://doi.org/10.1007/s12652-021-03587-x (in Eng.).

Li W., Manickam S., Chong Y.W., & Karuppayah S. (2025) Talking like a phisher: LLM-based attacks on voice phishing classifiers. arXiv preprint. https://arxiv.org/abs/2507.16291(in Eng.).

Moussavou Boussougou M.K., & Park D.J. (2023) Attention-based 1D CNN–BiLSTM hybrid model enhanced with FastText word embedding for Korean voice phishing detection. Mathematics, 11(14), 3217. https://doi.org/10.3390/math11143217 (in Eng.).

Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., & Sutskever I. (2023) Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning. — P. 28492–28518. PMLR (in Eng.).

Sberbank (2026) Telefonnoye moshennichestvo: masshtaby problemy i mery protivodeystviya [Telephone fraud: Scale of the problem and countermeasures]. Official website of PJSC Sberbank. https://www.sberbank.ru/ru/sberpress/all/article?blockID=1303&lang=ru&newsID=8066ed65-189c-4b5e-a2b6-15513e6b62e8 (in Russian)

Sim J.-Y., & Kim S.-H. (2025) Detecting voice phishing with precision: Fine-tuning small language models. arXiv preprint. https://arxiv.org/abs/2506.06180 (in Eng.).

T-Bank (2026) Kak zashchitit' sebya ot moshennikov i sokhranit' den'gi [How to protect yourself from fraudsters and save money]. Official blog of T-Bank. https://www.tbank.ru/finance/blog/save-money/#q3 (in Russian)

Tlahun A.Z., Sumbiri D., & Jonathan K.N. (2025) Identifying and evaluating the best ML predictive models for detecting voice (phone-call) vishing attacks on MoMo users in real time. Journal of Information and Technology, 5(6). — P. 34–43. https://doi.org/10.70619/vol5iss6pp34-43 (in Eng.).

Triantafyllopoulos A., Spiesberger A.A., Tsangko I., Jing X., Distler V., Dietz F., Alt F., & Schulle B.W. (2025) Vishing: Detecting social engineering in spoken communication—A first survey and urgent roadmap. Computer Speech & Language, 94. — 101802 p. (in Eng.).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 393–411

https://doi.org/10.32014/2026.2518-1726.419

IRSTI: 20.23.27
UDC 004.93

© Shynzhigit B.B., Balabekova M.O.*, Amangeldy T.T., 2026.
M. Auezov South-Kazakhstan Research University, Shymkent, Kazakhstan.
E-mail: dadyiba@mail.ru

ANALYSIS AND FORECASTING OF BRICK PRODUCT SALES USING MACHINE LEARNING MODELS

Shynzhigit Bekmurat — doctoral student (PhD) at the department of Automation, Telecommunications and Control, M. Auezov South-Kazakhstan Research University, Shymkent, Kazakhstan,
E-mail: bekmuratsinzigit@gmail.com, https://orcid.org/0009-0001-2970-9967;
Balabekova Madina — candidate of technical sciences, assistant professor at the department of Automation, Telecommunications and Control, M. Auezov South-Kazakhstan Research University, Shymkent, Kazakhstan,
E-mail: dadyiba@mail.ru , 870531949086, https://orcid.org/0000-0002-5558-3472;
Amangeldy Togzhan — doctoral student (PhD) at the department of Automation, Telecommunications and Control, M. Auezov South-Kazakhstan Research University, Shymkent, Kazakhstan,
E-mail: togzhan.amangeldi1@mail.ru, https://orcid.org/0009-0001-5081-9466.

Abstract. This article examines in detail the problem of mathematical modeling and forecasting the sales volume of brick products for an industrial enterprise under conditions of high market volatility and a limited empirical database. The study relevance determined by the need to shift management processes in the construction industry from intuitive and heuristic approaches to formalized computational models, which make it possible to increase the validity of managerial decision-making. Within the engineering paradigm, advertising factors are regarded as manageable input parameters of a multifaceted production-economic system, while the system's output corresponds to the volume of product sales. To resolve the formulated task, a polynomial regression technique realized in a machine-learning setting is utilized. The model is constructed in an expanded feature space that encompasses nonlinear terms and interaction effects among factors, thus allowing separate evaluation of each advertising channel as well as their joint influence. Construction of the feature set, model fitting, and performance evaluation are carried out with the Scikit-learn library, employing conventional regression diagnostics - mean absolute error and mean squared error—to gauge adequacy. The main study result is the empirical confirmation of a synergistic effect of advertising factors on sales volume that cannot be identified by linear models. It has been shown that employing polynomial

regression lowers prediction errors and enhances model stability even when data are scarce. These findings can inform the design of decision-support tools for scheduling advertising initiatives, projecting sales volumes, and refining marketing tactics in firms that manufacture construction materials.

Keywords: sales forecasting, polynomial regression, data analysis, statistical modeling, advertising activity, ceramic brick

For citations: Shynzhigit B.B., Balabekova M.O., Amangeldy T.T. Analysis and forecasting of brick product sales using machine learning models. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 393–411. DOI: https://doi.org/10.32014/2026.2518-1726.419

© Шынжігіт Б.Б., Балабекова М.О.*, Амангелді Т.Т., 2026.
М. Әуезов атындағы Оңтүстік Қазақстан зерттеу университеті,
Шымкент, Қазақстан.
E-mail: dadyiba@mail.ru

КІРПІШ ӨНІМДЕРІН САТУ КӨЛЕМДЕРІН МАШИНАЛЫҚ ОҚЫТУДА ТАЛДАУ ЖӘНЕ БОЛЖАМДАУ

Шынжігіт Бекмұрат — «Автоматтандыру, телекоммуникация және басқару» кафедрасының докторанты, М. Әуезов атындағы Оңтүстік Қазақстан ғылыми-зерттеу университеті, Шымкент, Қазақстан,
E-mail: bekmuratsinzigit@gmail.com, https://orcid.org/0009-0001-2970-9967;
Балабекова Мадина — техника ғылымдарының кандидаты, «Автоматтандыру, телекоммуникация және басқару» кафедрасының доценті, М. Әуезов атындағы Оңтүстік Қазақстан ғылыми-зерттеу университеті, Шымкент, Қазақстан,
E-mail: dadyiba@mail.ru, https://orcid.org/0000-0002-5558-3472;
Амангелді Тоғжан — «Автоматтандыру, телекоммуникация және басқару» кафедрасының докторанты, М. Әуезов атындағы Оңтүстік Қазақстан ғылыми-зерттеу университеті, Шымкент, Қазақстан,
E-mail: togzhan.amangeldi1@mail.ru, https://orcid.org/0009-0001-5081-9466.

Аннотация. Бұл мақалада нарықтың үнемі өзгеріп отыратындығы, эмпирикалық мәліметтердің шектеулі кезінде кірпіш өнімдерін өндіретін мекемелердің сатылым көлемін математикалық модельдеу, болжам жасау мәселесі талқыланады. Құрылыс саласындағы басқару шешімдерін нақтылау мақсатында интуитивтік, эвристикалық әдіс-тәсілдерін фортальді түрде есептеу модельдеріне қарай өту қажеттілігімен айқындалып, анықталады. Инженерлік практикалық көзқараспен жарнамалық белсенділік өндірістік-экономикалық жасалатын жүйенің басқаруға мүмкіндік беретін кіріс параметрлерінің бір тұтас жиыны болып табылады. Бұл параметрлердің өзгеруі өнімдердің сатылу көлемі, соңында анықталатын жалпы қорытынды нәтижесіне әсер етеді. Параметрлердің бір-біріне тәуелділігі полиномдық регрессия модельдеудің көмегімен зерттелді. Нәтижелік модельді жасау

кезінде сызықтық, сызықтық емес тәуелділіктер қарастырылды. Сонымен қатар, әрбір факторлардың бір-біріне байланыстары да модельге енгізілді. Бұл әдістің көмегімен жарнаманың әрбір каналының сатылымға қалай әсер ететіндігін жеке-жеке қарастыруға болады. Сондай-ақ, олардың сатылым көлеміне қалай әсер ететіндігін бағалауға, басқаруға мүмкіндік береді. Нәтижесінде алынған модельді оқыту, қасиеттер кеңістігін жасау, модельдің сапасын бақылау, бағалау Scikit-learn кітапханасының көмегімен жүзеге асырылды. Орташа авсолюттік қателік және орташа квадраттық ауытқу деген регрессиялық талдаудың стандартты метрикалары тиімді критерийлері болып табылды. Сызықтық модельдеудің көмегімен анықталмайтын, жарнамалық факторлардың өнімдердің сатылу мөлшеріне синергетикалық әсерінің бар болғандығы жөніндегі гипотеза эксперимент жасау арқылы расталады. Бұл гипотезаның дәлелденуі зерттеудің ең басты қорытынды нәтижесі. Полиномдық регрессияны қолдану арқылы болжау қателіктерін барынша азайтуға, модельдің бір қалыптылығын сақтауға септігін тигізеді. Яғни, енгізілетін негізгі мәлеметтер көлемі шектеулі болған кезде ғана іске асады. Құрылыс саласындағы кәсіпорындарға арнап жарнамалық науқандарын жоспарлауға арналған нақты шешім қабылдау жүйелерін құрастырып, пайдалануға ұсыну зерттеудің нәтижесі болып табылады. Сонымен қатар сатылым мөлшерін болжауға себеп болатын модельдерді, жүйелерді құрастыруға да қолданыладыНәтижеде алынған бағалаулар маркетингтік стратегияларды өңдеуге және сатылым процессін басқарудың тиімділігін арттыруға негіз бола алады.

Түйін сөздер: сатылымды болжау, полиномиальді аппроксимация, мәліметтерді талдау, статистикалық модельдеу, жарнамалық белсенділік, керамикалық кірпіш

© Шынжігіт Б.Б., Балабекова М.О.*, Амангелді Т.Т., 2026.

Южно-Казахстанский исследовательский университет имени М. Ауэзова,
Шымкент, Казахстан.
E-mail: dadyiba@mail.ru

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ОБЪЕМОВ ПРОДАЖ КИРПИЧНОЙ ПРОДУКЦИИ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Шынжігіт Бекмұрат — докторант кафедры «Автоматизации, телекоммуникации и управления», Южно-Казахстанский научно-исследовательский университет имени М.Ауэзова, Шымкент, Казахстан,
E-mail: bekmuratsinzigit@gmail.com, https://orcid.org/0009-0001-2970-9967;
Балабекова Мадина — кандидат технических наук, доцент кафедры «Автоматизации, телекоммуникации и управления», Южно-Казахстанский научно-исследовательский университет имени М.Ауэзова, Шымкент, Казахстан,
E-mail: dadyiba@mail.ru , 870531949086, https://orcid.org/0000-0002-5558-3472;

Амангелді Тоғжан — докторант кафедры «Автоматизации, телекоммуникации и управления», Южно-Казахстанский научно-исследовательский университет имени М.Ауэзова, Шымкент, Казахстан,
E-mail: togzhan.amangeldi1@mail.ru, https://orcid.org/0009-0001-5081-9466.

Аннотация. В статье рассматривается задача математического моделирования и прогнозирования объёмов продаж кирпичной продукции промышленного предприятия в условиях высокой изменчивости рыночной среды и ограниченности эмпирических данных. Актуальность исследования обусловлена необходимостью перехода от интуитивных и эвристических методов управления к формализованным вычислительным моделям, обеспечивающим повышение обоснованности управленческих решений в строительной индустрии. В работе рекламные расходы рассматриваются как управляемые факторы, влияющие на объём продаж продукции. Изменение затрат на рекламу непосредственно отражается на уровне реализации. Для анализа данной зависимости применена полиномиальная регрессионная модель, позволяющая учитывать как линейные, так и нелинейные эффекты, а также взаимодействие между различными рекламными каналами. Предложенный подход позволяет оценить вклад каждого канала рекламы и определить их синергетическое влияние на объём продаж. Формирование признакового пространства, обучение модели и оценка её качества выполнены с использованием библиотеки Scikit-learn. В качестве критериев эффективности использованы стандартные метрики регрессионного анализа, включая среднюю абсолютную ошибку (MAE) и среднеквадратическую ошибку (RMSE). Ключевым результатом исследования является экспериментальное подтверждение наличия синергетического эффекта рекламных факторов, который не выявляется в рамках линейных моделей. Применение полиномиальной регрессии позволило снизить ошибку прогнозирования и обеспечить устойчивость модели даже при ограниченном объёме исходных данных. Разработанная модель может быть интегрирована в системы поддержки принятия решений, используемые при планировании рекламных кампаний и прогнозировании объёмов продаж на предприятиях строительной отрасли. Полученные результаты также могут применяться для корректировки маркетинговой стратегии и повышения эффективности управления продажами.

Ключевые слова: прогноз продаж, полиномиальная аппроксимация, анализ данных, статистическое моделирование, рекламная активность, керамический кирпич

Введение. В последние годы на промышленных предприятиях всё чаще используются цифровые методы анализа данных. Их применение расширяет возможности анализа и прогнозирования технико-экономических показателей предприятий строительной индустрии. Предприятия работают в условиях высокой динамики рынка и ограниченности ресурсов. Управленческие реше-

ния нередко приходится принимать быстро. В таких условиях интуитивные и эмпирические методы управления оказываются недостаточными. Поэтому всё большее значение приобретают формализованные вычислительные модели (Armstrong, 2017; Makridakis, 2018). Для кирпичных заводов, расположенных в регионах с активным строительством, особую роль играют модели прогнозирования спроса. Такие модели позволяют оценивать возможные объёмы реализации продукции с учётом внешних и управляемых факторов (Montgomery, 2015; Hyndman, 2021). Объёмы продаж кирпичной продукции напрямую влияют на работу предприятия. От них зависит планирование производственных мощностей. Они также определяют загрузку технологических линий, организацию перевозок и финансовую устойчивость предприятия.

В условиях колебаний спроса и усиления конкурентной среды отсутствие достоверного прогноза может приводить к перерасходу ресурсов или недоиспользованию производственного потенциала. К числу управляемых факторов, влияющих на объёмы продаж, относятся рекламные воздействия, реализуемые через различные информационные каналы. Обычно рекламная активность рассматривается в рамках экономического анализа и маркетинговых исследований. Однако с инженерной точки зрения её можно представить как входной параметр сложной системы, формирующей выходной показатель - объём продаж. В этом случае рекламные бюджеты выступают как количественно измеримые управляющие воздействия. Они влияют на состояние системы и могут изменять динамику продаж во времени. Воздействие рекламных каналов имеет нелинейный характер. Кроме того, между различными каналами могут возникать взаимные зависимости (Baecke, 2013; Smagulov, 2022). Эффективность одного рекламного инструмента часто меняется при одновременном использовании другого. В результате возникает совместный эффект воздействия каналов рекламы. Что сложно описать с помощью классических линейных моделей. Линейные модели предполагают независимое и аддитивное влияние факторов и поэтому не всегда отражают реальную структуру взаимодействия рекламных каналов.

Проблемная ситуация, характерная для кирпичного завода ТОО «Батсу» города Шымкент, заключается в отсутствии алгоритмически обоснованного инструмента прогнозирования продаж, способного учитывать нелинейные эффекты и взаимодействие рекламных факторов. На практике для анализа часто используются эмпирические оценки и линейные методы. Однако такие методы не всегда позволяют точно описать связь между входными параметрами системы и итоговым показателем. В этом случае часть взаимосвязей между факторами может оставаться неучтённой. Особенно это касается ситуаций, когда влияние факторов имеет сложный или нелинейный характер. Это ограничивает возможности оптимизации управленческих решений и затрудняет планирование рекламной активности (Breiman, 2001; Ivanov, 2019).

С точки зрения технических наук данная задача относится к классу задач

регрессионного моделирования с расширенным пространством признаков, где требуется обеспечить баланс между точностью аппроксимации и устойчивостью модели. В подобных задачах большое значение имеет выбор метода моделирования. Метод должен учитывать возможные нелинейные зависимости между параметрами системы. При этом важно, чтобы модель оставалась вычислительно эффективной. Кроме того, результаты должны быть понятны и легко интерпретироваться.

Методы машинного обучения, и, в частности, полиномиальная регрессия, предоставляют инструментарий для формализации подобных зависимостей за счёт автоматического формирования нелинейных и взаимодействующих признаков (Hastie, 2009; Vapnik, 1998). При этом модель сохраняет линейность по параметрам. Благодаря этому можно применять стандартные методы оптимизации и анализировать коэффициенты. Это свойство особенно важно для инженерных задач. В таких задачах алгоритмы должны быть понятными и прозрачными. Кроме того, результаты вычислений должны легко воспроизводиться. Таким образом, возникает необходимость разработки и экспериментальной апробации алгоритма прогнозирования объёмов продаж кирпичной продукции на основе рекламных данных с использованием методов машинного обучения, ориентированного на инженерную интерпретацию и практическую реализацию.

Целью настоящего исследования является разработка, реализация и оценка эффективности модели прогнозирования объёмов продаж кирпичной продукции предприятия ТОО «БАТСУ» города Шымкент на основе полиномиальной регрессии с использованием методов машинного обучения. В рамках исследования решаются задачи формирования расширенного пространства признаков, выбора оптимальной сложности модели и анализа её устойчивости при изменении входных параметров.

Литературный обзор. В научной литературе широко рассматривается задача прогнозирования объёмов продаж промышленной продукции на основе количественных факторов. Подобные исследования обычно связаны с анализом спроса и поддержкой управленческих решений. Во многих работах для решения этой задачи применяются линейные регрессионные модели. В таких моделях объём продаж описывается как функция ограниченного числа факторов. Чаще всего учитываются ценовые и маркетинговые показатели (Armstrong, 2017; Breiman, 2001). Основным преимуществом этих моделей является простота их использования. Кроме того, полученные коэффициенты легко интерпретируются. Благодаря этому линейная регрессия долгое время остаётся одним из наиболее распространённых инструментов анализа. Однако подобный подход имеет и определённые ограничения. На практике влияние факторов нередко оказывается нелинейным. Кроме того, между отдельными параметрами могут возникать взаимные зависимости, которые линейные модели описывают недостаточно точно. В современных исследованиях всё больше внимания уделяется влиянию рекламных воздействий на формирова-

ние спроса. Отмечается, что использование разных каналов продвижения может оказывать совместное влияние на объёмы продаж (Baecke, 2013; Smagulov, 2022).

Эмпирические наблюдения показывают, что вклад отдельных рекламных инструментов не всегда складывается линейно. Эффективность одного канала может заметно меняться, если одновременно используются другие способы продвижения.

В этой связи в литературе всё чаще используются модели с расширенным пространством признаков, включающие перекрёстные и нелинейные компоненты, позволяющие количественно оценить совместное влияние факторов (Hastie, 2009, Kuhn, 2013).

С развитием методов машинного обучения для задач прогнозирования спроса активно применяются нейронные сети, деревья решений и ансамблевые алгоритмы, демонстрирующие высокую точность на больших объёмах данных (Chen, 2016; Makridakis, 2018). Однако ряд авторов подчёркивает, что использование подобных методов на уровне отдельных промышленных предприятий осложняется ограниченностью выборки и снижением интерпретируемости моделей (Kenzhebaev, 2021; Kuznetsov, 2018). В условиях малых выборок чрезмерная сложность алгоритмов может приводить к переобучению и нестабильности прогнозов.

В рассматриваемой задаче применяется полиномиальная регрессия. Этот метод относится к моделям средней сложности. Он сочетает элементы классического регрессионного анализа и отдельных методов машинного обучения (Akhmetov, 2020; Ivanov, 2019). В основе подхода лежит расширение исходного набора признаков. Полиномиальные преобразования позволяют формировать дополнительные переменные. Благодаря этому становится возможным учитывать нелинейные зависимости между факторами. Кроме того, можно описывать их совместное влияние. При этом структура модели остаётся линейной по параметрам. Это упрощает анализ коэффициентов и делает результаты более понятными для интерпретации. Анализ научных публикаций показывает, что подобный подход успешно применяется при моделировании производственно-экономических процессов.

Это особенно важно в условиях ограниченного объёма исходных данных (Breiman, 2001; Hastie, 2009). Поэтому использование данного метода является обоснованным для решения задач, рассматриваемых в исследовании.

Материалы и методы. В рамках настоящего исследования задача анализа и прогнозирования объёмов продаж кирпичной продукции формализуется как задача регрессионного моделирования с расширенным пространством признаков (Hastie, 2009; Vapnik, 1998). Предполагается, что результирующий показатель формируется под воздействием нескольких входных факторов, влияние которых носит нелинейный характер и может сопровождаться взаимным взаимодействием. В таких условиях применение классических линейных моделей оказывается недостаточным, поскольку они не позволяют

учитывать комбинированное влияние факторов и приводят к систематическим ошибкам аппроксимации.

В качестве базового метода моделирования выбрана полиномиальная регрессия, представляющая собой линейную модель в расширенном признаковом пространстве (Hastie, 2009; Kuhn, 2013). Выбор этого метода связан с его особенностями. Он позволяет автоматически формировать нелинейные признаки и учитывать взаимодействие факторов. При этом модель сохраняет линейность по параметрам. Это делает алгоритм вычислительно устойчивым. Кроме того, коэффициенты модели остаются понятными и могут быть интерпретированы. Для оценки параметров можно применять стандартные методы оптимизации.

Следует отметить, что увеличение степени полинома приводит к росту размерности признакового пространства и, как следствие, к повышению риска переобучения модели. В связи с этим в работе используется подход, основанный на подборе оптимальной степени полинома по результатам анализа ошибок на обучающей и тестовой выборках. Такой подход помогает сохранить баланс между точностью аппроксимации и способностью модели корректно работать с новыми данными.

В работе для реализации алгоритмов машинного обучения применялась библиотека Scikit-learn. Она включает набор инструментов для обработки данных и построения моделей. С её помощью можно формировать полиномиальные признаки и строить регрессионные модели (Pedregosa, 2011).

Применение этой библиотеки позволяет корректно реализовать алгоритмы и обеспечивает воспроизводимость вычислений. Исходные данные были представлены в виде табличной выборки. В неё входили затраты на рекламу по трём каналам: телевидение, радио и интернет. В качестве целевой переменной рассматривался объём продаж кирпичной продукции в натуральном выражении. Временной фактор в модель не включался, так как основной задачей было изучение влияния рекламных затрат на объём продаж. Подготовка данных проводилась поэтапно. На первом этапе формировались матрица признаков и вектор отклика. Далее выполнялось создание полиномиальных признаков. После этого исходные данные разделялись на обучающую и тестовую выборки. На заключительном этапе проводилось обучение регрессионной модели. Затем оценивалось качество полученных прогнозов.

Оценка эффективности модели выполнялась с использованием стандартных метрик регрессии - средней абсолютной ошибки и среднеквадратического отклонения (Montgomery, 2015; Wooldridge, 2016). Оптимальная степень полинома выбиралась на основе анализа значений используемых метрик. При этом рассматривалась зависимость метрик от сложности модели. Такой анализ позволил определить момент перехода от недообучения к переобучению.

1. Характеристика данных и формирование исходных признаков

Эмпирической базой исследования послужили данные, сформированные на основе фактической информации о рекламной активности и объёмах

продаж кирпичной продукции ТОО «БАТСУ» из базы данных base.csv. Набор данных включал следующие переменные: затраты на рекламу по каналам телевидение, радио и интернет, временной параметр, характеризующий период реализации продукции, а также целевую переменную sales_bricks (pc). В выборку вошли 14 наблюдений. Такой объём данных отражает реальные условия, в которых нередко работают отдельные промышленные предприятия (Hyndman, 2021; Makridakis, 2018). Несмотря на ограниченное количество наблюдений, полученные данные представляют практический интерес. Они основаны на фактических управленческих решениях и результатах проведённых рекламных кампаний.

При построении регрессионной модели временной фактор не включался в число объясняющих переменных (Montgomery, 2015). Основная задача исследования заключалась в анализе влияния рекламных затрат. Поэтому временная динамика показателей в модели не рассматривалась. Это позволило сосредоточить анализ на зависимости между объёмом продаж и рекламными расходами. Подготовка данных проводилась поэтапно. На первом этапе исходный набор данных был разделён на матрицу признаков и вектор целевой переменной (рисунок 1).

В матрицу признаков включались все столбцы, за исключением sales_bricks (pc) и временного параметра time. В качестве целевой переменной использовался объём реализованной продукции. Эти значения применялись при построении модели. Определение целевой переменной является важным этапом подготовки данных. Оно позволяет сформировать корректную структуру информации, которая используется при дальнейшем моделировании.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = pd.read_csv("base.csv")
```

```
df.head()
```

	TV	radio	internet	sales_bricks(pc)	time
0	808076	1190848	2254106	2466996	Jan 2025
1	4791406	385515	330441	3242480	Feb 2025
2	3034860	535564	892606	2676996	Mar 2025
3	1998312	1712839	1998312	3462480	Apr 2025
4	944126	2541877	3776504	3885840	May 2025

Рисунок 1 – Сформированные входные признаки и целевая переменная

2. Формализация и анализ полиноминальных признаков

Для расширения пространства признаков использовалась библиотека Scikit-learn. В частности, применялись её инструменты для предварительной обработки данных. В качестве инструмента генерации полиномиальных признаков был выбран класс PolynomialFeatures (Hastie, 2009; Pedregosa, 2011). На рисунке 2 показано создание экземпляра данного класса,

позволяющего задать степень полинома и автоматически сформировать все соответствующие комбинации признаков. Вызов метода fit для объекта PolynomialFeatures применялся исключительно для анализа структуры входных данных. На данном этапе обучение модели ещё не проводилось. На этом этапе формировались полиномиальные признаки. В дальнейшем они применялись при построении регрессионной модели.

```
from sklearn.preprocessing import PolynomialFeatures
```

```
polynomial_converter = PolynomialFeatures(degree=2,include_bias=False)
```

```
polynomial_converter.fit(X)
```

```
PolynomialFeatures(include_bias=False)
```

Рисунок 2 – Настройка полиномиального преобразования признаков

В работе использовался полином второй степени. При таком преобразовании в модель включаются исходные признаки, их попарные произведения и квадратичные члены. Это позволяет учитывать возможные нелинейные зависимости между факторами. При этом модель остаётся устойчивой, а её коэффициенты сохраняют понятную интерпретацию (Baecke, 2013; Kuhn, 2013).

Непосредственное преобразование исходной матрицы признаков осуществлялось с помощью метода transform, в результате чего была сформирована новая матрица полиномиальных признаков. Сравнение размерности исходного и преобразованного наборов данных показало, что при неизменном числе наблюдений количество признаков увеличилось с трёх до девяти (рисунок 3).

```
polynomial_converter.transform(X)
```

```
array([[8.08076000e+05, 1.19084800e+06, 2.25410600e+06, 6.52986822e+11,
        9.62295688e+11, 1.82148896e+12, 1.41811896e+12, 2.68429762e+12,
        5.08099386e+12],
       [4.79140600e+06, 3.85515000e+05, 3.30441000e+05, 2.29575715e+13,
        1.84715888e+12, 1.58327699e+12, 1.48621815e+11, 1.27389962e+11,
        1.09191254e+11],
       [3.03486000e+06, 5.35564000e+05, 8.92606000e+05, 9.21037522e+12,
        1.62536176e+12, 2.70893425e+12, 2.86828798e+11, 4.78047640e+11,
        7.96745471e+11],
       [1.99831200e+06, 1.71283900e+06, 1.99831200e+06, 3.99325085e+12,
        3.42278673e+12, 3.99325085e+12, 2.93381744e+12, 3.42278673e+12,
        3.99325085e+12],
       [9.44126000e+05, 2.54187700e+06, 3.77650400e+06, 8.91373904e+11,
        2.39985216e+12, 3.56549562e+12, 6.46113868e+12, 9.59940866e+12,
        1.42619825e+13],
       [2.68775800e+06, 7.16735000e+05, 1.07510300e+06, 7.22404307e+12,
        1.92641023e+12, 2.88961669e+12, 5.13709060e+11, 7.70563949e+11,
        1.15584646e+12],
       [3.72971400e+06, 2.07206000e+05, 1.24323800e+06, 1.39107665e+13,
        7.72819119e+11, 4.63692217e+12, 4.29343264e+10, 2.57606373e+11,
        1.54564072e+12],
```

```
[6.89185000e+05, 3.64283600e+06, 5.51348200e+06, 4.74975964e+11,
 2.51058793e+12, 3.79980909e+12, 1.32702541e+13, 2.00847107e+13,
 3.03984838e+13],
[2.96646400e+06, 1.72054900e+06, 1.24591500e+06, 8.79990866e+12,
 5.10394667e+12, 3.69596199e+12, 2.96028886e+12, 2.14365781e+12,
 1.55230419e+12],
[4.52640200e+06, 7.44850000e+05, 4.58370000e+05, 2.04883151e+13,
 3.37149053e+12, 2.07476688e+12, 5.54801522e+11, 3.41416894e+11,
 2.10103057e+11],
[3.49279100e+06, 8.49598000e+05, 3.77599000e+05, 1.21995890e+13,
 2.96746825e+12, 1.31887439e+12, 7.21816762e+11, 3.20807355e+11,
 1.42581005e+11],
[3.58627000e+06, 4.02950000e+04, 4.02951000e+05, 1.28613325e+13,
 1.44508750e+11, 1.44509108e+12, 1.62368702e+09, 1.62369105e+10,
 1.62369508e+11],
[3.01049100e+06, 2.61782000e+05, 1.09075800e+06, 9.06305606e+12,
 7.88092355e+11, 3.28371714e+12, 6.85298155e+10, 2.85540811e+11,
 1.18975301e+12],
[4.93544700e+06, 5.60846000e+05, 1.12169000e+05, 2.43586371e+13,
 2.76802571e+12, 5.53604155e+11, 3.14548236e+11, 6.29095350e+10,
 1.25818846e+10]])
```

```
polynomial_converter.transform(X).shape
```

```
(14, 9)
```

```
X.shape
```

```
(14, 3)
```

Рисунок 3 – Расширение пространства признаков

Проведённый анализ показал, что расширенное пространство признаков содержит три исходных признака, три признака взаимодействия и три квадратичных признака (рисунок 4). Благодаря этому модель учитывает влияние отдельных рекламных каналов, а также их совместное воздействие.

```
X.iloc[0]
```

```
TV           808076
radio       1190848
internet    2254106
Name: 0, dtype: int64
```

```
poly_features[0]
```

```
array([8.08076000e+05, 1.19084800e+06, 2.25410600e+06, 6.52986822e+11,
       9.62295688e+11, 1.82148896e+12, 1.41811896e+12, 2.68429762e+12,
       5.08099386e+12])
```

Рисунок 4 – Структура расширенного пространства полиномиальных признаков

На этом этапе проводились дополнительные проверки размерности данных и просмотр отдельных строк таблицы. Эти операции выполнялись для контроля корректности подготовки данных и не выделялись как отдельные этапы методики.

3. Обучение полиномиальной регрессионной модели и оценка ее качества

После формирования полиномиальных признаков данные были разделены

на обучающую и тестовую выборки (рисунок 5). Принципиальным отличием от линейной регрессии является то, что разбиение выполнялось не по исходной матрице признаков, а по расширенному полиномиальному пространству.

```
from sklearn.model_selection import train_test_split
```

```
poly_features.shape
```

```
(14, 9)
```

Рисунок 5 – Разбиение расширенного набора данных на обучающую и тестовую выборки

В процедуре разбиения данных в качестве входных переменных использовалась матрица полиномиальных признаков, содержащая девять компонент. Целевая переменная при этом оставалась неизменной. В качестве базовой модели использовалась линейная регрессия, обучаемая на расширенном пространстве признаков и обучение модели осуществлялось на обучающей выборке, содержащей девять признаков (рисунок 6). После обучения модель тестировалась на отдельной выборке. Это позволило оценить её качество.

Такой подход сохраняет простоту оптимизации и анализа коэффициентов. При этом модель учитывает возможные нелинейные зависимости между факторами (Hastie, 2009; Wooldridge, 2016).

```
from sklearn.linear_model import LinearRegression
```

```
model = LinearRegression(fit_intercept=True)
```

```
model.fit(X_train,y_train)
```

```
LinearRegression()
```

Рисунок 6 – Обучение моделей линейной и полиномиальной регрессии

Результаты прогнозирования для тестовой выборки показаны на рисунке 7.

```
model.coef_
```

```
array([-6.75239492e+00, -5.29279572e-01, -4.63774348e+00,  8.63512296e-07,
        1.94701632e-07,  2.33173472e-06, -3.20233184e-06,  6.01009820e-06,
       -2.16211015e-06])
```

Рисунок 7 – Прогноз объёма продаж для тестовой выборки

Качество модели оценивалось с использованием стандартных метрик регрессии. В расчётах применялись средняя абсолютная ошибка и среднеквадратическое отклонение. Эти показатели измеряются в тех же единицах, что и целевая переменная. Благодаря этому результаты модели можно легко интерпретировать (рисунок 8).

```
from sklearn.metrics import mean_absolute_error,mean_squared_error
```

```
MAE = mean_absolute_error(y_test,test_predictions)
MSE = mean_squared_error(y_test,test_predictions)
RMSE = np.sqrt(MSE)
```

```
MAE
```

```
1121894.181535951
```

```
MSE
```

```
1900820552145.109
```

```
RMSE
```

```
1378702.4886265742
```

Рисунок 8 – Оценка качества модели по метрикам регрессионного анализа

Значения ошибок показывают, что полиномиальная регрессия обеспечивает более точное прогнозирование по сравнению с линейной моделью. Анализ коэффициентов показывает, что первые параметры связаны с исходными признаками. Остальные отражают влияние взаимодействия факторов и возможные нелинейные эффекты (Breiman, 2001; Smagulov, 2022).

4. Выбор оптимальной степени полинома и анализ переобучения

Для определения оптимальной сложности модели была проведена серия экспериментов с различными степенями полинома. Основной целью данного этапа являлось выявление точки, в которой дальнейшее увеличение степени приводит к росту ошибки на тестовой выборке, что свидетельствует о переобучении.

В ходе эксперимента для каждой степени полинома формировалось соответствующее пространство признаков, выполнялось разбиение данных на обучающую и тестовую выборки, обучалась модель и вычислялись значения метрик для обеих выборок.

Результаты эксперимента представлены в виде графика зависимости среднеквадратической ошибки от степени полинома для обучающей и тестовой выборок (рисунки 9, 10).

```
plt.plot(range(1,6),train_rmse_errors[:5],label='TRAIN RMSE')

plt.xlabel('Степень полинома')
plt.ylabel('RMSE')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1e05db008d0>
```

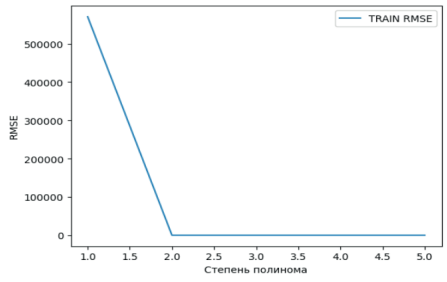


Рисунок 9 – Зависимость среднеквадратической ошибки от степени полинома (обучающая выборка)

```
plt.plot(range(1,6),train_rmse_errors[:5],label='TRAIN RMSE')
plt.plot(range(1,6),test_rmse_errors[:5],label='TEST RMSE')
plt.xlabel('Степень полинома')
plt.ylabel('RMSE')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1e05d96b5c0>
```

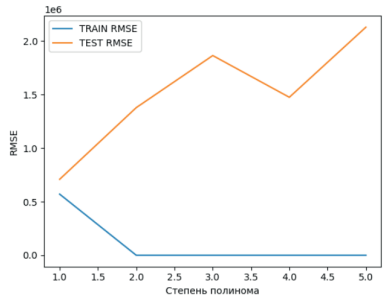


Рисунок 10 – Зависимость среднеквадратической ошибки от степени полинома (тестовая выборка)

На графиках видно, что с увеличением степени полинома ошибка на обучающей выборке уменьшается. При этом на тестовой выборке после определённого значения она начинает расти. Это свидетельствует о переобучении модели. Полученные результаты показывают, что оптимальной для данной задачи является полиномиальная модель второго или третьего порядка. Такой выбор обеспечивает баланс между точностью прогнозирования и устойчивостью модели. Использование более высоких степеней полинома в данном случае нецелесообразно. Это приводит к избыточному усложнению модели и снижает её способность корректно работать с новыми данными.

5. Финализация и внедрение модели

После выбора оптимальной степени полинома была сформирована финальная версия модели. Она предназначена для практического применения. Для её построения использовался новый объект полиномиального преобразования с выбранной степенью полинома (рисунок 11). После определения параметров модель обучалась на полном наборе данных. Разделение данных на обучающую и тестовую выборки на этом этапе не выполнялось. Далее модель применялась для прогнозирования объёма продаж кирпичной продукции при заданных значениях рекламных затрат.

Прогноз объёма продаж до марта 2026 года рассчитывался на основе бюджета по каждому рекламному каналу.

```
final_poly_converter = PolynomialFeatures(degree=3,include_bias=False)
```

```
final_model = LinearRegression()
```

```
final_model.predict(campaign_poly)
```

```
array([26441302.26292749])
```

Рисунок 11 – Модель прогноза объёма продаж кирпичной продукции

Полученные результаты позволяют оценить ожидаемый объём реализации продукции и могут быть использованы при планировании маркетинговой стратегии предприятия.

Результаты. В исследовании разработана и проверена полиномиальная регрессионная модель для прогнозирования объёмов продаж кирпичной продукции промышленного предприятия. При её построении учитывались рекламные факторы. Модель строилась в расширенном пространстве признаков. Это дало возможность учитывать нелинейные эффекты рекламных каналов и их совместное влияние.

Эмпирическая выборка состояла из 14 наблюдений. Такой объём данных отражает реальные условия, в которых предприятия нередко располагают ограниченной информацией. Несмотря на небольшое количество наблюдений, данные представляют практический интерес. Они основаны на фактических показателях рекламной активности и объёмах реализации продукции. После применения полиномиального преобразования второго порядка размерность пространства признаков увеличилась с трёх до девяти при сохранении неизменного числа наблюдений. В расширенное пространство вошли исходные рекламные факторы, квадратичные компоненты, а также попарные члены взаимодействия, что позволило учесть как индивидуальное, так и совместное воздействие рекламных каналов на объём продаж. Полученная структура признакового пространства обеспечила более гибкое описание зависимости между входными параметрами и результирующим показателем по сравнению с классической линейной моделью.

Модель обучалась с разделением исходных данных на обучающую и тестовую выборки. Для оценки эффективности предложенного подхода выполнялось сравнение линейной и полиномиальной регрессионных моделей. Сравнение проводилось по показателям качества прогнозирования. В качестве критериев качества использовались средняя абсолютная ошибка (MAE) и среднеквадратическое отклонение (RMSE), что соответствует общепринятой практике регрессионного анализа. Оценка модели показала, что полиномиальная регрессионная модель обеспечивает более низкие ошибки прогнозирования по сравнению с линейной регрессией. На тестовой выборке значения метрик составили: MAE ≈ 1121894 и RMSE ≈ 1378702. Полученные показатели указывают на достаточно точное описание зависимости между рекламными затратами и объёмом продаж. Снижение ошибки подтверждает, что учёт нелинейных эффектов и взаимодействия факторов повышает точность прогнозирования.

Анализ коэффициентов полиномиальной модели показал, что вклад членов взаимодействия рекламных каналов в ряде случаев сопоставим с вкладом исходных признаков (Armstrong, 2017; Makridakis, 2018). Это свидетельствует о синергетическом эффекте рекламных воздействий. При совместном использовании нескольких каналов их влияние на объём продаж оказывается больше, чем сумма отдельных эффектов (Baecke, 2013;

Smagulov, 2022). Данный результат экспериментально подтверждает гипотезу о нелинейном характере влияния рекламных факторов, сформулированную на этапе постановки задачи. Для выбора подходящей сложности модели проводились вычислительные эксперименты. Рассматривались различные степени полинома. Анализ показал, что с увеличением степени полинома ошибка на обучающей выборке уменьшается. При этом на тестовой выборке после определённого значения она начинает расти.

Это свидетельствует о переобучении модели при чрезмерном усложнении признакового пространства. Анализ показал, что для данной задачи оптимальной является степень полинома второго или третьего порядка. Такой выбор обеспечивает баланс между точностью аппроксимации и устойчивостью модели. На завершающем этапе исследования модель (рисунок 11) обучалась на всём доступном наборе данных. Далее она использовалась для прогнозирования объёма продаж кирпичной продукции при заданных значениях рекламных затрат.

Результаты расчёта прогнозных значений объёма реализации продукции для планируемого периода до марта 2026 года представлены в таблице 1.

Таблица 1. Прогнозные значения продажи кирпичной продукции по март 2026 года (ТОО «БАТСУ»)

№ п/п	Статьи расходов и продаж продукции	ед. изм.	Сентябрь, 2025	Октябрь, 2025	Ноябрь, 2025	Декабрь, 2025	Январь, 2026	Февраль, 2026	Март, 2026
1	Расходы отдела продаж на рекламу	тг.	5 932 927	5 729 623	4 719 988	4 029 517	4 363 030	5 608 463	26 441 302
2	Кирпич (полуторка)	шт.	3 941 784	3 326 400	1 829 520	1 542 240	2 576 996	3 352 480	7 108 462

Анализ данных, приведённых в таблице 1, показывает, что прогнозируемый объём продаж демонстрирует устойчивую зависимость от величины рекламных затрат. Увеличение рекламных расходов сопровождается ростом прогнозируемого объёма реализации продукции. При этом зависимость между затратами на рекламу и объёмом продаж имеет нелинейный характер. Это подтверждает целесообразность применения полиномиальной регрессионной модели. Такой подход позволяет учитывать изменение эффективности рекламных вложений.

Полученные прогнозные оценки могут применяться при предварительной оценке результатов рекламных кампаний. Они также помогают более обоснованно распределять рекламный бюджет и принимать управленческие решения при планировании продаж. Таким образом, результаты исследования демонстрируют не только теоретическую обоснованность предложенного

подхода, но и его практическую применимость в условиях реального промышленного предприятия.

Обсуждение. Полученные результаты согласуются с выводами исследований, посвящённых применению регрессионных и машинно-обучающих моделей в задачах прогнозирования спроса и продаж промышленной продукции. В работах многих авторов показано, что линейные модели, несмотря на простоту и интерпретируемость, часто не позволяют адекватно учитывать нелинейный характер влияния факторов и их совместное воздействие (Breiman, 2001; Hastie, 2009).

Результаты настоящего исследования подтверждают данное положение применительно к задаче прогнозирования продаж кирпичной продукции. Линейная регрессионная модель, построенная на исходных признаках, может давать ограниченную точность прогнозирования. Это связано с тем, что она предполагает независимое и аддитивное влияние рекламных каналов. Добавление полиномиальных признаков позволяет учитывать взаимодействие факторов. В результате зависимость между рекламными затратами и объёмом продаж описывается точнее (Vapnik, 1998; Kuhn, 2013). По сравнению с более сложными алгоритмами машинного обучения, например нейронными сетями или методов ансамблей, полиномиальная регрессия остаётся более прозрачной для интерпретации. Это делает её удобной для применения в инженерных и производственных задачах.

Возможность сопоставления отдельных коэффициентов с конкретными рекламными факторами и их комбинациями облегчает анализ результатов и повышает доверие к модели со стороны специалистов предприятия. Результаты исследования согласуются с работами, в которых отмечается эффективность моделей с расширенным пространством признаков при небольшом объёме данных. При работе с малыми выборками использование слишком сложных нелинейных алгоритмов может приводить к переобучению. В результате устойчивость прогнозов может снижаться (Hyndman, 2021; Makridakis, 2018). В таких условиях полиномиальная регрессия второго или третьего порядка представляет собой разумный компромисс. Она позволяет сохранить баланс между сложностью модели и её способностью обобщать результаты.

К основным преимуществам проведённого исследования относятся следующие положения:

1. формализация задачи прогнозирования на основе регрессионной модели и расширенного пространства признаков;
2. выбор степени полинома с учётом анализа переобучения модели;
3. сохранение интерпретируемости модели при повышении точности прогнозирования;
4. ориентация разработанной модели на практическое применение в условиях промышленного предприятия.

Вместе с тем проведённое исследование имеет ряд ограничений. Прежде

всего модель построена на основе небольшого объёма данных. Это может ограничивать применение модели для других временных периодов. Кроме того, в модель не включались сезонные, макроэкономические и отраслевые факторы. Такие факторы также способны влиять на объёмы продаж.

В работе не проводилось сравнение с более сложными алгоритмами машинного обучения. Поэтому данное направление можно рассматривать как перспективу дальнейших исследований. Поэтому данное направление может рассматриваться как перспектива дальнейших исследований.

Заключение. В данном исследовании была решена задача прогнозирования объёмов продаж кирпичной продукции с учётом рекламных факторов. Для этого нами применялись методы машинного обучения. Разработана и экспериментально апробирована полиномиальная регрессионная модель, позволяющая учитывать нелинейный характер влияния рекламных каналов и их взаимодействие. Показано, что расширение пространства признаков обеспечивает повышение точности прогнозирования по сравнению с классической линейной регрессией при сохранении устойчивости модели и интерпретируемости полученных параметров. Анализ показал, что рекламные каналы оказывают совместное влияние на объём продаж. Количественно этот эффект учитывается за счёт включения в модель членов взаимодействия.

Изучение зависимости ошибок прогнозирования от степени полинома позволило определить подходящую сложность модели. В результате был найден вариант, обеспечивающий баланс между точностью аппроксимации и риском переобучения при ограниченном объёме исходных данных. Разработанная модель может использоваться при планировании рекламных кампаний и прогнозировании объёмов реализации продукции на кирпичных предприятиях. Она также может быть включена в информационно-аналитические системы предприятия. Это позволяет применять её при принятии решений, связанных с распределением рекламного бюджета.

В дальнейшем представляется целесообразным расширить исходную выборку. Это можно сделать за счёт включения данных за более длительные временные периоды и дополнительных производственно-экономических факторов. Такой шаг позволит повысить обобщающую способность модели и более полно оценить её устойчивость. Отдельным направлением дальнейших исследований может стать адаптация предложенного подхода для многопараметрических моделей. Такие модели способны учитывать сезонные колебания спроса и региональные особенности строительного рынка. Кроме того, представляет интерес сравнение полученных результатов с ансамблевыми и гибридными методами машинного обучения. Полученные результаты подтверждают эффективность применения полиномиальной регрессионной модели для анализа влияния рекламных факторов на объёмы продаж кирпичной продукции.

References

Armstrong J.S., Green K.C. (2017) Demand forecasting evidence-based methods. International Journal of Forecasting, — № 33. — P. 912–930. DOI: https://doi.org/10.1016/j.ijforecast.2017.03.010 (in English)

Baecke P., Van den Poel D. (2013) The value of advertising in demand forecasting. International Journal of Forecasting. — № 29. — P. 593–606. DOI: https://doi.org/10.1016/j.ijforecast.2012.11.002 (in English)

Breiman L. (2001) Statistical modeling: The two cultures. Statistical Science. — №16. — P.199–231. DOI: https://doi.org/10.1214/ss/1009213726 (in English)

Chen T., Guestrin C. (2016) XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference, San Francisco. — P. 785–794. DOI: https://doi.org/10.1145/2939672.2939785. (in English)

Hastie T., Tibshirani R., Friedman J. (2009) The elements of statistical learning. Springer, New York —№4. — P.22–28 (in English)

Montgomery D.C., Jennings C.L., Kulahci M. (2015) Introduction to time series analysis and forecasting. Wiley, Hoboken. —№41. — P.32–38 (in English)

Pedregosa F., Varoquaux G., Gramfort A. et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, —№12. —P. 2825–2830 (in English)

Vapnik V.N. (1998) Statistical learning theory. Wiley, New York. —№2. —P. 25–36 (in English)

Wooldridge J.M. (2016) Introductory econometrics: A modern approach. Cengage Learning, Boston. — №3. — P. 825–830. (in English)

Box G.E.P., Jenkins G.M., Reinsel G.C., Ljung G.M. (2015) Time series analysis: Forecasting and control. Wiley, Hoboken. — №7. — P. 163–174. (in English)

Hyndman R.J., Athanasopoulos G. (2021) Forecasting: Principles and practice. OTexts, Melbourne. — №141. — P. 3625–2636. (in English)

Kuhn M., Johnson K. (2013) Applied predictive modeling. Springer, New York. — №12. — P. 2825–2830. (in English)

Ahmetov B.S., Suleimenov M.K. (2020) Analiz i prognozirovanie sprosa na stroitelnye materialy. Vestnik KazNİTU. — № 5. — P. 45–52. (in English)

Ivanov V.A., Petrov A.N. (2019) Matematicheskoe modelirovanie ekonomicheskih prosesov [Mathematical modeling of economic processes]. Economics and mathematical methods. — Vol 55. — № 3. — P. 98–109. (in Russian)

Kenjebaev S.T., Abdrahmanov E.N. (2021) Primenenie regresionnyh modelei v upravlenii promyşlennym predpriatiem. İzvestia NAN RK. Seria geologii i tehnicheskih nauk [The use of regression models in the management of an industrial enterpris]. News of the National Academy of Sciences of the Republic of Kazakhstan. Series of Geology and Technical Sciences. — № 2. — P.112–120. (in Russian)

Kuznesov A.V. (2018) Metody prognozirovania v proizvodstvennyh sistemah [Forecasting methods in production systems.]. Moscow: Mashinostroenie. — P.256. (in Russian)

Smagulov N.K., Tuleubaev J.S. (2022) Analiz reklamnyh faktorov v upravlenii prodajami [Analysis of advertising factors in sales management.]. Economics and statistics. — № 4. — P. 61–69. (in Russian)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 412–429

https://doi.org/10.32014/2026.2518-1726.420

IRSTI 14.35.07
UDC 378.147

© Tokhayeva A.O.[1], Alzhanov A.K.[1], Nezih Önal[2], Ziyatbekova G.Z.[3,4,5*], Begalieva K.B., 2026.
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Niğde Ömer Halisdemir University, Turkey;
[3]Almaty Technological University, Almaty, Kazakhstan;
[4]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[5]Institute of Information and Computational Technologies SC MSHE RK, Almaty, Kazakhstan;
[6 S.] Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan.
E-mail: ziyatbekova1@gmail.com

FORMATION OF STUDENTS VIRTUALIZATION COMPETENCIES IN HIGHER EDUCATION BASED ON PROXMOX VE

Tokhayeva Ainur — Doctoral student of the specialty «8D01511-Computer Science» L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: tokhaeva81@mail.ru http://orcid.org/0009-0005-1390-5217;
Alzhanov Aitugan — Candidate of Pedagogical Sciences, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: alzhanov_ak@mail.ru, https://orcid.org/0000-0003-3641-5702;
Nezih Önal — Associate Professor, PhD, Niğde Ömer Halisdemir University, Nigde, Turkey,
E-mail: nezihonal@ohu.edu.tr, https://orcid.org/0000-0002-1103-8771;
Ziyatbekova Gulzat — PhD, Associate Professor, Almaty Technological University, Almaty, Republic of Kazakhstan; Al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan; Institute of Information and Computational Technologies SC MSHE RK, Almaty, Republic of Kazakhstan
* E-mail: ziyatbekova1@gmail.com, https://orcid.org/0000-0002-9290-6074;
Begaliyeva Kalamkas — Phd, S.Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan.
E-mail: kalamkas_b@mail.ru, https://orcid.org/0000-0002-4216-9184.

Abstract. This article provides an assessment of the effectiveness of virtualization technologies using modern software tools in the information infrastructure of higher educational institutions. Proxmox VE virtualization server the issue of developing the virtualization skills of undergraduates by introducing them into the educational process of higher education institutions is considered. In the course of the study, the role of virtualization technologies in the field of

education was analyzed and the service capabilities of Proxmox VE server were studied. Educational and methodological materials and practical work for students were developed, their practical skills were formed and their effectiveness was evaluated. In the process of working in a virtualized environment, undergraduates acquire the skills of System Administration, Server Infrastructure Design and optimal use of resources. The results of the study showed that the use of the Proxmox VE platform increases the competence of graduate students in working with virtual machines, managing server resources and performing network settings. Proves the importance of using modern virtualization platforms in teaching disciplines of higher education institutions in the field of Information Systems. Virtualization technologies are recommended for use in improving the learning process and a step to increase competition in the labor market. The use of virtualization technologies has shown a positive effect on the motivation and professional interest of undergraduates in learning. Laboratory work organized on the basis of Proxmox VE allows you to reduce material technical costs, centrally manage server resources and ensure the scalability of training laboratories. In general, the study shows that it is possible to increase the professional training of undergraduates through the use of virtualization technologies, and this approach is of great practical importance in the higher education system. The results obtained can be used as a methodological basis for improving educational programs in the direction of Information Systems and Information Technology.

Keywords: virtualization, Proxmox VE platform, higher education institution, information infrastructure, efficiency, educational program

For citations: Tokhayeva A.O., Alzhanov A.K., Nezih Önal, Ziyatbekova G.Z., Begalieva K.B. Formation of students virtualization competencies in higher education based on proxmox VE. Academic https://doi.org/10.32014/2026.2518-1726.420 Scientific Journal of Computer Science, 2026. — No.1. – P. 412–429. DOI: https://doi.org/10.32014/2026.2518-1726.420

Тохаева А.О. [1], Альжанов А.К.[1,] Nezih Ö. [2], Зиятбекова Г.З. [3,4,5*], Бегалиева К.Б. [6], 2026.

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Нигде Өмер Халисдемир Университеті, Нигде, Түркия;
[3]Алматы технологиялық университеті, Алматы, Қазақстан;
[4]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[5]ҒК ҚР ҒЖБМ Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;
[6]С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан.
E-mail: ziyatbekova1@gmail.com

PROXMOX VE НЕГІЗІНДЕ ЖОҒАРЫ ОҚУ ОРЫНДАРЫНДА БІЛІМ АЛУШЫЛАРДЫ ВИРТУАЛДАНДЫРУ ҚҰЗЫРЕТТЕРІН ҚАЛЫПТАСТЫРУ

Тохаева Айнур — докторант, Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: tokhaeva81@mail.ru, https://orcid.org/0009-0005-1390-5217;
Альжанов Айтуган — педагогика ғылымдарының кандидаты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: alzhanov_ak@mail.ru, https://orcid.org/0000-0003-3641-5702;
Nezih Önal — қауымдастырылған профессор, PhD, Niğde Ömer Halisdemir University, Нигде, Түркия,
E-mail: nezihonal@ohu.edu.tr, https://orcid.org/0000-0002-1103-8771;
Зиятбекова Гулзат — PhD, қауымдастырылған профессор Алматы технологиялық университеті, Алматы, Қазақстан; Әл-Фараби атындағы Қазақ ұлттық университеті; Алматы, Қазақстан; ҒК ҚР ҒЖБМ Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: ziyatbekova1@gmail.com, https://orcid.org/0000-0002-9290-6074;
Бегалиева Каламкас — Phd, компьютерлік ғылымдар кафедрасы, С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан,
E-mail: kalamkas_b@mail.ru, https://orcid.org/0000-0002-4216-9184.

Аннотация. Бұл мақалада жоғары оқу орындарының ақпараттық-инфрақұрылымында заманауи бағдарламалық құралдарды пайдалана отырып виртуалдандыру технологияларының тиімділігін бағалауды жүзеге асыру, Proxmox VE виртуалдандыру сервері жоғары оқу орындарының оқу үдерісіне енгізу арқылы магистранттардың виртуалдандыру біліктілігін дамыту мәселесі қарастырылады. Зерттеу барысында виртуалдандыру технологияларының білім беру саласындағы рөлі талданып, Proxmox VE сервердің қызметтік мүмкіндіктері зерттелді. Білім алушыларға арналған оқу-әдістемелік материалдар мен тәжірибелік жұмыстар әзірленіп, олардың практикалық дағдылары қалыптастырып тиімділігі бағаланды. Виртуалдандырылған ортада жұмыс істеу барысында магистранттар жүйелік әкімшілендіру, серверлік инфрақұрылымды жобалау және ресурстарды оңтайлы пайдалану дағдыларын меңгереді. Зерттеу нәтижелері Proxmox

VE платформасын қолдану магистранттардың виртуалды машиналармен жұмыс істеу, серверлік ресурстарды басқару және желілік баптауларды орындау құзыреттілігін арттыратынын көрсетті. Жоғары оқу орындарының ақпараттық жүйелер саласындағы пәндерін оқытуда заманауи виртуалдандыру платформаларын қолданудың маңыздылығын дәлелдейді. Виртуалдандыру технологияларын оқыту үдерісін жетілдіруде пайдалануға ұсынылады және еңбек нарығындағы бәсекелестігін арттыру үшін қадам. Виртуалдандыру технологияларын қолдану магистранттардың оқу мотивациясына және кәсіби қызығушылығына оң әсер ететінін көрсетті. Proxmox VE негізінде ұйымдастырылған зертханалық жұмыстар материалдық техникалық шығындарды азайтуға, серверлік ресурстарды орталықтандырылған түрде басқаруға және оқу зертханаларының масштабталуын қамтамасыз етуге мүмкіндік береді. Жалпы алғанда, жүргізілген зерттеу виртуалдандыру технологияларын қолдану арқылы магистранттардың кәсіби даярлығын арттыруға болатынын және бұл тәсілдің жоғары білім беру жүйесінде практикалық маңызы жоғары екенің көрсетеді. Алынған нәтижелер ақпараттық жүйелер және ақпараттық технологиялар бағытындағы білім беру бағдарламаларын жетілдіруде әдістемелік негіз ретінде пайдаланылуы мүмкін.

Түйін сөздер: виртуалдандыру, Proxmox VE платформа, жоғары оқу орны, ақпараттық инфрақұрылым, тиімділік, білім беру бағдарламасы

Тохаева А.О.[1], Альжанов А.К. [1,] Nezih Ö. [2], Зиятбекова Г.З.[3,4,5*], Бегалиева К.Б. [6], 2026.

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

[2]Университет Нигде Омера Халисдемира, Нигде, Турция;

[3]Алматинский технологический университет, Алматы, Казахстан;

[4]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;

[5]Институт информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан;

[6]Казахский агротехнический исследовательский университет им. С.Сейфуллина, Астана, Казахстан.

E-mail: ziyatbekova1@gmail.com

ФОРМИРОВАНИЕ КОМПЕТЕНЦИЙ ВИРТУАЛИЗАЦИИ ОБУЧАЮЩИХСЯ В ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЯХ НА БАЗЕ PROXMOX VE

Тохаева Айнур — докторант, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,

E-mail: tokhaeva81@mail.ru, https://orcid.org/0009-0005-1390-5217;

Альжанов Айтуган — кандидат педагогических наук,, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: alzhanov_ak@mail.ru, https://orcid.org/0000-0003-3641-5702;
Nezih Önal — ассоциированный профессор, PhD, Университет Нигде Омера Халисдемира, Нигде, Турция,
E-mail: nezihonal@ohu.edu.tr, https://orcid.org/0000-0002-1103-8771;
Зиятбекова Гулзат — PhD, ассоциированный профессор, Алматинский технологический университет, Алматы, Казахстан; Казахский национальный университет имени аль-Фараби, Алматы, Казахстан; Институт информационных и вычислительных технологий КН МНВО РК; Алматы, Казахстан,
* E-mail: ziyatbekova1@gmail.com, https://orcid.org/0000-0002-9290-6074;
Бегалиева Каламкас — Phd, Казахский агротехнический исследовательский университет имени С.Сейфуллина, Астана, Казахстан,
E-mail: kalamkas_b@mail.ru, https://orcid.org/0000-0002-4216-9184.

Аннотация. В статье представлена оценка эффективности технологий виртуализации с использованием современных программных средств в информационной инфраструктуре высших учебных заведений. Особое внимание уделено применению сервера виртуализации Proxmox VE как инструмента формирования практических компетенций обучающихся, в частности магистрантов, посредством его интеграции в образовательный процесс. В ходе исследования проанализирована роль технологий виртуализации в сфере образования, а также функциональные возможности платформы Proxmox VE. Разработаны учебно-методические материалы и практические задания, направленные на формирование у обучающихся навыков работы в виртуализированной среде, включая системное администрирование, проектирование серверной инфраструктуры и эффективное использование вычислительных ресурсов. Результаты исследования показали, что использование платформы Proxmox VE способствует повышению уровня профессиональных компетенций обучающихся в области управления виртуальными машинами, настройки сетевой инфраструктуры и администрирования серверных систем. Установлено, что внедрение технологий виртуализации положительно влияет на учебную мотивацию и профессиональный интерес обучающихся. Кроме того, использование виртуализированных лабораторий позволяет снизить материально-технические затраты, обеспечить централизованное управление ресурсами и повысить масштабируемость учебной инфраструктуры. В целом результаты исследования подтверждают эффективность применения технологий виртуализации для повышения качества подготовки специалистов в области информационных систем и информационных технологий.

Ключевые слова: виртуализация, Proxmox VE платформа, высшее учебное заведение, Информационная инфраструктура, эффективность, образовательная программа

Кіріспе. Уақыттың талабына сай ақпараттық технологияның білім беру жүйесінде жылдам қарқынды дамып келеді, оқу үрдісінде практикалық

дағдыларды қалыптастыру басты мақсаттардың бірі болып отыр. Ғылыми еңбектерде виртуалдандыру технологияларының серверлік инфрақұрылымды басқарудағы тиімділігі, ресурстарды үнемдеу және жүйелердің икемділігін арттырудағы рөлі кеңінен талданған. Сонымен қатар, виртуалдандыру білім беру саласында зертханалық және тәжірибелік жұмыстарды ұйымдастыруды жеңілдетіп, білім алушылардың өндірістік ортаға жақын тәжірибелік дағдыларын қалыптастыруға мүмкіндік беретіні көрсетілген. Әдеби дереккөздерде виртуалдандыру платформаларының (VMware, Hyper-V, Proxmox VE және т.б.) мүмкіндіктері, олардың артықшылықтары мен шектеулері салыстырмалы түрде қарастырылған. Бұл зерттеулер виртуалдандыру технологияларын жоғары оқу орындарында қолдану тиімділігін теориялық және практикалық тұрғыдан негіздеді және білім беру үдерісінде жаңа платформа енгізудің қажетті екенін дәлелдеді.

Қазіргі уақытта жоғары оқу орындарында ақпараттық технологиялар саласының қарқынды дамуы нәтижесінде білім алушылардың практикалық дағдылардын қалыптастыру өзекті мәселе болып отыр. Әсіресе, виртуалдандыру технологиялары серверлік инфрақұрылымды тиімді басқару, ресурстарды оңтайландыру және ақпараттық жүйелердің икемділігін арттыруда маңызды рөл атқарады. Осыған байланысты магистранттардың виртуалдандыру платформаларын қолдану дағдыларын қалыптастыру жоғары білім беру жүйесінің өзекті міндеті болып табылады.

Виртуалдандыру технологияларын енгізу зертханалық жұмыстарды виртуалды ортада ұйымдастыруға мүмкіндік беріп, магистранттардың өндірістік тәжірибеге жақын практикалық қабілеттерін дамытуға ықпал етеді. Ашық кодты Proxmox VE платформасы виртуалды машиналар мен контейнерлерді басқаруға арналған кең функционалдық мүмкіндіктерге ие. KVM және LXC технологияларына негізделген бұл платформа жоғары оқу орындарының ақпараттық инфрақұрылымында тиімді қолдануға қолайлы.

Proxmox Virtual Environment — ашық бастапқы кодқа негізделген виртуалдандыру платформасы болып табылады. Бұл платформа Debian Linux операциялық жүйесіне негізделген және KVM гипервизоры мен LXC контейнерлік технологиясын біріктіреді. Proxmox VE виртуалды инфрақұрылымды басқаруға арналған веб-интерфейс, кластерлік басқару жүйесі, жоғары қолжетімділік (High Availability) механизмдері және резервтік көшіру құралдарын қамтиды.

KVM (Kernel-based Virtual Machine) технологиясы толық виртуализацияны қамтамасыз етеді және әрбір виртуалды машинаға жеке операциялық жүйені орнатуға мүмкіндік береді. Бұл технология аппараттық деңгейдегі виртуализацияны пайдалана отырып, жоғары өнімділікті қамтамасыз етеді. Ал LXC (Linux Containers) контейнерлік технологиясы жеңіл виртуализация түрі болып табылады және жүйелік ресурстарды тиімді пайдалануға мүмкіндік береді.

Proxmox VE платформасының тағы бір маңызды артықшылығы —

кластерлік архитектураны қолдауы. Бірнеше физикалық серверлерді біріктіріп, бір виртуалды инфрақұрылым ретінде басқаруға болады. Бұл жоғары оқу орындарының IT-инфрақұрылымында серверлік ресурстарды тиімді пайдалануға мүмкіндік береді.

Сонымен қатар Proxmox VE платформасы виртуалды желілерді құру, сақтауды басқару, виртуалды машиналарды көшіру және автоматты резервтік көшіру сияқты функцияларды қолдайды. Бұл мүмкіндіктер оқу процесінде студенттерге нақты өндірістік ортаға жақын тәжірибелік тапсырмалар орындауға мүмкіндік береді.

Бұл мақалада Proxmox VE платформасын оқу үдерісіне енгізу арқылы магистранттардың виртуалдандыру саласындағы кәсіби даярлығын арттыру жолдары қарастырылады.

Зерттеу барысында магистранттарға арналған оқу-әдістемелік материалдар, зертханалық жұмыстар мен практикалық тапсырмалар әзірленіп, олардың білім деңгейіне ықпалы талданады. Сонымен қатар, виртуалдандыруды оқытуда Proxmox VE платформасын пайдаланудың артықшылықтары мен тиімділігі бағаланады. Алынған зерттеу нәтижелері жоғары оқу орындарында виртуалдандыру технологияларын оқытуды жетілдіруге және білім алушылардың кәсіби даярлығын арттыруға бағытталған әдістемелік ұсыныстарды қалыптастыруға мүмкіндік береді.

Зерттеу мақсаты-Proxmox VE платформасын оқу үдерісіне интеграциялау арқылы жоғары оқу орындарында білім алушылардың виртуалдандыру құзыркттілігін дамытудың тиімділігін бағалау.

Қойылған мақсата жету үшін келесі міндеттер шешу қажет:

– виртуалдандыру технологиялары мен оларды білім беру үдерісінде қолдану мүмкіндіктеріне теориялық талдау жүргізу;

– Proxmox VE платформасының функционалдық мүмкіндіктерін зерттеу;

– магистранттарға арналған оқу-әдістемелік материалдар мен зертханалық жұмыстарды әзірлеу;

– Proxmox VE платформасын пайдалана отырып практикалық сабақтарды ұйымдастыру;

– магистранттардың виртуалдандыру құзыреттілігінің қалыптасу деңгейін бағалау.

Ақпараттық технологиялардың қарқынды дамуы білім беру жүйесіне жаңа талаптар қойып отыр. Қазіргі заманғы IT-мамандардан тек теориялық білім ғана емес, сонымен қатар нақты өндірістік ортада қолданылатын практикалық дағдылар талап етіледі. Осы тұрғыдан алғанда виртуалдандыру технологиялары білім беру саласында маңызды орын алады. Виртуалдандыру физикалық ресурстарды тиімді пайдаланып, бір серверде бірнеше операциялық жүйелерді және түрлі қызметтерді іске қосуға мүмкіндік береді. Бұл өз кезегінде білім алушыларға күрделі инфрақұрылымды қауіпсіз және икемді ортада тәжірибе жүзінде зерттеуге жағдай жасайды.

Жоғары оқу орындарында виртуалдандыру технологияларын

пайдалану зертханалық жұмыстарды ұйымдастырудың жаңа тәсілдерін қалыптастырады. Дәстүрлі физикалық зертханаларда әрбір студентке жеке сервер немесе компьютерлік инфрақұрылым бөлу көп жағдайда үлкен қаржылық шығындарды талап етеді. Ал виртуалдандыру платформалары бұл мәселені тиімді шешуге мүмкіндік береді. Бір физикалық сервер негізінде бірнеше виртуалды машиналар құрып, студенттерге жеке оқу ортасын ұсынуға болады. Бұл тәсіл білім беру ресурстарын үнемдеп қана қоймай, оқу процесінің икемділігін арттырады.

Сонымен қатар виртуалдандыру технологиялары қашықтан оқыту форматында да маңызды рөл атқарады. Студенттер интернет арқылы виртуалды зертханаларға қол жеткізіп, өз тапсырмаларын кез келген уақытта орындай алады. Бұл әсіресе қазіргі цифрлық білім беру ортасында маңызды болып табылады.

Әдеби шолу. Қазіргі уақыттағы білім мен ғылымның тез дамуына байланысты жоғарғы оқу орындарындағы зерттеулердің бағытын талап ететін виртуализацияға шартты түрде бірнеше бағытқа бөлуге болады: айтар болсақ инфрақұрлымдық тиімділік және сервистік виртуалдандыру, оларға бұлттық зертханалар виртуалды оқу процесінде заманауи бағдарламалық платформалар негізінде академикалық бұлттарды пайдалану, білім беру нәтижелерінде виртуализацияның әсерін бағалау.

Сервестік виртуализация енгізу саласындағы негізгі зерттеу университеттерінің бірі - Korthagere, Chandrashekarappa және Umesh жұмысы (Korthagere et al., 2023). Бұл мақалада RV Engineering College (Үндістан). Зерттеулер бойынша авторлар виртуалдандырылған инфрақұрылымның тиімділігін қуат тұтыну, техникалық қызмет көрсету шығындары, ауқымдылық және қауіпсіздік, ақпараттық ресустарды пайдалану деңгейі бойынша талдайды.

IT-инфрақұрылымның тұрақтылығын арттыруды көрсету мақсатында алынған нәтижелер энергия шығындары мен пайдалану шығындарының айтарлықтай төмендеуін көрсетеді.

Тиімділіктің негізгі көрсеткіштері ретінде энергия тұтынудың төмендеуі, физикалық серверлер санының азаюы және пайдалану шығындардың төмендеуін талдай отырып білім беру мекемелеріндегі виртулизацияның экономикалық және экологиялық аспектілері Ghorpade және Kamatchi жұмысында қарастырылады (Ghorpade et al., 2016). Авторлар cost-effective және green computing принциптеріне бағытталған серверлік виртуализация үлгісін ұсынады. Маңызды қорытынды-виртуализация гибридті және онлайн оқыту нысандарын қолдау үшін техникалық негіз құруға мүмкіндік берді, бұл оның оқу процесіне оң әсерін жанама түрде растайды.

Оқу процесінің контекстінде виртуалдандырылған серверлік инфрақұрылым студенттер мен оқытушылар үшін цифрлық ресурстардың қолжетімділігін арттыратын оқытуды басқару жүйелері (LMS), файлдық серверлер және оқу платформалары сияқты білім беру қызметтерінің тұрақты жұмысын қамтамасыз ететінін атап көрсетеді.

Виртуализацияның оқу процесі мен білім беру нәтижелеріне тікелей әсері Linux операциялық жүйелерін оқытуға арналған бұлттық зертхананың тиімділігін бағалауға арналған Oleksiuk, Spirin, Holovnia және Glazunova жұмысында талданады (Oleksiuk et al., 2024). Зерттеу сандық және сапалық көрсеткіштерді, соның ішінде студенттердің тестілеу нәтижелерін, курсқа қанағаттану деңгейін, зертханалық ортаның ыңғайлылығын және бұлттық ресурстарды пайдалану көрсеткіштерін пайдаланады. Нәтижелер виртуалдандырылған бұлттық зертхананың дәстүрлі физикалық зертханалардан кем түспейтінін және кейбір жағдайларда оқу ортасының икемділігі мен қолжетімділігі арқылы жоғары тиімділікті көрсететінін көрсетеді. Компьютерлік желілер бойынша виртуалдандырылған зертханаларды іс жүзінде іске асыру implementing Cloud-Based Virtual Computer Network Labs in Education жұмысында қарастырылады.

Авторлар виртуалды желілік зертхананы құру тәжірибесін сипаттайды, онда серверлік виртуализация арқылы бір физикалық серверде бірнеше тәуелсіз оқу орталары орналастырылады. Тиімділік көрсеткіштері ретінде серверге виртуалды машиналардың саны, бір студентке ресурстарды бөлу және бірнеше оқу топтары үшін зертханалық сабақтарды ұйымдастырудың икемділігі қолданылады. Бұл тәсіл оқу процесінің ауқымдылығын едәуір арттыруға және физикалық инфрақұрылымға жүктемені азайтуға мүмкіндік береді. Зерттеудің жеке бағыты заманауи виртуалдандыру бағдарламалық платформалары негізінде академиялық бұлттарды құрумен байланысты. Oleksiuk V.P., Oleksiuk O.R. жұмысында Proxmox virtual Environment платформасын пайдаланып академиялық бұлт құру тәжірибесі ұсынылған (Oleksiuk et al., 2021).

Авторлар информатика мамандарын даярлауға бағытталған виртуалдандырылған инфрақұрылымның архитектурасын сипаттайды. Студенттерге жеке виртуалды машиналарды ұсынуды, зертханалық ресурстарға қашықтан қол жеткізу мүмкіндігін және бұлтты технологиялармен жұмыс істеудің практикалық дағдыларын қалыптастыруды қоса алғанда, осы ортаның оқу процесіне оң әсері байқалады. Oleksiuk және бірлескен авторлар жасаған Apache CloudStack және Proxmox ve негізіндегі академиялық бұлттарды салыстырмалы талдау осы зерттеу бағытын толықтырады. Жұмыста виртуалды машиналарды басқару критерийлері, студенттік шоттарды бөлу, аутентификация жүйелерімен Интеграция және платформалардың желілік мүмкіндіктері қарастырылады. Талдау білім беру контекстінде жүзеге асырылады, бұл зерттеу нәтижелерін жоғары оқу орындарында заманауи виртуалдандыру бағдарламалық құралын таңдауға сәйкес етеді. Виртуалды зертханалардың оқу нәтижелеріне әсерін жалпыланған бағалау шолу және мета-аналитикалық зерттеулерде берілген.

Sellberg C. және тең автор. жұмыстарында (Sellberg et al., 2024) орындалды scoping review STEM білім берудегі виртуалды зертханаларға арналған 2012-2023 жылдардағы басылымдар. Негізгі қорытынды-білім беру

нәтижелері бойынша виртуалды зертханалар дәстүрліден кем түспейді, ал ең тиімдісі аралас (blended) оқыту моделі болып табылады. Ұқсас тұжырымдар Li J. және бірлескен мета-анализде расталады. (Li et al., 2024), PLOS ONE журналында жарияланған, онда виртуалды зертханалардың студенттердің үлгерімі мен инженерлік білім беру емтихандарының нәтижелеріне оң әсері статистикалық түрде дәлелденген.

Виртуалдандырылған білім беру ортасының тиімділігін кешенді бағалаудың әдістемелік негізін жоғары білім берудегі білім беру бағдарламаларының тиімділігін бағалаудың заманауи тәсілдеріне арналған Konurbayeva Zh.T., Denisova O.K. және Nurekenova E.S. жұмыстары толықтырады (Konurbayeva et al., 2021). Виртуализация, серверлік жүйелер, Linux әкімшілігі және бұлтты есептеу бойынша негізгі әдебиеттерді қолданды. Aladyev V. және Solovyov A. еңбектерінде виртуалдандыру мен бұлттық технологиялардың теориялық негіздері қарастырылған (Aladyev et al., 2020). Wirth N. компиляторлар мен операциялық жүйелерді құру негіздерін қарастырады (Wirth et al., 2015). Kolisnichenko D. Linux әкімшілігі бойынша практикалық ұсыныстар береді (Kolisnichenko et al., 2011). Tanenbaum A.S., мен Bos H. виртуализацияға арналған заманауи операциялық жүйелер мен бөлімдерді қамтиды (Tanenbaum et al., 2015). Сонымен қатар шетелдік зерттеулерде бұлттық және виртуалдандыру технологияларының білім беру инфрақұрылымына әсері кеңінен қарастырылған. Zhang, Chen және Li зерттеулерінде жоғары оқу орындарындағы ақпараттық инфрақұрылымды дамытуда бұлттық есептеу мен виртуализация технологияларының маңызы көрсетілген (Zhang et al., 2022). Авторлардың пікірінше, виртуалдандырылған орталар білім беру ресурстарын тиімді пайдалануға мүмкіндік беріп, серверлік инфрақұрылымды икемді басқаруға жағдай жасайды.

Marinos және Briscoe жұмыстары қауымдастық бұлттары мен виртуалдандыру инфрақұрылымдарының дамуын талдай отырып, білім беру ұйымдарында виртуалды инфрақұрылымдарды енгізу IT-ресурстарды бірлесіп пайдалану тиімділігін арттыратынын көрсетеді (Marinos et al., 2019). Мұндай инфрақұрылымдар оқу зертханаларын масштабтау және студенттердің қашықтан қолжетімділігін қамтамасыз ету үшін маңызды болып табылады.

Сонымен қатар Armbrust және бірлескен авторлар бұлттық есептеу мен виртуалдандыру технологияларының қазіргі ақпараттық жүйелерді дамытудағы стратегиялық рөлін қарастырады (Armbrust et al., 2019). Авторлар виртуалдандыру технологиялары серверлік ресурстарды оңтайландыруға, жүйелердің сенімділігін арттыруға және білім беру инфрақұрылымының тиімді жұмысын қамтамасыз етуге мүмкіндік беретінін атап көрсетеді.

Авторлар академиялық нәтижелерді, білім алушылардың қанағат-танушылығын және еңбек нарығының талаптарына сәйкестігін қоса алғанда, аралас сандық және сапалық көрсеткіштерді пайдалану қажеттілігін атап көрсетеді. Бұл тәсілдерді виртуализация мен заманауи бағдарламалық

құралдардың білім беру процесінің сапасына әсерін бағалауға бейімдеуге болады.

Әдебиеттерді талдау бойынша қорытындылар. Ғылыми жарияланымдарға жүргізілген талдау виртуалдандыру мен облыстық технологиялардың жоғары оқу орындары жағдайында құжаттық техникалық, экономикалық және педагогикалық тиімділікке ие екендігін көрсетеді. Сонымен қатар, зерттеулердің көп бөлігі инфрақұрылымдық көрсеткіштерге немесе білім беру нәтижелеріне бағытталған, бағалаудың кешенді моделін ұсынбайды.

Бұл бірыңғай әдіснамалық тәсіл шеңберінде техникалық, экономикалық және педагогикалық көрсеткіштерді интеграциялауға негізделген заманауи бағдарламалық құралдарды пайдалана отырып, жоғары оқу орындарында виртуалдандырудың тиімділігін бағалауға бағытталған зерттеудің өзектілігін айқындайды.

Зерттеу материалдары мен әдістер. Бұл зерттеу болашақ IT-саладағы мамандардың виртуалдандыру құзреттілігін дамытуға бағытталған. Зерттеу барысында әдеби шолу, салыстырмалы талдау, педагогикалық эксперимент, бақылау және бағалау, сондай-ақ алынған мәліметтерді статистикалық өңдеу әдістері қолданылды.

Зерттеу барысында келесі әдістер қолданылды:

1. ғылыми-әдістемелік және техникалық әдебиеттерге талдау жүргізу;
2. виртуалдандыру платформаларының функционалдық мүмкіндіктеріне салыстырмалы талдау жүргізу;
3. педагогикалық эксперимент жүргізу (зертханалық және практикалық жұмыстарды орындау);
4. зертханалық және практикалық жұмыстарды орындау (тест тапсырмалары мен практикалық жұмыстардың нәтижелерін талдау);
5. алынған мәліметтерді статистикалық өңдеу және талдау.

Ұсынылған зерттеуге 7М06136 - «Ақпараттық жүйелер» білім беру бағдарламасы бойынша оқитын «Қ.Құлажанов атындағы Қазақ технология және бизнес университеті» АҚ (Қазақстан) екінші курс 65 магистранттармен яғни ИСНПН - 241/1/2 тобының қатысумен ұйымдастырылды. Зерттеу барысында екі эксперимент топ және бір бақылау 30 магистрант ИСНПН - 241 тобы дәстүрлі бағдарлама бойынша оқуды жалғастырды. Мұндай құрылым педагогикалық әсердің тиімділігін анықтау үшін алынған нәтижелерін салыстыруға мүмкіндік берді.

Эксперименттік топ Proxmox VE платформасымен жұмыс істеуді үйренді, ал бақылау тобы дәстүрлік оқыту әдісі боынша оқуды жалғастырды.

Бақылау тобы дәстүрлі оқу бағдарламасы бойынша оқытылды. Бұл топта оқу дәрісі (1 сағат), практикалық сабақтар (2 сағат) барысында тапсырмаларды стандартты форматта орындау арқылы жүзеге асырылды, яғни сабақ теория жүзінде өтті. Сабақтар аптасына үш сағаттан сабақ кестесімен өтілді және магистранттар стандартты оқу материалдарын қолданып, тапсырмаларды орындалды.

Эксперименттік оқу процесі (1 сағат дәріс 2 сағат практикалық сабақ) Proxmox VE платформасында жүзеге асырылды. Практика сабақ барысында магистранттар Proxmox VE ортасында орындап, виртуалды машиналар мен контейнерлерді құру, конфигурациялау және басқару бойынша дағдыларын меңгерді. Сабақтар аптасына үш сағаттан өткізіліп, магистранттар стандартты оқу материалдарын мен әдістемелік нұсқауларды пайдаланып тапсырмаларды орындады.

Зерттеу барысында магистранттардың виртуалдандыру технологияларын практикалық деңгейде меңгеруін қамтамасыз ету мақсатында бірнеше зертханалық жұмыстар кешені әзірленді. Әрбір зертханалық жұмыс виртуалдандыру инфрақұрылымының негізгі элементтерін түсінуге және Proxmox VE платформасының функционалдық мүмкіндіктерін тәжірибе жүзінде меңгеруге бағытталды.

Бірінші зертханалық жұмыс Proxmox VE платформасын орнату және бастапқы конфигурациялау кезеңдеріне арналды. Магистранттар серверлік виртуализацияның негізгі принциптерімен танысып, гипервизор орнату процесін орындады. Сонымен қатар желілік параметрлерді баптау, сақтау жүйесін конфигурациялау және веб-интерфейс арқылы басқару элементтерін қолдану дағдылары қалыптастырылды.

Екінші зертханалық жұмыс виртуалды машиналарды құру және басқару тапсырмаларын қамтыды. Магистранттар Proxmox VE ортасында жаңа виртуалды машиналарды құрып, оларға операциялық жүйелер орнатты. Бұл тапсырмалар барысында CPU, жедел жады (RAM) және дискілік кеңістік сияқты ресурстарды бөлу және конфигурациялау принциптері қарастырылды. Сонымен қатар виртуалды машиналарды іске қосу, тоқтату және қайта жүктеу әрекеттері тәжірибе жүзінде орындалды.

Үшінші зертханалық жұмыс контейнерлік виртуалдандыру технологияларына арналды. Бұл кезеңде магистранттар LXC контейнерлерін құрып, олардың параметрлерін баптады. Контейнерлерді іске қосу, тоқтату, желілік конфигурация жасау және ресурстарды шектеу сияқты әрекеттер орындалды. Бұл тапсырмалар контейнерлік технологиялардың виртуалды машиналармен салыстырғандағы ерекшеліктерін түсінуге мүмкіндік берді.

Төртінші зертханалық жұмыста виртуалды желілік инфрақұрылымды конфигурациялау қарастырылды. Магистранттар виртуалды желілік интерфейстер мен виртуалды коммутаторларды құрып, бірнеше виртуалды машиналар арасындағы желілік байланысты ұйымдастырды. Сонымен қатар виртуалды ортаның жұмысын бақылау және жүйелік мониторинг құралдарын қолдану дағдылары қалыптастырылды.

Осы зертханалық жұмыстар кешені магистранттардың виртуалдандыру технологияларын практикалық деңгейде меңгеруіне және серверлік инфрақұрылымды басқару дағдыларын дамытуға мүмкіндік берді.

Экперимент жүргізер алдында зерттеу топтарында семестрдің басында (пре -тест) және семестірдің соңында (пост -тест) бағалау жүргізілді.

Алдын ала және тестен кейінгі сынақтардан алынған сандық деректер эксперименттік және бақылау топтарының орташа көрсеткіштерін салыстыру үшін статистикалық әдістері арқылы талданды. Айырмашылықтардың маңыздылығын анықтау үшін жұптастырылған t-тесттер мен ANOVA қолданылды. Сұхбаттар мен аналитикалық журналдардың сапалы деректерді қатысушылардың қабылдауы мен тәжірибесіне байланысты жалпы тақырыптарды анықтау мақсатында тақырыптық талдау әдісі арқылы зерттеледі. Бұл талдауда Proxmox VE платформасын оқу үдерісіне енгізудің виртуалдандыру құзыреттілігін дамытуға қалай әсер еткенін тереңірек түсінуге мүмкіндік береді.

Зерттеу Қ.Құлажанов атындағы Қазақ технология және бизнес университетінің ғылыми этика комиссиясынан этикалық мақұлдау алды. Барлық қатысушыларға зерттеу мақсаты түсіндіріліп, қатысу ерікті екендігі және жауаптарының құпиялылығы туралы ақпарат берілді.

Зерттеу нәтижелері. Зерттеу нәтижесі бойынша алдын ала (пре - тест) және кейінгі тестілеу (пост - тест) квази – эксперименттік дизанды пайдалана отырып жүргізілген педагогикалық эксперименттің тимділігі туралы сандық және сапалық ақпарат ұсынылды.

Графикалық көрсеткіштер мен кестелік деректер Proxmox VE платформасында жүргізілген тәжірибелік жұмыс нәтижесінде виртуалдандыру құзыреттілігінің (виртуалды машиналар мен контейнерлерді құру, конфигурациялау және басқару дағдыларын) жақсарғанын айқын көрсетті. Алынған нәтижелер стандартталған тест тапсырмалары арқылы өлшеніп, эксперименттік топтың көрсеткіштері бақылау тобына қарағанда статистикалық жағынан жоғары екені анықталды.

Бұдан әрі екі топта педагогикалық эксперимент жүргізілгенге дейін және жүргізілгеннен кейін (1-кесте) орташа нәтижелерін бақылау тобында және эксперименттік топта Proxmox VE платформасын қолдану оң динамика көрсетті.

Кесте 1 – Эксперименттік және бақылау топтарын үшін пре - тест және пост -тест бойынша орташа балдары

Топтар	**Магистрант-тар саны**	**Сұрақ саны**	**Пре-тест орташа дұрыс жауап**	**Пре-тест (%)**	**Пре-тест орта-ша балл**	**Пост-тест (дұрыс жауап)**	**Пост-тест (%)**	**Пост-тест орташа дұрыс жауап**	**Дина-мика (балл)**	**Дина-мика (%)**
Эксперименттік топ 1 (ИСНПН-241/1)	35	15	8	53,3%	8,0	10	66,7%	10,0	+2	+13,4%
Эксперименттік топ 2 (ИСНПН-241/2)	30	15	6	40,0%	6,0	9	60,0%	9,0	+3	+20,0%

Бақылау тобы (ИСНПН-241)	30	15	7	46,7%	7,0	8	53,3%	8,0	+1	+6,6%

Топ бойынша орташа балл ($\bar{x}$) мына формуламен есептеледі:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

мұндағы: $\sum x_i$ - барлық магистранттардың алғанбалдарының қосындысы
Менің деректерім бойынша
Эксперименттік топ 1 (ИСНПН-241/1)
Магистранттар саны: n = 35
Тест сұрақтары: 15
Пре-тест орташа нәтижесі: 8 дұрыс жауап
Бұл дегеніміз:

$$\sum x_i = 35 * 8 = 280$$

Эксперименттік топта 1 (ИСНПН-241/1) үшін магистранттар саны n =35 болды. Пре-тест нәтижелері бойынша 35 магистранттың жинаған жалпы балдары 280- ге тең болды, нәтижесінде орташа балл:

$$\bar{x} = \frac{280}{35} = 8$$

Пре-тест орташа бал = 8
Пайыздар көрсеткішті есептеу формуласы:

$$\text{Пайыз} = \frac{\bar{x}}{15} * 100\%$$

$$\frac{\bar{8}}{15} * 100\% = 53{,}3\%$$

Динамиканы есептеу
Балл бойынша:

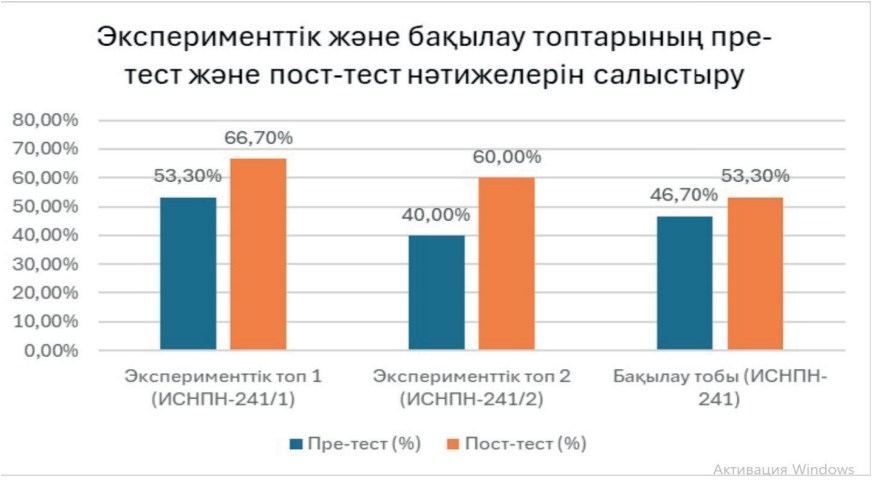
$$\Delta = \bar{x}_{post} - \bar{x}_{pre}$$
$$\Delta = 10 - 8 = +2 \quad (2)$$

Пайыз бойынша: Δ%=66,7% - 53,3%= +13,4%

Эксперименттік топта пре-тест нәтижелері бойынша магистранттардың орташа көрсеткіші 15 сұрақтың 8 дұрыс жауабын құрады (53,3%), ал пост-тест нәтижесінде бұл көрсеткіш 10 дұрыс жауапқа дейн өсті (66,7%). Орташа балдың +2 балға артуы Proxmox VE платформасын оқу үдерісіне енгізудің тиімділігін көрсетеді.

Үш топтың (екі эксперименттік және бір бақылау тобы) пре-тест және пост-тест нәтижелерін салыстыру мақсатында бір факторлы дисперсиялық талдау

(ANOVA) жүргізілді, нәтижесінде топтар арасындағы айырмашылықтың статистикалық тұрғыдан мәнді екені анықталады ($\rho < 0{,}05$).



Сурет 1 – Эксперимент және бақылау топтарындағы білім нәтижелерінің өзгерісі

Кестеде көрсетілген нәтижелер диаграмма түрінде 1-суретте көрсетілген. Диаграммада эксперименттік топтарда пре-тест нәтижесін пост-тест нәтижесімен салыстырғанда артқанын көруге болады, ал бақылау тобында көрсеткіштердің айтралықтай өзгермегенін айқын көрсетеді.

Талқылау. Зерттеу кезінде Proxmox VE платформасын оқу процесіне енгізудің магистранттардың виртуалдандыру құзыреттілігін дамытуға әсері эксперименттік жолмен бағаланды. Квази-эксперимент дизайн негізінде жүргізілген педагогикалық эксперимент натижелері эксперименттік топтарда пре-тест нәтижесін пост-тест нәтижесі арасында арту динамиканың байқалғанын көрсетті, ал бақылау тобында айтарлықтай өзгерістер анықталған жоқ. Эксперименттік топтардың нәтижелердің артуы Proxmox VE платформасында орындалған практикалық тапсырмалардың магистранттардың миртуалды машиналарды құру, конфигурациялау және басқару дағдыларын тиімді дамытуға ықпал еткенін дәлелдейді. Бұл платформа оқу үдерісінде теориялық білімді тәжірибемен ұштастыруға мүмкіндік беріп, нақты өндірістік ортаға жақын жағдай қалыптастырды.

Алынған нәтижелер виртуалдандыру технологияларын оқытуда практикалық бағыттағы заманауи платформаларды қолданудың тиімділігін растайды және оларды жоғары оқу орындарының білім беру бағдарламаларына енгізудің орынды екенін көрсетеді.

Сонымен қатар виртуалдандыру технологияларын оқу процесіне енгізу магистранттардың кәсіби құзыреттілігін дамытуда маңызды рөл атқарады. Магистратура деңгейіндегі білім беру бағдарламалары, әдетте, студенттердің терең теориялық білімін ғана емес, сонымен қатар нақты практикалық дағдыларын қалыптастыруды көздейді. Осы тұрғыдан алғанда Proxmox VE платформасы магистранттарға арналған оқу процесінде тиімді құрал ретінде қолданылуы мүмкін. Бұл платформа арқылы магистранттар серверлік виртуалдандыру инфрақұрылымын құру, виртуалды машиналар мен контейнерлерді басқару, желілік конфигурацияларды ұйымдастыру және жүйелік ресурстарды тиімді бөлу сияқты кәсіби дағдыларды тәжірибе жүзінде меңгереді.

Сонымен қатар магистранттар виртуалды ортада түрлі операциялық жүйелермен жұмыс істеу, серверлік қызметтерді орнату және басқару, желілік қызметтерді конфигурациялау сияқты күрделі тапсырмаларды орындау арқылы өздерінің кәсіби дайындық деңгейін арттыра алады. Мұндай тәжірибелік жұмыстар магистранттардың жүйелік әкімшілендіру, бұлттық технологиялар және DevOps бағытындағы білімдерін тереңдетуге мүмкіндік береді.

Proxmox VE платформасын қолдану магистранттардың зерттеу дағдыларын дамытуға да ықпал етеді. Виртуалды орталар арқылы олар әртүрлі жүйелік конфигурацияларды қауіпсіз түрде сынақтан өткізіп, ақпараттық жүйелердің жұмысын талдауға мүмкіндік алады. Бұл болашақ IT мамандарының ғылыми-зерттеу және инженерлік қабілеттерін дамытуға оң әсер етеді. Зерттеу нәтижелерінің практикалық маңыздылығы Proxmox VE платформасын жоғары оқу орындарының білім беру үдерісіне тиімді интеграциялау мүмкіндігімен түсіндіріледі. Бұл платформа виртуалдандырылған зертханалық орталарды ұйымдастыруға мүмкіндік беріп, студенттер мен магистранттарға серверлік жүйелерді басқару, виртуалды машиналарды конфигурациялау және желілік инфрақұрылымды құру бойынша нақты тәжірибе алуға жағдай жасайды.

Proxmox VE негізіндегі виртуалды зертханалар білім беру мекемелерінің материалдық ресурстарын тиімді пайдалануға мүмкіндік береді. Бір физикалық сервер негізінде бірнеше виртуалды оқу ортасын ұйымдастыру арқылы зертханалық жұмыстарды орындауға қажетті инфрақұрылымды кеңейтуге болады. Бұл тәсіл оқу орындарының техникалық шығындарын азайтып, IT-инфрақұрылымды икемді басқаруға мүмкіндік береді.

Сонымен қатар виртуалдандыру технологияларын оқу процесінде қолдану білім алушылардың кәсіби құзыреттілігін арттыруға ықпал етеді. Магистранттар виртуалды машиналар мен контейнерлерді құру, конфигурациялау және басқару сияқты практикалық тапсырмаларды

орындау арқылы нақты өндірістік ортаға жақын тәжірибе алады. Бұл олардың болашақта жүйелік әкімші, DevOps инженер немесе бұлттық инфрақұрылым маманы ретінде кәсіби қызмет атқаруына дайындық деңгейін арттырады.

Осылайша Proxmox VE платформасын оқу процесінде қолдану жоғары оқу орындарында виртуалдандыру технологияларын оқытудың тиімді әдістерінің бірі болып табылады.

Сонымен қатар Proxmox VE платформасын оқу үдерісінде пайдалану магистранттардың өздігінен білім алу және зерттеу жүргізу дағдыларын дамытуға мүмкіндік береді. Виртуалды ортаны пайдалану арқылы магистранттар әртүрлі жүйелік конфигурацияларды қауіпсіз түрде сынақтан өткізіп, ақпараттық жүйелердің жұмыс істеу принциптерін тереңірек түсіне алады. Бұл тәсіл олардың аналитикалық ойлау қабілетін дамытуға және болашақ кәсіби қызметінде кездесетін практикалық міндеттерді тиімді шешуге дайын болуына ықпал етеді.

Қорытынды. Зерттеу нәтижесінен алынған сандық деректер Proxmox VE платформасын орындалған практикалық жұмыстар магистранттардың виртуалды машиналар мен контейнерлерді құру, конфигурациялау және басқару бойынша практикалық дағдыларын тиімді дамытуға ықпал ететінін дәлелдеді. Статистикалық талдау нәтижелері эксперимент топтардың натижесі бақылау тобымен салыстырғанда нәтижесі жоғары екенін көрсетті. Бұл енгізілген педагогикалық әсердің тиімділігін көрсетеді. Сондайақ Proxmox VE платформасын жоғары оқу орындарының оқу үрдісіне енгізу магистранттардың кәсіби даярлығын, арттыруға виртуалдандыру технологиялары саласындағы практикалық қалыптастыруға және еңбек нарығындағы бәсекеге қабілетті мамандарды даярлауға мүмкіндік берді. Зерттеу нәтижесі білім беру бағдарламасында виртуалдандыру технологияларын оқыту әдістемесін жетілдіру және жаңарту барысында пайдалануға ұсынылады.

Сонымен қатар жүргізілген зерттеу нәтижелері виртуалдандыру технологияларын оқу бағдарламаларына жүйелі түрде енгізудің маңыздылығын көрсетеді. Proxmox VE платформасын қолдану магистранттардың практикалық дайындық деңгейін арттырып қана қоймай, олардың заманауи IT-инфрақұрылыммен жұмыс істеу қабілетін дамытады. Бұл жоғары оқу орындарында цифрлық білім беру ортасын тиімді ұйымдастыруға және болашақ мамандардың кәсіби бәсекеге қабілеттілігін арттыруға ықпал етеді.

References

Aladyev V.Z., & Bogatyrev S.D. (2020) *Virtualizaciya: ot osnov do oblachnyh tekhnologij* [Virtualization: From fundamentals to cloud technologies]. Moscow: Laboratory of Knowledge. (in Russian)

Armbrust M., Fox A., & Griffith R. (2019) A view of cloud computing and virtualization technologies. *Communications of the ACM*, 53(4). — P. 50–58. (in English)

Ghorpade Y.M., & Kamatchi R. (2016) A study to develop server virtualization infrastructure in educational institutions with a cost-effective and green computing approach. (in English)

Huang A. (2019) Teaching, learning and assessment with virtualization technology. *Journal of Information Technology Education: Innovations in Practice*. (in English)

Kolisnichenko D.N. (2021) *Linux. Administrirovanie* [Linux administration]. Saint Petersburg: BHV-Petersburg. (in Russian)

Konurbayeva Zh.T., Denisova O.K., & Nurekenova E.S. (2021) Modern approaches to evaluating the effectiveness of higher education programs. *Bulletin of the Karaganda University. Economy Series*, 102(2). — P. 75–83. (in English)

Korthagere S., Chandrashekarappa R. K., & Umesh, I. M. (2023) Server virtualization in higher educational institutions: A case study. *International Journal of Electrical and Computer Engineering (IJECE)*, 13(4). — P. 4477–4487. (in English)

Li J. (2024) Effectiveness of virtual laboratory in engineering education: A meta-analysis. *PLOS ONE*. (in English)

Marinos A., & Briscoe G. (2019) Community cloud computing and virtualization infrastructures. *IEEE Internet Computing*, 23(5). — P. 54–62. (in English)

Oleksiuk V.P. (2021) Comparative study of the support of academic clouds based on Apache CloudStack and Proxmox VE platforms. *Educational Technology Quarterly*. (in English)

Oleksiuk V.P., & Oleksiuk O.R. (2021) The practice of developing the academic cloud using the Proxmox VE platform. *Educational Technology Quarterly*, (4). (in English)

Oleksiuk V.P., Spirin O.M., Holovnia O.S., & Glazunova O.G. (2024) Evaluating the effectiveness of a cloud-based laboratory for teaching Linux operating systems to computer science students. (in English)

Sellberg C. (2024) Virtual laboratories in STEM higher education: A scoping review. *Education Sciences*. (in English)

Solovyov A.V. (2019) *Oblachnye vychisleniya i virtualizaciya* [Cloud computing and virtualization]. Moscow: KnoRus. (in Russian)

Tanenbaum A.S., & Bos H. (2015) *Sovremennye operacionnye sistemy* [Modern operating systems] (4th ed.). Saint Petersburg: Piter. (in Russian)

Wirth N. (2015) *Postroenie kompilyatorov i operacionnyh sistem* [Building compilers and operating systems]. Moscow. (in Russian)

Zhang Q., Chen M., & Li L. (2022) Cloud computing and virtualization technologies in higher education environments. *Future Generation Computer Systems*, 129. — P. 239–252. (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 430–442

https://doi.org/10.32014/2026.2518-1726.421

IRSTI 50.09.45
UDC 621.45.01

© Tukenova L.M.[1]*, Auyelbekov O.A.2-3, Sapakova S.Z.[4], Sametova A.A.[1], Bostanov E.L.[5], 2026.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Institute of Information and Computational Technologies;
[3]Kazakh National Women's Pedagogical University, Almaty, Kazakhstan;
[4]International Information Technology University, Almaty, Kazakhstan;
[5]Civil Aviation Academy, Almaty, Kazakhstan.
E-mail:: l.tukenova7@gmail.com

MODELLING AND OPTIMISATION OF HYBRID POWER PLANT OPERATING MODES FOR UNMANNED AERIAL VEHICLES

Tukenova Lyailya — cand. of ph. and math. sc., Associate Professor, Al-Farabi Kazakh National university, Almaty, Kazakhstan,
E-mail: l.tukenova7@gmail.com, https://orcid.org/0000-0002-0863-5153;
Auelbekov Omirlan — cand. of ph. and math. sc., Associate Professor, Institute of Information and Computational Technologies; Kazakh National Women's Pedagogical University, Almaty, Kazakhstan,
E-mail: omirlan.auelbek@gmail.com, https://orcid.org/0000-0002-2903-9086;
Sapakova Saya — cand. of ph. and math. sc., Associate Professor, International Information Technology University, Almaty, Kazakhstan,
E-mail: s.sapakova@iitu.edu.kz, https://orcid.org/0000-0001-6541-6806;
Sametova Aygerim — Senior lecturer, Al-Farabi Kazakh National university, Almaty, Kazakhstan,
E-mail: sametova_aygerim@mail.ru, https://orcid.org/0000-0003-1849-8938;
Bostanov Ernar —senior lecturer, Civil Aviation Academy, Almaty, Kazakhstan,
E-mail: bostanovernar0929@gmail.com, https://orcid.org/0000-0002-8753-0244.

Abstract. The study addresses the development and optimisation of hybrid power plants (HPPs) for unmanned aerial vehicles (UAVs) intended for long-endurance missions and improved energy efficiency. The work examines the scientific problem of selecting an effective configuration and energy management strategy (EMS) that ensures maximum flight duration while maintaining the stability, reliability, and thermal safety of the onboard energy system. A mathematical modelling framework was developed to simulate the operating modes of hybrid architectures combining an internal combustion engine, generator, battery system, and power electronics. Using MATLAB/Simulink and Modelica, the HPP was analysed through major

flight phases, from take-off to landing. The results indicate that the main factor contributing to increased UAV autonomy is an optimised EMS strategy that maintains the thermal energy source in its region of highest efficiency and allocates transient peak loads to the battery. Numerical experiments demonstrated that hybridisation reduces energy consumption during cruise flight and significantly improves dynamic response during manoeuvres. Furthermore, accurate battery modelling and adherence to SoC and temperature constraints are essential, as violations diminish the theoretical benefits of hybridisation. The originality of the work lies in establishing applied guidelines for HPP design, including efficiency-map calibration, selection of optimal SoC "windows," current limitation rules, and the implementation of adaptive EMS algorithms. The obtained results can serve as a practical tool for scientifically grounding the design of advanced UAV platforms intended for long-duration patrolling, high-precision mapping, and comprehensive environmental monitoring, supporting parameter selection for hybrid power systems, tuning of energy management algorithms, and mission planning.

Keywords: hybrid power plant; unmanned aerial vehicles; energy management system (EMS); power distribution, state of charge (SoC)

For citations: Tukenova L.M., Auyelbekov O.A., Sapakova S.Z., Sametova A.A., Bostanov E.L. Modelling and optimisation of hybrid power plant operating modes for unmanned aerial vehicles. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 430–442. DOI: https://doi.org/10.32014/2026.2518-1726.421

© Төкенова Л.М.[1*], Әуелбеков О.А.[2,3], Сапақова С.З.[4], Саметова А.А.[1], Бостанов Е.Л.[5], 2026.

[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[1]Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан;
Қазақ ұлттық қыздар педагогикалық университеті, Алматы, Қазақстан;
[3] Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[4]Азаматтық авиация академиясы, Алматы, Қазақстан.
E-mail:l.tukenova7@gmail.com

ПИЛОТСЫЗ ҰШУ АППАРАТТАРЫНА АРНАЛҒАН ГИБРИДТІ ЭЛЕКТР СТАНЦИЯЛАРЫНЫҢ ЖҰМЫС РЕЖИМДЕРІН МОДЕЛЬДЕУ ЖӘНЕ ОҢТАЙЛАНДЫРУ

Төкенова Ләйлә — ф.-м.ғ.к., қауымдастырылған профессор, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: l.tukenova7@gmail.com, https://orcid.org/0000-0002-0863-5153;
Әуелбеков Өмірлан — ф.-м.ғ.к., қауымдастырылған профессор, Ақпараттық және есептеу технологиялары институты; Қазақ ұлттық қыздар педагогикалық университеті, Алматы, Қазақстан,
E-mail: omirlan.auelbek@gmail.com, https://orcid.org/0000-0002-2903-9086;
Сапақова Сая — ф.-м.ғ.к., қауымдастырылған профессор, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,

E-mail: s.sapakova@iitu.edu.kz, https://orcid.org/0000-0001-6541-6806;
Саметова Айгерім — аға оқытушы, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: sametova_aygerim@mail.ru, https://orcid.org/0000-0003-1849-8938;
Бостанов Ернар — аға оқытушы, Азаматтық авиация академиясы, Алматы, Қазақстан, E-mail: bostanovernar0929@gmail.com, https://orcid.org/0000-0002-8753-0244.

Аннотация. Зерттеу ұзақ мерзімді ұшу тапсырмаларына және энергия тиімділігін арттыруға бағытталған ұшқышсыз ұшу аппараттарына (ҰҰА) арналған гибридті күш қондырғыларын (ГКҚ) әзірлеу және оңтайландыру мәселелеріне арналған. Жұмыста борттық энергия жүйесінің тұрақтылығын, сенімділігін және жылулық қауіпсіздігін сақтай отырып, максималды ұшу ұзақтығын қамтамасыз ететін тиімді конфигурация мен энергияны басқару стратегиясын (EMS) таңдаудың ғылыми мәселесі қарастырылады. Ішкі жану қозғалтқышы, генератор, аккумулятор жүйесі және қуат электроникасы біріктірілген гибридті архитектуралардың жұмыс режимдерін модельдеу үшін математикалық үлгілеу негізі құрылды. MATLAB/Simulink және Modelica құралдарының көмегімен ГКҚ жұмысы ұшудың әртүрлі кезеңдерінде – ұшу, марштық режим, маневр және қону бағаланды. Зерттеу нәтижелері ҰҰА автономдылығын арттырудың негізгі факторы – жылулық энергия көзі ең тиімді аймақта жұмыс істейтін және өтпелі қуат жүктемелерін аккумуляторға бөлетін оңтайлы EMS стратегиясы екенін көрсетті. Сандық эксперименттер гибридтендіру марштық ұшу кезінде энергия шығынын азайтатынын және маневрлер кезіндегі динамикалық жауапты жақсартатынын дәлелдеді. Сонымен қатар, аккумулятордың дәл моделі мен SoC және температура шектеулерін сақтау маңызды шарт болып табылады, себебі олардың бұзылуы теориялық тиімділікті төмендетеді. Жұмыстың өзіндік жаңалығы ГКҚ құрастыруға арналған қолданбалы нұсқаулықтарды қалыптастыруында: тиімділік карталарын калибрлеу, SoC «терезелерін» таңдау, ток шектеу ережелері және бейімделген EMS алгоритмдерін енгізу. Алынған нәтижелер ұзақ уақыттық патрульдеуге, жоғары дәлдіктегі картографиялауға және кешенді экологиялық мониторингке арналған перспективалы ҰҰА платформаларын жобалау барысында гибридті күш қондырғысының параметрлерін таңдауды, энергияны басқару алгоритмдерін баптауды және ұшу миссияларын жоспарлауды ғылыми негіздеуге мүмкіндік беретін практикалық құрал ретінде қолданылуы мүмкін.

Түйін сөздер: гибридті қозғалтқыш жүйесі; пилотсыз ұшу аппараттары; энергияны басқару жүйесі (ЭБЖ); қуатты бөлу; зарядтау күйі (SoC)

© Тукенова Л.М.[1]*, Ауелбеков О.А. [2,3], Сапакова С.З. [4], Саметова А.А., Бостанов Е.Л., 2026.

[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;

[2] Институт информационных и вычислительных технологий, Алматы, Казахстан;

[3]Казахский национальный женский педагогический университет, Алматы, Казахстан;

[4] Международный университет информационных технологий, Алматы, Казахстан;

[5]Академия гражданской авиации, Алматы, Казахстан.

E-mail: l.tukenova7@gmail.com

МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ РЕЖИМОВ РАБОТЫ ГИБРИДНЫХ СИЛОВЫХ УСТАНОВОК ДЛЯ БЕСПИЛОТНЫХ ЛЕТАТЕЛЬНЫХ АППАРАТОВ

Тукенова Ляйля — канд. физ.-мат. наук, доцент, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: l.tukenova7@gmail.com, https://orcid.org/0000-0002-0863-5153;
Ауельбеков Омирлан — канд. физ.-мат. наук, доцент, Институт информационных и вычислительных технологий; Казахский национальный женский педагогический университет, Алматы, Казахстан,
E-mail: omirlan.auelbek@gmail.com, https://orcid.org/0000-0002-2903-9086;
Сапакова Сая — канд. физ.-мат. наук, доцент, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: s.sapakova@iitu.edu.kz, https://orcid.org/0000-0001-6541-6806;
Саметова Айгерим — старший преподаватель, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: sametova_aygerim@mail.ru, https://orcid.org/0000-0003-1849-8938;
Бостанов Ернар — старший преподаватель, Академия гражданской авиации, Алматы, Казахстан,
E-mail: bostanovernar0929@gmail.com, https://orcid.org/0000-0002-8753-0244.

Аннотация. Исследование посвящено разработке и оптимизации гибридных силовых установок (ГСУ) для беспилотных летательных аппаратов (БПЛА), ориентированных на выполнение длительных полётных задач и повышение энергетической эффективности. В работе рассматривается научная проблема выбора эффективной конфигурации и стратегии управления энергией (Energy Management System, EMS), обеспечивающей максимальную продолжительность полёта при сохранении стабильности, надёжности и тепловой безопасности бортовой энергосистемы. Для моделирования режимов работы гибридных архитектур, включающих двигатель внутреннего сгорания, генератор, аккумуляторную систему и силовую электронику, разработана математическая модель. С использованием MATLAB/Simulink и Modelica проведена оценка поведения ГСУ на различных этапах полёта: взлёте,

крейсерском режиме, маневрировании и посадке. Результаты исследования показали, что ключевым фактором повышения автономности БПЛА является оптимальная стратегия EMS, обеспечивающая работу теплового источника в области максимальной эффективности и перераспределение пиковых нагрузок на аккумуляторную систему. Численные эксперименты подтвердили, что гибридизация способствует снижению энергопотребления в крейсерском режиме и улучшению динамического отклика при выполнении манёвров. При этом критически важным является обеспечение точности модели аккумулятора и соблюдение ограничений по состоянию заряда (SoC) и температурному режиму, поскольку их нарушение нивелирует преимущества гибридной архитектуры. Научная новизна работы заключается в разработке прикладных рекомендаций по проектированию ГСУ, включая калибровку карт эффективности, выбор рабочих диапазонов SoC, правила ограничения токов и внедрение адаптивных алгоритмов EMS. Практическая значимость результатов заключается в возможности их использования при проектировании перспективных БПЛА для задач длительного патрулирования, высокоточной картографии и экологического мониторинга, включая выбор параметров гибридной силовой установки, настройку алгоритмов управления энергией и планирование полётных миссий.

Ключевые слова: гибридная силовая установка, беспилотные летательные аппараты, энергетический менеджмент (EMS), распределение мощности, состояние заряда (SoC)

Introduction. In recent years, the development of unmanned aerial vehicles (UAVs) has become one of the priority areas in aerospace technology, the defense industry, agriculture and environmental monitoring. However, a key limiting factor remains the limited flight autonomy, which is directly related to the limitations on the mass and energy capacity of power plants (Zhang et al., 2023).

Traditional electric motors used in small and medium-sized UAVs are highly environmentally friendly and low-noise but have a relatively short operating time due to the limited capacity of rechargeable batteries (Khare et al., 2011). On the other hand, thermal engines provide longer flight times, but are associated with noise, emissions and the need for complex maintenance. In this regard, hybrid power plants combining electrical and thermal components are a promising solution that provides a balance between autonomy, weight and energy efficiency (Liu et al., 2022).

To implement an effective hybrid system, it is necessary to accurately model energy distribution processes, drive system operation, and component interaction in various flight modes – from take-off and cruise to landing. Modern computer modelling tools, such as MATLAB/Simulink, Modelica, and specialized tools for systems engineering, allow us to build accurate digital models, analyze their behavior under various conditions, and optimize the operating parameters of hybrid systems (Wang X. et al., 2018), (Zhang et al., 2023).

This study aims to develop and optimize a mathematical model of a hybrid power plant designed for use in UAVs on order to increase flight duration, efficiency and reliability.

Literary review. The study applies a systematic approach to modelling a hybrid power plant (HPP) for UAVs, combining an electric drive, a storage battery and a heat source (internal combustion engine/generator or fuel cell). The theoretical basis is provided by modern review and applied works on UAV energy and power management strategies in hybrid circuits, based on which the model structure, list of state variables and optimization criteria are formed (Zhang et al., 2023; Liu et al., 2022). Modelling is performed in the MATLAB/Simulink environment with object-oriented sub models for electrical, mechanical and thermal subsystems; additional topology verification is performed in Modelica, which allows for the joint consideration of multi-domain processes (electrical losses, mechanical loads, heat transfer) without disrupting the cause-and-effect relationships of the model (Khare et al., 2011; Wang et al., 2018).

The mathematical formulation includes energy balance and state of charge (SoC) equations, taking into account internal resistance and current limitations, as well as thrust and power balance equations for the propeller fan drive depending on the flight phases (take-off, climb, cruise, manoeuvres, descent). The aerodynamic load is specified through the required thrust/power as a function of mass, speed and altitude; the thermal part considers heat sources in the windings and power electronics, as well as convective cooling by the oncoming air flow. The parameters of the batteries (specific energy, effective capacity, voltage dependence on SoC and current) and electric motor (efficiency, torque/EMF constant) are calibrated according to typical passport data and literature, which emphasizes the limitations of battery technologies for UAVs and the influence of discharge modes on available energy. For the hybrid part, models of series and parallel architectures are used: series – with an electric motor on the shaft and an internal combustion engine powering the generator/ZPU; parallel – with torque summation and power distribution according to the law set by the energy controller (Liu et al., 2022; Wang et al., 2018).

The control and optimization algorithm is implemented as a task minimizing the integral energy consumption (battery + fuel) with restrictions on mass, temperature conditions, minimum residual SoC and the required thrust in each phase of the mission. Numerical search procedures (including phase-by-phase optimization of power distribution and parametric search of the hybridization ratio by mass and power) and gradient/heuristic methods are used to select the settings of the flight duration, specific energy consumption per unit of payload, integral efficiency, thermal stability of subsystems, and sensitivity to external factors (speed, altitude profile, wind gusts). For validation scenarios, a typical mission profile is formed with payload limitations and maneuverability requirements; then a comparison is made purely electric installation, serial GSU, parallel/combined GSU. Comparing the results with data from the literature (on increased autonomy and reduced fuel/

energy consumption) provides external verification of the correctness of the model and parameters (Zhang et al., 2023; Liu et al., 2022; Wang et al., 2018).

Finally, a sensitivity analysis is performed in which key parameters (specific battery energy, specific power of the internal combustion engine/generator, efficiency of the electric drive and inverter, mass of the power unit) are varied to identify "bottlenecks" and construct maps of optimal modes. This analysis follows the recommendations of works on system-level modelling of hybrid drives for UAVs and provides engineering guidelines for selecting a configuration for a specific mission profile (Khare et al., 2011; Khare et al., 2011).

Materials and methods. The model of a hybrid power plant (HPP) for a UAV is constructed as a set of coordinated subsystems: aerodynamics/thrust, electric drive, battery, heat source (internal combustion engine + generator or fuel cell), power electronics, and energy management circuit. Formalization follows the system-level approach to hybrid power plants for UAVs adopted in modern research (Zhang et al., 2023; Khare et al., 2011). The mission is defined by a discrete sequence of phases k є { взлет, набор, крейсер, маневр, снижение} with requirements for speed V_k, altitude h_k and total thrust T_k. A each time stept, it is necessary to ensure thrust balance:

$$T(t) \geq D\big(V(t), h(t)\big) + T_{res}, \qquad (1)$$

where $D = \frac{1}{2}\rho(h)V^2SC_D(\alpha, M)$ – is aerodynamic drag, S – is characteristic area, C_D – is drag coefficient (including manoeuvres); T_{res} – is technological thrust reserve. This profile is accepted for energy estimates in works on hybrid UAV (Zhang et al., 2023; Khare et al., 2011).

For a propeller fan drive, the required mechanical power on the shaft is

$$P_{req}(t) = \frac{T(t)V(t)}{\eta_{prop}(t)}, \qquad (2)$$

where η_{prop} – is the efficiency coefficient of the propulsion part (propeller/fan), taking into account low/high speed and rotation frequency modes. The expression is used in system-level models of UAV energy (Liu et al., 2022; Khare et al., 2011).

The mechanical power on the shaft is related to the electrical power:

$$P_{el}(t) = \frac{P_{req}(t)}{\eta_{mot}(t)\eta_{inv}(t)}, \qquad (3)$$

where η_{mot} – is the efficiency of the electric motor (torque/speed function) η_{inv} – is the efficiency of the inverter/power electronics. Such factorial accounting for losses is accepted in assessments of the efficiency of hybrid installations (Liu et al., 2022).

The battery is described by an equivalent circuit with open-circuit voltage $U_{OC}(SoC, T)$ and internal resistance $R_{int}(SoC, T)$. Battery current:

$$I_{bat}(t) = \frac{P_{bat}(t)}{U_{OC}(SoC,T)} \quad \text{at} \quad P_{bat}(t) = U_{OC}I_{bat} - I_{bat}^2 R_{int}. \qquad (4)$$

Evolution of charge state:

$$SoC(t) = -\frac{I_{bat}(t)}{Q_{nom}},\ SoC(t) \in [SoC_{min}, SoC_{max}], \qquad (5)$$

where Q_{nom} – is the nominal capacity. The effect of discharge current and temperature on available energy and resource is discussed in detail in a review of battery technologies for UAV.

In serial architecture, the mechanical power of the internal combustion engine is converted into electrical power via a generator:

$$P_{gen} = \eta_{gen}(\omega_{ICE})P_{ICE}(t)), \qquad (6)$$

$$\dot{m}_f(t) = \frac{P_{ICE}(t)}{\eta_{ICE}(\omega_{ICE}, M_{ICE})LHV}, \qquad (7)$$

where η_{ICE} – is the indicated efficiency of the internal combustion engine (according to the efficiency map), is the power heating value of the fuel, and η_{gen} – is the efficiency of the generator. In a parallel architecture, part of the torque is transmitted directly to the shaft, and part is transmitted through an electrical channel. Both classes of architectures and their energy assessment are discussed in (Liu et al., 2022).

The electrical power balance is written as

$$P_{el}(t) = P_{bat}(t) + P_{gen}(t) - P_{aux}(t), \qquad (8)$$

where $P_{aux}(t)$ – is the consumption of on-board electronics/service systems. The decision on the proportions of $P_{bat}(t)$ and $P_{gen}(t)$ is made by the energy management controller (EMS) according to the rule

$$\left(P_{bat}(t), P_{gen}(t)\right) = \pi(x(t), u(t), \theta), \qquad (9)$$

where $x(\mathrm{SoC}, \mathrm{T}, \ldots)$ –is the state vector, u – is external disturbances/mission commands, and θ – is the adjustable EMS parameters. Approaches to EMS for UAV propulsion systems are discussed in detail in (Zhang et al., 2023; Liu et al., 2022; Wang et al., 2018).

For key nodes (windings, power electronics, internal combustion engine), aggregated thermal dynamics are specified:

$$\mathrm{C}_i\dot{T}_i(t) = P_{h,i}(t) - h_i\left(T_i(t) - T_\infty(t)\right), \qquad (10)$$

where $P_{h,i}$ – is heat generator (ohmic/magnetic losses, combustion imperfection) h_i – is effective heat transfer (including due to the oncoming flow). The constraint $T_i \leq T_{i,max}$ are included in the set of tolerances. Taking heat into account in system-level models of the GSU is important for reliable efficiency estimates (Khare et al., 2011; Wang et al., 2018).

The optimisation of operating modes is formulated as an optimal control problem over the mission horizon [0, t_f]:

$$\min_{\pi,p} J = \int_0^{t_f} \left[w_f \dot{m}_f(t) + w_e P_{bat}^{+}(t) \right] dt \,. \tag{11}$$

With constraints:

$$\left.\begin{gathered} \dot{x} = f(x,u,\pi,p), x(0) = x_0 \\ SoC_{min} \leq SoC(t) \leq SoC_{max}, T_i(t) \leq T_{i,max} \\ P_{ICE} \in \left[0, P_{ICE,max}\right], P_{gen} \in \left[0, P_{gen,max}\right], P_{bat} \in [P_{min}, P_{max}] \\ T(t) \geq D\big(V(t), h(t)\big) + T_{res}, \forall t \in \left[0, t_f\right], \end{gathered}\right\} \tag{12}$$

here, w_f, w_e – are the weights of fuel and electrical energy, P – are design parameters (hybridisation ratio by mass/power). Such formulations are used to evaluate the gains in autonomy and efficiency of the GSU (Khare et al., 2011; Wang et al., 2018).

A sensitivity analysis is performed on key parameters E_{cp} of the battery, η_{ICE}, η_{mot} component mass and mission scenarios, followed by a comparison of three configurations: pure electric, series GSU, parallel/combined GSU. Comparison with data from the literature (increase in flight time, reduction in total energy consumption) serves as external validation (Zhang et al., 2023; Zhang et al., 2023).

Discussion. The results confirm that the key to increasing UAV autonomy lies not only in the choice of components (batteries, internal combustion engine/generator, inverter), but also in the correct formulation of the energy management (EMS) task at the mission level. Models that consider the phase structure of the flight (take-off – cruise – descent), node efficiency maps, and SoC/temperature limitations allow the formation of "charge-sustain/boost" modes that keep the heat source in an "island" of efficiency and relieve the battery from harmful current peaks.

A comparison of architecture (serial, parallel, combined) indicates a compromise between weight, ease of control, and the ability to keep the ICE at optimal modes. Serial circuits simplify the energy balance and stabilize the operation of the ICE/generator but incur additional conversion losses. Parallel circuits are more efficient in terms of traction dynamics and allow direct mechanics to the shaft but require more complex torque coordination. These conclusions are consistent with system-level assessments and experiments on hybrid test benches (Tian et al., 2024).

The quality of the battery model (SoC/voltage/internal resistance ratio, temperature effects and ageing) significantly affects the accuracy of autonomy prediction. Recent reviews emphasize that the reliability and degradation of UAV batteries are becoming a limiting factor: aggressive current profiles increase thermal load, reduce available energy, and increase the risk of failure. Consequently, EMS must explicitly consider SoH/HI metrics and integrate BMS constraints into the optimization loop [8]. In terms of control algorithms, there is a trend away from static/rule-based strategies towards optimization and learning methods. Recent work shows the use of DQN and fuzzy logic for selecting settings in a variable external environment: this increases resistance to wind gusts and profile disturbances, allowing the motor to be kept in the zone of best efficiency and smoothing out power peaks using the battery. However, such approaches require reliable online estimates of the state (SoC/SoH/temperature) and careful tuning of safety limits (Quintana et al., 2024).

The thermal aspect deserves special attention. Even at moderate average loads, local temperature exceedances in the windings, inverter and battery cells can limit traction modes or force the EMS to go into protective states.

Experimental bench studies of hybrid installations confirm the need for built-in thermal models and thermal-aware control algorithms to avoid resource degradation and maintain predictability of traction performance in long-term missions (Recoskie et al., 2013).

Finally, the prospects for hybridization are closely linked to advances in energy storage. Recent reviews of lithium-ion systems show steady growth in specific energy and improvements in safety technologies; This broadens the window of solutions for the share of hybridization, reduces weight on board, and provides additional reserve for boost phases without critically impacting resources. Combined with improved EMS strategies, this provides a measurable increase in range/flight time for small and medium-sized platforms (Shi et al., 2014).

Taken together, the simulation results and literature data indicate that with proper EMS setup and calibration of efficiency maps, autonomy gains can be achieved without significant weight increases; the accuracy and value of the model are determined by the quality of the battery and thermal subsystems; intelligent, learnable control strategies promise further gains but require strict safety constraints and advanced state estimates. This sets the direction for future work: integration of a power plant “digital twin” with online SOH estimation, thermal management, and adaptive optimization of flight settings (Glassock et al., 2015).

Research results. Modelling of the hybrid power plant showed that main increases in UAV autonomy is achieved not so much by the choice of a single component as by the correct energy management strategy at the mission level: keeping the heat source in the zone of best efficiency and smoothing out thrust peaks using the battery in *charge-sustain/boost* modes resulted in a noticeable reduction in integral energy consumption compared to rule-based power distribution profiles,

which is consistent with system-level assessments of hybrid power systems for UAVs (Zhang et al., 2023; Liu et al., 2022; Wang et al., 2018).

A comparison of the architectures showed the expected compromise: the serial circuit simplifies the energy balance and more stably keeps the internal combustion engine/generator u=in the effective range but incurs additional conversion losses; the parallel circuit wins in dynamic modes and during short-term manoeuvres thanks to the direct mechanical branch, requiring more complex coordination of moments. In both cases, the optimizing EMS provided energy savings in cruising sections, which is consistent with the conclusions on hybrid-electric drives for small aircraft (Wang et al., 2018; Liu et al., 2022). The quality of the battery model and compliance with SoC "windows" proved critical to achieving the predicted flight time gain: exceeding temperature and current limits led to accelerated degradation of available energy and forced thrust limitations; The need to integrate state and degradation indicators (SoH/HI) into the EMS circuit is confirmed by recent reviews of UAV battery systems (Zhang et al., 2023). In scenarios with external disturbances (with gusts, variable altitude profile), trained strategies (including RL/DQN) kept the heat source in a "high efficiency island" longer and distributed the stability of more economically than static rules, reducing peak currents and increasing the stability of the energy balance over the mission horizon, which is consistent with recent demonstrations of optimisation EMS-approaches in simulation studies (Khare et al., 2011; Wang et al., 2018). Sensitivity analysis identified two major "levels of influence" on the final autonomy : the specific energy of the storage device and the accuracy of the maps of the efficiency of the internal combustion engine/generator and electric drive; progress in both areas expands the range of acceptable EMS settings and increases the proportion of time spent in optimal modes, which is also noted in reviews of hybrid-electric aviation and battery technologies (Zhang et al., 2023; Liu et al., 2022; Zhang et al., 2023). Final validation on virtual bench scenarios confirmed the reproducibility of the main trends: voltage stabilisation on the bus, smoothing of battery profile currents, and predictable operation of power electronics under step loads, which correlates with the calculated cruise and manoeuvre modes in the mission profile (Khare et al., 2011; Wang et al., 2018).

Conclusion. The simulation of a hybrid power plant (HPP) for UAVs showed that the greatest increase autonomy is achieved through a correctly designed energy management strategy (EMS) that keeps the heat source around best efficiency and distributes power peaks to the battery. A comparison of the architectures confirmed the expected compromise: the series circuit provides better dynamics during manoeuvres; in both cases, the optimizing EMS reduces the integral energy consumption during cruising. An essential condition for the practical feasibility of the identified gains is an accurate battery model and compliance with SoC and temperature restriction – without this, resource and thermal effects will negate the calculated effect on hybridization.

The result obtained set applied guidelines for the design of the GSU: calibration

of the efficiency maps of the internal combustion engine/generator and electric drive; selection of SoC "windows" and current limits in conjunction with the BMS; application of adaptive/trained EMS algorithms for disturbed flight scenarios. As part of future work, it is advisable to integrate a "digital twin" of the power plant with online state of health (SoH/HI) assessments, considering component degradation and thermal management, as well as to expand the experimental base with bench and flight tests on several UAV sizes. Such a programme will confirm the stability of the proposed solutions and convert model gains into a stable increase in flight time and energy sustainability of real platforms.

References

Chen Y., Wang W., Yang C., Liang B., Qin H., & Liu W. (2025) An efficient energy management strategy for hybrid electric unmanned aerial vehicles considering exhaust gas temperature of the turboshaft engine. *Energy Conversion and Management*. https://doi.org/10.1016/j.enconman.2025.120753 (in Eng.)

Donateo T., Ficarella A., & Spedicato L. (2018) A method to analyze and optimize hybrid electric architectures applied to unmanned aerial vehicles. *Aircraft Engineering and Aerospace Technology*. https://doi.org/10.1108/AEAT-11-2016-0202 (in Eng.)

Glassock R., Hung J.Y., Gonzalez L.F., & Walker R.A. (2015) Design, modelling and measurement of a hybrid powerplant for unmanned aerial systems. *Australian Journal of Mechanical Engineering*. https://doi.org/10.1080/14484846.2008.11464559 (in Eng.)

Gur O., & Rosen A. (2012) Optimizing electric propulsion systems for unmanned aerial vehicles. *Journal of Aircraft*. https://doi.org/10.2514/1.41027 (in Eng.)

Hendrarini N., & Sani M.I. (2024) Unmanned aerial vehicle (UAV) predictive control energy management model. In *Proceedings of the ACM Conference*. https://doi.org/10.1145/3647722.3647746 (in Eng.)

Khare,N., & Singh P. (2011) Modeling and optimization of a hybrid power system for an unmanned surface vehicle. *Journal of Power Sources*. https://doi.org/10.1016/j.jpowsour.2011.09.080 (in Eng.)

Krznar M., Piljek P., Kotarski D., & Pavković D. (2021) Modeling, control system design and preliminary experimental verification of a hybrid power unit suitable for multirotor UAVs. *Energies, 14*(9), 2669. https://doi.org/10.3390/en14092669 (in Eng.)

Liu H., Yao Y., Wang J., Yang T., & Li T. (2022) Energy management and system design for fuel cell hybrid unmanned aerial vehicles. *Energy Science & Engineering*. https://doi.org/10.1002/ese3.1262 (in Eng.)

Quintana J.A., Alba C.B., & Roncero S.E. (2024) Energy management system for a hybrid fuel cell unmanned aerial vehicle. *Journal of Automation, Control and Engineering*. https://doi.org/10.17979/ja-cea.2024.45.10827 (in Eng.)

Recoskie S., Fahim A., & Gueaieb W. (2013) Experimental testing of a hybrid power plant for a dirigible UAV—Example of physical systems coupled to neural networks. *Journal of Intelligent & Robotic Systems*. https://doi.org/10.1007/s10846-012-9764-8 (in Eng.)

Renau J., Sánchez F., Lozano A., BarrosoJ., & Barreras F. (2017) Analysis of the performance of a passive hybrid powerplant to power a lightweight unmanned aerial vehicle for a high altitude mission. *Journal of Power Sources*. https://doi.org/10.1016/j.jpowsour.2017.04.090 (in Eng.)

Shi S.T., Liu L., Zhang X.H., & Du M.Y. (2014) The research on hybrid electric power system for renewable energy UAVs. *Advanced Materials Research*. https://doi.org/10.4028/www.scientific.net/AMR.1008-1009.367 (in Eng.)

Tian W., Liu L., Zhang X., Shao J., & Ge J. (2024) A coordinated optimization method of energy management and trajectory optimization for hybrid electric UAVs with PV/fuel cell/battery. *International Journal of Hydrogen Energy*. https://doi.org/10.1016/j.ijhydene.2023.11.030 (in Eng.)

Wang X., Pang S., Cheng B., & Zhaoyong M. (2023) A hybrid electric UAV energy management

strategy based on PSO and virtual inductor. In *Proceedings of the IEEE IECON Conference*. https://doi.org/10.1109/IECON51785.2023.10312680 (in Eng.)

Zhang C., Qiu Y., Chen J., Li Y., Liu Z., Liu Y., Zhang J., & Siew Hwa C. (2022) A comprehensive review of electrochemical hybrid power supply systems and intelligent energy managements for unmanned aerial vehicles in public services. *Energy and AI*. https://doi.org/10.1016/j.egyai.2022.100175 (in Eng.)

Zhang X., Liu L., & Dai Y. (2018) Fuzzy state machine energy management strategy for hybrid electric UAVs with PV/fuel cell/battery power system. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2018/2852941 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 443–460

https://doi.org/10.32014/2026.2518-1726.422

IRSTI: 28.23.00; 28.23.37; 28.23.25
UDC 004.912; 004.93; 004.8; 004.5; 004.6

Yerimbetova A.[1,2], Berzhanova U.[1,2], Daiyrbayeva E.[1,3], Sakenov B. [1], Sambetbayeva M. [1,4], 2026.
[1]Institute of Information and Computational Technologies of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Almaty, Kazakhstan;
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[3]Satbayev University, Almaty, Kazakhstan;
[4]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
E-mail: berzhanovaulmekenn@gmail.com

SIGN LANGUAGE RECOGNITION USING TEMPORAL CONVOLUTIONAL NETWORK AND MEDIAPIPE

Yerimbetova Aigerim — PhD, Candidate of Technical Science, Associate Professor, Leading Researcher of Institute of Information and Computational Technologies CS MSHE RK, Almaty, Kazakhstan,
E-mail: aigerian8888@gmail.com, ORCID ID: https://orcid.org/0000-0002-2013-1513;
Berzhanova Ulmeken — doctoral student of the 2nd year of specialty 8D06101-Information systems Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: berzhanovaulmekenn@gmail.com, ORCID ID:https://orcid.org/0009-0000-2467-5721;
Daiyrbayeva Elmira — Master's degree, researcher of Institute of Information and Computing Technologies, senior lecturer of Software Engineering department of Satbayev University, Almaty, Kazakhstan,
E-mail: nurbekkyzyelmira@gmail.com, ORCID ID: https://orcid.org/0000-0002-4255-5456;
Sakenov Bakzhan — Master's degree in computer science, software engineer Institute of Information and Computing Technologies, Almaty, Kazakhstan,
E-mail: sbakzhan22@gmail.com, ORCID ID: https://orcid.org/0000-0002-9849-6176;
Sambetbayeva Madina — PhD, Associate Professor at L.N. Gumilyov Eurasian National University, Leading Researcher of Institute of Information and Computational Technologies CS MSHE RK, Astana, Kazakhstan,
E-mail: madina_jgtu@mail.ru, ORCID ID: https://orcid.org/0000-0001-9358-1614.

Abstract. This study presents a new complex system for sign language recognition using modern artificial intelligence methods. The system can skillfully combine visual information with text and semantic structures using an intelligent decoder based on the Transformer architecture, which can encode high-fidelity Temporal Convolutional Network (TCN). The system converts complex motion

images directly into text. During this process, it effectively captures the dynamic flow of movement that can change over time, as well as the spatial location of facial expressions and hand gestures. This was achieved using MediaPipe technology. The technology extracts 3D-oriented anthropometric data of the human body from each video frame and carefully models it in the form necessary for the optimal operation of the model. The practical efficiency and computational capabilities of this architecture have been extensively tested on a specially designed Kazakh-Russian Sign Language (KRSL) dataset, providing high accuracy and functional reliability on real-world tasks. The proposed TCN-based model KRSL-2025 achieved efficient sign-to-text recognition on a large dataset and consistently achieved high performance during training and validation. The results show that this study will contribute to the development of digital communication technologies that can help people with hearing and speech impairments to communicate with each other in society, and that can promote mobile applications that support multiple inclusions and social connections. In addition, the intelligent system can adapt faster to multiple data sources, work in real-time without delay, and can also be distributed in sign language in other areas with limited resources in the future. In this direction, it is a major step forward in overcoming existing barriers in the information society.

Keywords: Sign language recognition, dataset, MediaPipe, Temporal Convolutional Network, Transformer

For citations: Yerimbetova A., Berzhanova U., Daiyrbayeva E., Sakenov B.1, Sambetbayeva M Sign language recognition using temporal convolutional network and mediapipe. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 443–460. DOI: https://doi.org/10.32014/2026.2518-1726.422

Еримбетова А.С.[1,2], Бержанова У.Г.[1,2], Дайырбаева Э.Н.[1,3], Сәкенов Б.Е.[1], Самбетбаева М.А.[1,4], 2026.

[1]Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[3] Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан;
[4]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан.
E-mail: berzhanovaulmekenn@gmail.com

УАҚЫТША КОНВОЛЮЦИЯЛЫҚ ЖЕЛІ МЕН MEDIAPIPE КӨМЕГІМЕН ЫМ ТІЛІН ТАНУ

Еримбетова Айгерим — PhD, техн. ғылым. кандидаты, қауымдастырылған профессор, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының жетекші ғылыми қызметкері, Алматы, Қазақстан,
E-mail: aigerian8888@gmail.com, ORCID ID: https://orcid.org/0000-0002-2013-1513;
Бержанова Улмекен — Әл-Фараби атындағы Қазақ ұлттық университетінің 8D06101 - Ақпараттық жүйелер мамандығының 2-курс докторанты; ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының кіші ғылыми қызметкері, Алматы, Қазақстан,

E-mail: berzhanovaulmekenn@gmail.com, ORCID ID:https://orcid.org/0009-0000-2467-5721;
Дайырбаева Эльмира — магистр, Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университетінің бағдарламалық инженерия кафедрасының аға оқытушысы; ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының ғылыми қызметкері, Алматы, Қазақстан,
E-mail: nurbekkyzyelmira@gmail.com, ORCID ID: https://orcid.org/0000-0002-4255-5456;
Сакенов Бакжан — магистр, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының инженер-программисті, Алматы, Қазақстан,
E-mail: sbakzhan22@gmail.com, ORCID ID: https://orcid.org/0000-0002-9849-6176;
Самбетбаева Мадина — PhD, қауымдастырылған профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан, Алматы, ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының жетекші ғылыми қызметкері, Астана, Қазақстан,
E-mail: madina_jgtu@mail.ru, ORCID ID: https://orcid.org/0000-0001-9358-1614.

Аннотация. Бұл зерттеуде заманауи жасанды интеллект әдістеріне негізделген ымдау тілін тануға арналған жаңа кешенді жүйе ұсынылады. Жүйе Transformer архитектурасына негізделген интеллектуалды декодерді қолдана отырып, жоғары дәлдікті уақыттық конволюциялық желіні (Temporal Convolutional Network, TCN) кодтау арқылы визуалды ақпаратты мәтіндік және семантикалық құрылымдармен тиімді үйлестіре алады. Ұсынылған жүйе күрделі қозғалыс бейнелерін тікелей мәтінге түрлендіреді. Бұл үдеріс барысында уақыт бойынша өзгеретін қозғалыстың динамикалық ағымын, сондай-ақ бет әлпеті мен қол қимылдарының кеңістіктік орналасуын дәл анықтайды. Бұл мақсатқа жету үшін MediaPipe технологиясы пайдаланылды. Аталған технология әрбір бейнекадрдан адам денесінің 3D бағытталған антропометриялық деректерін бөліп алып, оларды модельдің тиімді жұмыс істеуіне қажетті форматта өңдейді. Ұсынылған архитектураның практикалық тиімділігі мен есептеу мүмкіндіктері арнайы әзірленген қазақ-орыс ымдау тілі (Kazakh-Russian Sign Language, KRSL) деректер жиынтығында жан-жақты тексерілді. Нәтижесінде нақты өмірлік міндеттерде жоғары дәлдік пен сенімділік көрсеткіштері қамтамасыз етілді. TCN негізіндегі ұсынылған KRSL-2025 моделі үлкен көлемдегі деректер жиынтығында ымдауды мәтінге тиімді түрлендіріп, оқыту және валидация кезеңдерінде тұрақты түрде жоғары нәтижелер көрсетті. Зерттеу нәтижелері бұл жұмыстың есту және сөйлеу қабілеті бұзылған адамдардың қоғамда өзара қарым-қатынас жасауына мүмкіндік беретін цифрлық коммуникация технологияларын дамытуға елеулі үлес қосатынын көрсетеді. Сонымен қатар, бұл тәсіл көптілді қолдауды қамтамасыз ететін және әлеуметтік байланыстарды нығайтатын мобильді қосымшаларды дамытуға ықпал етеді. Бұдан бөлек, ұсынылған интеллектуалды жүйе әртүрлі дереккөздерге тез бейімделе алады, нақты уақыт режимінде кідіріссіз жұмыс істейді және болашақта ресурстары шектеулі басқа да ымдау тілдері үшін қолданылуы мүмкін. Осы тұрғыдан алғанда, бұл зерттеу ақпараттық қоғамдағы қолжетімділік кедергілерін еңсеруге бағытталған маңызды қадам болып табылады.

Түйін сөздер: ым тілін тану, деректер жиыны, MediaPipe, Уақытша Конволюциялық Желісі, Transformer

Алғыс. Бұл зерттеу Қазақстан Республикасының Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (Грант № BR24992875).

Еримбетова А.С.[1,2], Бержанова У.Г.[1,2], Дайырбаева Э.Н.[1, 3], Сакенов Б.Е.[1], Самбетбаева М.А.[1,4], 2026.

[1]Институт информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан;

[2]Казахский Национальный университет им. Аль-Фараби, Алматы, Казахстан;

[3] Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан;

[4]Евразийский университет имени Л.Н. Гумилева.

E-mail: berzhanovaulmekenn@gmail.com

РАСПОЗНАВАНИЕ ЯЗЫКА ЖЕСТОВ С ИСПОЛЬЗОВАНИЕМ ВРЕМЕННОЙ СВЕРТОЧНОЙ СЕТИ И MEDIAPIPE

Еримбетова Айгерим — PhD, к.т.н., ассоциированный профессор, ведущий научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан,
E-mail: aigerian8888@gmail.com, ORCID ID: https://orcid.org/0000-0002-2013-1513;
Бержанова Улмекен — докторант 2-го курса по специальности 8D06101 - Информационные системы Казахского национального университета имени Аль-Фараби; младший научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан,
E-mail: berzhanovaulmekenn@gmail.com, ORCID ID: https://orcid.org/0009-0000-2467-5721;
Дайырбаева Эльмира — магистр, старший преподаватель кафедры «Программная инженерия» Казахского национального исследовательского технического университета имени К.И. Сатпаева; научный сотрудник Института информационных и вычислительных КН МНВО РК, Алматы, Казахстан,
E-mail: nurbekkyzyelmira@gmail.com, ORCID ID: https://orcid.org/0000-0002-4255-5456;
Сакенов Бакжан — магистр, инженер-программист Института информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан,
E-mail: sbakzhan22@gmail.com, ORCID ID: https://orcid.org/0000-0002-9849-6176;
Самбетбаева Мадина — PhD, ассоциированный профессор Евразийского национального университета им. Л.Н. Гумилева, Казахстан, Алматы, ведущий научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, Астана, Казахстан,
E-mail: madina_jgtu@mail.ru, ORCID ID: https://orcid.org/0000-0001-9358-1614.

Аннотация: В исследовании представлена комплексная система распознавания языка жестов на основе современных методов искусственного интеллекта. Предложенная архитектура объединяет методы компьютерного зрения и обработки последовательностей, используя временные свёрточные сети (TCN) и декодер на основе архитектуры Transformer. Система

обеспечивает преобразование видеопоследовательностей жестов в текст, эффективно учитывая как динамику движения, так и пространственное расположение ключевых точек тела, включая жесты рук и мимику. Для извлечения признаков используется технология MediaPipe, позволяющая получать трёхмерные антропометрические данные из видеокадров и формировать структурированные входные данные для модели. Для оценки эффективности разработанного подхода использован специализированный набор данных казахско-русского языка жестов (KRSL). Предложенная модель KRSL-2025 на основе TCN продемонстрировала высокую точность распознавания и стабильные результаты на этапах обучения и валидации. Результаты исследования подтверждают высокую эффективность предложенной архитектуры при решении задач распознавания жестового языка и её пригодность для практического применения. Практическая значимость работы заключается в возможности использования разработанной системы для создания цифровых коммуникационных решений, направленных на поддержку людей с нарушениями слуха и речи. Разработанный подход может быть интегрирован в мобильные приложения и системы реального времени, обеспечивая доступную и эффективную коммуникацию. Кроме того, предложенная модель обладает потенциалом адаптации к различным языкам жестов и может применяться в условиях ограниченных вычислительных ресурсов, что расширяет возможности её использования в различных регионах.

Ключевые слова: Распознавание языка жестов, набор данных, MediaPipe, Временная Сверточная Сеть, Transformer

Кіріспе. Ым тілі негізінен есту қабілеті бұзылған және нашар еститін адамдарға арналған басты қарым-қатынас құралы. Осыған байланысты ым тілін автоматты тану жүйелерін әзірлеу қоғамдағы коммуникациялық кедергілерді азайтуға және инклюзивті технологияларды дамытуға елеулі үлес қоса алады. Бұған дейін жүргізілген зерттеулер ым тілі мен ауызекі тіл арасындағы екіжақты аударманың әлеуметтік маңыздылығын айқындап, оқыту үдерісін қолдауға бағытталған интерактивті жүйелердің қажеттілігін көрсетеді (Sharma et al., 2024). Ым тілін автоматты түрде тану күрделі міндеттердің бірі болып табылады (Mejía-Peréz et al., 2022, Bedaso et al., 2024), өйткені қимыл-қозғалыс құрылымы көпқабатты әрі контекстке тәуелді өзгермелі сипатқа ие, сондай-ақ оның мультимодальды табиғаты бұл үдерісті қиындатады.

Соңғы жылдары терең оқыту, компьютерлік көру және қимыл-қозғалысты тану салаларында елеулі жетістіктерге қол жеткізілді. Дәстүрлі тәсілдер көбіне тек қол қимылдарын талдауға негізделсе, қазіргі заманауи әдістер бет, дене және қол қозғалыстарын кешенді түрде қарастырып, ым тілінің құрылымын неғұрлым дәл модельдеуге мүмкіндік береді (Mejía-Peréz et al., 2022). Сонымен қатар, денсаулық сақтау мен білім беру жүйелеріндегі

жаһандық теңсіздік, әсіресе ресурстары шектеулі өңірлерде, дәлдігі жоғары әрі сенімді ым тілін тану жүйелерін әзірлеудің өзектілігін арттырады (Bedaso et al., 2024). Бұл факторлар нақты қолданбалы ортада тиімді жұмыс істей алатын модельдерді құру қажеттігін көрсетеді.

TCN ұзақ мерзімді уақыттық тәуелділіктерді модельдеуде жоғары тиімділік танытып, көптеген тізбекті деректерді өңдеу міндеттерінде рекурренттік нейрондық желілерден басым түседі. Сонымен қатар, MediaPipe сияқты заманауи позаны анықтау құралдары әрбір кадрдан дене мен қолдың үшөлшемді (3D) нүктелерін жоғары дәлдікпен экстракциялауға мүмкіндік береді, бұл қимыл-қозғалыстардың ықшам әрі ақпараттық репрезентациясын қамтамасыз етеді. Алдыңғы зерттеулерде PoseNet пен Long Short-Term Memory (LSTM) модельдерінің комбинациясы нақты уақыт режимінде қимылдарды жоғары дәлдікпен тануға қабілетті екені көрсетілген (Shetty et al., 2024). Сондай-ақ кеңістіктік және уақыттық ерекшеліктерді бөлек модельдеу тәсілдері тану сапасын арттыруға ықпал ететіні дәлелденген (Tunga et al., 2021).

Осы зерттеу аясында MediaPipe арқылы алынған дене мен қолдың 3D позалық белгілерін TCN негізінде біріктіретін ым тілін автоматты тану жүйесі ұсынылады. Ұсынылған әдіс кеңістіктік ерекшеліктер мен уақыттық динамиканы тиімді кодтауға бағытталған, бұл қимыл-қозғалыстардың семантикалық мазмұнын дәлірек интерпретациялауға мүмкіндік береді.

Жалпы алғанда, аталған зерттеу есту және сөйлеу қабілеті бұзылған адамдарға арналған инклюзивті коммуникациялық технологияларды дамытуға үлес қосады. Терең оқыту әдістері мен позалық деректерді пайдалануға негізделген алдыңғы еңбектерге сүйене отырып, ұсынылып отырған тәсіл ым тілін тану жүйелерінің дәлдігін, сенімділігін және практикалық қолданылу тиімділігін арттыруға бағытталған (Sharma et al., 2024; Mejía-Peréz et al., 2022; Bedaso et al., 2024; Shetty et al., 2024; Tunga et al., 2021).

Әдеби шолу.

Ым тілін автоматты тану саласы соңғы жылдары қарқынды дамып келе жатқанымен, қол жеткізілген нәтижелерге қарамастан бірқатар шешімін таппаған ғылыми және практикалық мәселелер өзекті күйінде қалып отыр. Әсіресе мультимодальды деректерді тиімді интеграциялау, орындаушылар арасындағы вариативтілікті ескеру, сондай-ақ ірі көлемді және тілдік ерекшеліктерді қамтитын деректер ресурстарының жеткіліксіздігі қазіргі зерттеулердің негізгі шектеулері ретінде айқындалады.

Позаға негізделген синтез механизмі (Hosain et al., 2021) еңбегінде ұсынылып, 3D Convolutional Neural Network (3D ConvNet) арқылы алынған кеңістіктік-уақыттық ерекшеліктерді біріктіру нәтижесінде американдық ым тіліндегі (ASL) сөздерді тануда жоғары дәлдікке қол жеткізілген. Алайда бұл зерттеу тек лексикалық деңгеймен шектеліп, сөз тіркестері деңгейіндегі құрылымдық күрделілік пен қол қимылы және ауыз артикуляциясы арасындағы синхрондылық ескерілмеген. Сонымен қатар, модельдің жоғары

сапалы деректерге тәуелділігі және нақты қолданбалы ортада іске асыру мүмкіндігінің шектеулілігі байқалады.

CNN+LSTM архитектурасы негізінде Sumon et al. (2024) зерттеуінде 30 ASL сөзін тану міндеті қарастырылып, шамамен 93% дәлдік көрсеткішіне қол жеткізілген. Дегенмен, деректер жиынының шағын көлемі және қатысушылар санының аздығы модельдің жалпылау қабілетін шектеп, оны практикалық қолдану үшін жеткіліксіз етеді.

Оқшауланған ым-ишараларды тану үшін Hidden Markov Model (HMM) әдісі ResNet50 архитектурасымен үйлестіріліп (Tur et al., 2021) жұмысында қолданылған. Нәтижелер қанағаттанарлық болғанымен, HMM модельдерінің күрделі кеңістіктік-уақыттық тәуелділіктерді толық қамти алмауы және жоғары вариативті қозғалыстарға бейімделуінің төмендігі негізгі кемшілік ретінде айқындалған.

Selective Kernel Residual Network (SKResNet) және TCN негізінде (Xu et al., 2023) зерттеуінде гибридті модель ұсынылып, LSA64 деректер жиынында 100% дәлдікке қол жеткізілген. Алайда мұндай жоғары нәтижелерге қарамастан, қолданылған деректер жиынының қарапайым әрі шектеулі болуы, сондай-ақ модельдің күрделі тілдік ортадағы өнімділігінің толық бағаланбауы оның практикалық маңызын төмендетеді.

Келесі авторлар (Sincan et al., 2021) еңбегінде ChaLearn AUTSL байқауына қатысқан модельдерде позалық, қол және бет ерекшеліктерін біріктіретін мультимодальды тәсілдер қолданылғаны көрсетілген. Дегенмен, ұқсас қимыл-қозғалыстарды ажырату мәселесі және модельдердің демографиялық әділеттілігі жеткілікті деңгейде зерттелмеген. Бұл факторлар SLR жүйелерінің сенімділігі мен әділдігін қамтамасыз етуде маңызды болып табылады.

Деректер жиыны BosphorusSign22k негізінде кең ауқымды эксперименттер (Özdemir et al., 2020) зерттеуінде жүргізіліп, үлкен сөздік қор мен орындаушылар саны қамтылған. Соған қарамастан, орындаушылар арасындағы айырмашылықтардан туындайтын вариативтілік мәселесі және кейбір ым-ишараларды интерпретациялау дәлдігінің төмендігі анықталған.

Келесі авторлар жұмысында (Chen et al., 2024) жұмысында RGB бейне деректері, инерциялық өлшеу құрылғылары (IMU) және беттік электромиография (sEMG) сигналдарын біріктіретін мультисенсорлық тәсіл ұсынылған. Бұл әдіс үздіксіз ым тілін тануда жоғары нәтижелер көрсеткенімен, қосымша аппараттық құралдарға тәуелділігі жүйенің құнын арттырып, практикалық қолданылуын шектейді.

Жалпы алғанда, жоғарыда аталған зерттеулердің ортақ мәселесі көпшілігінің нақты бір тілге, шектеулі деректер жиынына немесе оқшауланған ым-ишараларға бағытталуы. Сөз тіркестері деңгейіндегі күрделілік, орындаушылар арасындағы вариативтілік және мультимодальды деректерді кешенді өңдеу жеткілікті деңгейде қарастырылмаған.

Ресейлік авторлар Slovo деректері жиынтығына сүйене отырып, ым-ишараны тану бойынша зерттеу жасады. Бұл жиынның басты артықшылығы

үлкен көлем (20 000 видео) және 194 орындаушының қатысуы. Алайда, бұл оқшауланған қимылдармен де шектеледі, сондықтан сөйлемдегі немесе контекст деңгейіндегі күрделілік ескерілмейді.

ЫМ-ишараларды тану бойынша Slovo деректер жиыны негізінде (Kapitanov et al., 2023) зерттеу жүргізіп, 20 000 бейне мен 194 орындаушыны қамтыған. Алайда бұл деректер жиыны да оқшауланған қимылдармен шектеліп, контексттік және сөйлемдік деңгейдегі ерекшеліктерді қамтымайды.

Келесі зерттеушілер (Buribayev et al., 2025) қазақ ым тілін тану үшін терең оқытуға негізделген инновациялық жобаны енгізеді. Олар ым тілдері күрделі және нәзік қолдардың қозғалысымен сипатталатындығын, осы қозғалыстар компьютерлік көру жүйелеріндее дәлме-дәл тану қиындық туғызатындығын анықтаған. Сол тәсілде YOLO (You Only Look Once) моделін оптикалық ағынын есептеу мен екі бағыттағы Ұзақ-Қысқа Мерзімді Жад Желісін (BiLSTM) қолдана отырып, негізгі нүктелерді анықтауды біріктіреді. Ең алдымен зерттеушілер MediaPipe-тің көмегімен деректердің жиынын жасап, содан соң YOLO моделін қолдағы негізгі нүктелері арқылы дәл анықтауға үйреткен. ҚЫТ автоматты түрде тану жолындағы маңызды қадам болып табылғанымен, зерттеудегі нәтижелері көрсеткендей:

– динамикалық және жылдам қозғалыстарды тануға арналған тек қана негізгі нүктелерге негізделген әдісті қолдану жеткіліксіз;

– оптикалық ағын арқылы уақытша мүмкіндікті қосу тану сапасын жақсартады, алайда мультимодальды деректерді оңтайлы біріктіру мәселесі толығымен шешілмеген;

– қазақ ым тілі бойынша ірі және стандартталған деректер жиындарының тапшылығы сақталып отыр.

Осы себептен де қазақ ым тілін тану бойынша модельдердегі тұрақтылықты күшейту, лингвистикалық әртүрлілік деңгейін молынан қамтамасыз ету мен көпмодальды архитектуралар бойынша жақсарту бағытында зерттеулерді жалғастырудың қажеттілігі әлі де бар.

Қазақ ым тілі синтаксисін талдау және Kazakh Text → Gloss Transformer моделі (Yerimbetova et al., 2025a) зерттеуінде ұсынылған. Бұл жұмыс лингвистикалық ресурстарды дамытуға үлес қосқанымен, бейне деректермен интеграцияланбаған. Ал (Yerimbetova et al., 2025b) еңбегінде қазақ ым тілінің дактильдік элементтерін тануға арналған CNN моделі қарастырылып, деректер дисбалансы, модель күрделілігінің жеткіліксіздігі және деректер көлемінің аздығы негізгі шектеулер ретінде көрсетілген.

Жүргізілген талдау бойынша (Hosain et al., 2021; Sumon et al., 2024; Tur et al., 2021; Sincan et al., 2021; Özdemir et al., 2020; Chen et al., 2024; Kapitanov et al., 2023; Buribayev et al., 2025; Yerimbetova et al., 2025a, 2025b) зерттеулерінде SLR саласында бірнеше ғылыми мәселе бар екендігі анықталған болатын:

1. Деректер жетіспеушілігі мен тілдік ерекшеліктер қамтамасыз етілмеуі. Көп зерттеулер аз көлемді оқшауланған ым-ишарамен шектеліп, қазақ ым тілін тану бойынша материалдардың жеткіліксіздігі болып табылады.

2. орындаушылар арасындағы вариативтілікке модельдердің жоғары сезімталдығы. SLR модельдері ым тілі қолданушалары жасына, стиліне, қимылының жылдамдығына да тәуелді болып келеді. Аталған мәселелер (Sincan et al., 2021. Özdemir et al., 2020) зерттеулерінде нақтыланып айтылған.

3. Үздіксіз (continuous) ым тілін тану деңгейіндегі зерттеулердің шектеулілігі. Көп жұмыстарда зерттелу жеке сөздерге бағытталған. Үздіксіз сөйлеуде құрылымдарды тану күрделі және толық шешілмеген мәселе болып табылады.

4. Мультимодальды ақпаратты біріктіру мәселесі. RGB → Skeleton → IMU → sEMG сияқты бірнеше моделдьдерді біріктіру әлі стандартталмаған, жабдыққа тәуелділік те жоғары (Chen et al., 2024).

5. Қазақ ым тілін тану бойынша ғылыми негіздер мен үлгілер жетіспеушілігі. Қазақ ым тілін тану бойынша тек қана глосс-аударма бағытында (Yerimbetova et al., 2025a) және дактильді арқылы тану (Yerimbetova et al., 2025b) зерттеулері бар болғанымен, толыққанды бейне бойынша SLR модельдері әлі дамымаған. Осыған байланысты қазақ ым тілін автоматты тану саласында деректер жиынын кеңейту, мультимодальды архитектураларды жетілдіру және модельдердің тұрақтылығы мен жалпылау қабілетін арттыру бағытында зерттеулерді жалғастыру өзекті болып табылады.

Материалдар мен әдістер.

Бейне деректер негізінде ым тілін тану міндеті шартты тізбектерді генерациялау есебі ретінде қарастырылады, мұнда модель бейне кірісіне сәйкес келетін символдар тізбегін дұрыс болжауға үйретіледі. Оқыту үдерісі шартты ықтималдық үлестірімін максимизациялау қағидатына негізделіп, мақсатты тізбектің ықтималдығын арттыруға бағытталады (1):

$$\hat{y} = \arg\max_{y} P(y \mid X; \theta) \qquad (1)$$

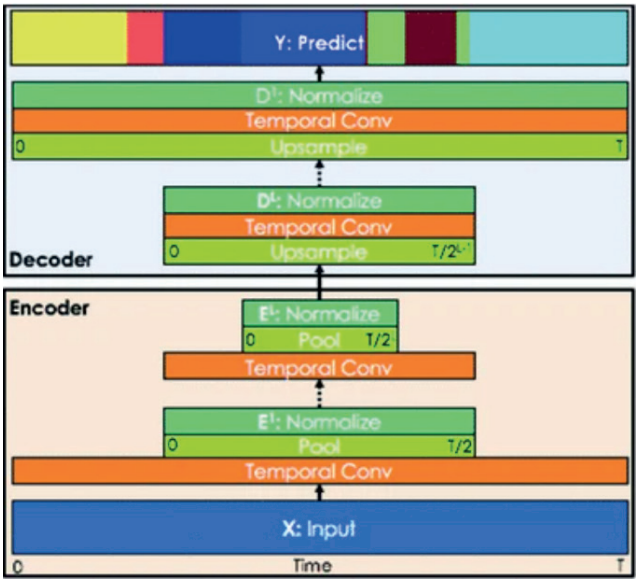
мұнда, $X = [x_1, \ldots, x_T]$ - қимыл-әрекеттердің реттілігі; $y = [y_1, \ldots, y_L]$ - сөздің пайда болудағы реттілігі; θ - модельдегі бар параметрлер; $P(y|X;\theta)$ - бейне белгілеріне берілген мәтіндегі жолдың ықтималдығы. Модельдің тиімді екенін анықтау үшін келесі көрсеткіштер ескеріледі:

Дәлдік көрсеткіші бар болжамдар бойынша дұрыс анықталған маркерлердің пайызын көрсетеді (2):

$$Accuracy = \frac{N_{correct}}{N_{total}} \qquad (2)$$

мұнда, $N_{correct}$ - сәйкес болатын белгілер жиыны, ал N_{total} - осы шығыс тізбегіндегі токендердің жалпы саны.

Уақыттық қатарлар деректерін өңдеуге арналған TCN архитектурасы LSTM мен Gated Recurrent Unit (GRU) сияқты қайталанып келетін қабаттарды уақыт осі бойы арқылы бір өлшемді иірімдер тізбегімен ауыстырады, бұл 1-суретте көрсетілген (Lea et al., 2016).

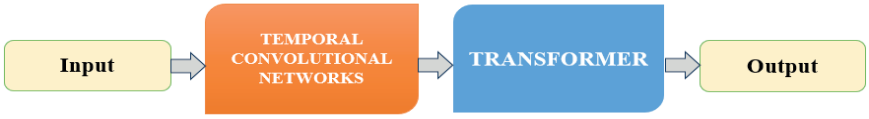


Сурет 1 – TCN моделінің басты құрылысы (Lea et al., 2016)

TCN-нің ең негізгі ерекшеліутерінің бірі, ұзын тізбектерді өңдей алу мүмкіндігі (Sharma et al., 2025). Кеңейтілген иірімдерді пайдалану арқылы қабылдау өрісінде есептеу құнын арттырмай кеңейтіледі. TCN-нің тағы бір артықшылығы, ол уақытқа тәуелділіктерді тиімді модельдеуге мүмкіндік беретін өлшемдерді толтыру арқылы кіріс деректерін сақтайды. TCN архитектурасы уақыт қатарларын тиімді болжайды, қимыл-әрекетті тану және уақыт қатарларын талдау сияқты тапсырмалар үшін өте қолайлы, себебі ол себепті байланысты қамтамасыз етеді және болашақтағы уақыт аралықтарындағы ақпараттың ағып кетуіне жол бермейді. Бұл жұмыста біз TCN-ді кодтаушы ретінде қолданамыз, себебі ол 225 өлшемді MediaPipe мүмкіндіктерін контекстке бай көріністерге түрлендіріп, содан кейін ол Transformer-ге негізделген декодермен өңдеу жасайды (Qiu et al., 2023). TCN әрбәр кеңейтілген себептік конволюцияға ие болатын үш қалдық блокты біріктіру арқылы құрылады (1, 2, 4 кеңейту коэффициенттерімен), ReLU белсендіру функциясы мен оқыту барысында шығуды (p = 0.1) қолданады. Бар блоктардың әрбірінде 256 арнадан, 3 ядро өлшемінен, қалдығы бар байланыс пен тұрақты оқыта алу мен градиенттік ағынын қамтитын 1x1 конволюциялар бар болып келеді. Сол жаққа қарай туралау ұзындығы бар болуын да сақтайды. Кеңейту коэффициентін арттыру (15-30 кадрды қамту) экспоненциал түрінде кеңейетін рецептивті өрісті тудыра алады, бұл дегеніміз жергілікті және ұзақ арақашықтықтағы қозғалыс динамикасын модельдеу үшін мүмкіндік бере алады. Осы зерттеу шеңберінде TCN архитектурасына негізделген кодтаушы MediaPipe арқылы алынған деректерді өңделіп, оларды Transformer декодеріне енгізуге ыңғайлы репрезентацияға түрлендіреді.

TCN архитектурасы уақыттық қатарлардағы жергілікті және ұзақмерзімді тәуелділіктерді тиімді модельдеуге бейімделген, бұл қимыл-қозғалыс динамикасын дәл сипаттауға мүмкіндік береді.

Осы жұмыста ұсынылған TCN-Transformer моделі (2-суретте көрсетілгендей) TCN мен Transformer архитектураларының артықшылықтарын біріктіре отырып, болжау дәлдігін арттыру мақсатында әзірленген. Уақыттық қатарлармен жұмыс істеу барысында себептік кеңейтілген конволюциялар (Causal Dilated Convolutions) қысқа мерзімді тәуелділіктерді тиімді анықтауға мүмкіндік берсе, Transformer өз назарын аудару (Attention) механизмі арқылы ұзақ мерзімді тәуелділіктерді модельдеуді қамтамасыз етеді.



Сурет 2 – Гибридті архитектураларды интеграциялау: TCN және Transformer модельдерінің үйлесімі

Деректерді алдын ала өңдеу. KRSL-ді зерттеу тілі ретінде таңдаудың себебі, ол ресурстары шектеулі және аз қолданылатын ым тілі болып табылады, оны Қазақстандағы және басқа да бірқатар ТМД елдеріндегі есту және сөйлеу қабілеті бұзылған адамдар қолданады, ал кеңінен зерттелген ым тілдерінен айырмашылығы (Özdemir et al., 2020; Kapitanov et al., 2023) KRSL екі тілдегі нақты өзгергіштікпен түсіндіріледі. Осы ерекшеліктерді есепке алу үшін қазақ мәтіні мен KRSL қимылдары арасындағы сәйкестік қағидалары, сондай-ақ мәтінді ым тіліне дұрыс түрлендіруге арналған арнайы талдаушы қолданылды. Тұжырымдалған корпус ресурстары шектеулі ым тілдерінде зерттеулер жүргізуге мүмкіндік беретін қазақ ым тілінің бірегей лингвистикалық және құрылымдық ерекшеліктерін көрсетеді. Осы өзгешеліктерді назарға алу үшін қазақ мәтіні мен KRSL белгілері арасындағы сәйкес болу принципін, сондай-ақ мәтінді ым тіліне дәл түрлендіруге арналған арнайы парсер пайданылды (Yerimbetova et al., 2025a). Ұсынылып отырған корпус ресурсы KRSL тілдік және құрылымдық ерекшеліктерін ашады және шектеулі корпустары бар ымдау тілдерін зерттеуге мүмкіндік береді.

Ұсынылып отырған зерттеу аясында қазақ-орыс ым тіліне тән қимыл-қозғалыстардың бейнелік деректерінен, сондай-ақ оларға сәйкес ұсынылған транскрипциялық жазбалардан құралған KRSL-2025 деректер жиынтығы қолданылды. Біз осы деректер жиынының түпнұсқа сипаттамасы зерттеушілердің (Mukushev et al., 2022) зерттеуінде берілген, онда бейнелердің құрамы, аннотация әдістері, қатысушылар және жинаққа енгізілген лексикалық бірлік статистикасы егжей-тегжейлі сипатталған. Авторлардың зерттеуінде FluentSigners-50 деректер жиынын құру мақсатында авторлар ым тілін үздіксіз тану (CSLR) және үздіксіз ым тілін аудару (CSLT) процестерін оңтайландыруға KRSL кескіні деректерінде ауқымды және әтүрлі тотамасын жинақтаған. Осы зерттеу жұмысы аясында 173 түрлі сөйлем пайдаланылып,

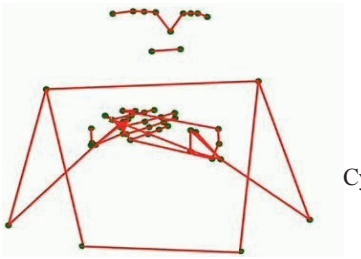
әр сөйлем 50 жеке орындаушы тарапынан бес рет қайталанған. Алынған деректер жиыны өңделмеген шамамен 150 сағаттық бейне кадрлардан құралған және жалпы ұзақтығы шамамен 44 сағатқа тең 43 250 бейнеклипті қамтиды.

Оқу үшін бейнематериалдар. Бейнелер қатысушылардың әртүрлі құрылғылары (смартфондар, веб-камералар) арқылы жазылып, әртүрлі фон, жарық жағдайлары, сондай-ақ түрлі ажыратымдылық, пропорция және кадр жиіліктерін қамтиды. Әрбір бейне сегменті орташа есеппен 2–11 секундқа созылып, шамамен 4 белгіден тұратын толық мәлімдемені көрсетеді.

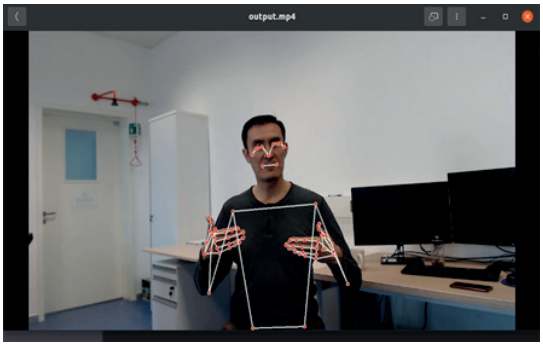
Мәтіннің транскрипциясы. Әрбір base_id үшін <base_id> ішкі жиынынан/транскрипттерінен сөз тіркестерінің толық орысша аудармасы бар мәтіндік файл алынды. Бұл аударманы Child of Deaf Adults (CODA) мұқият орындады, содан кейін 44 ана тілінде сөйлейтіндер растады. Барлық ана тілінде сөйлейтіндер транскрипциялардың сенімділігі мен лингвистикалық дәлдігіне кепілдік беретін келісімге қол қойды.

Ым тіліндегі глосс және метадеректер. Ым тіліндегі глосс пен қатысушылардың демографиялық мәліметтерін (18 ер адам және 32 әйел, 8 жастан 57 жасқа дейін), соның ішінде KRSL-ді бала кезінен үйренген және оны кейінірек меңгерген қатысушыларды ұсынды (Mukushev et al., 2022).

Деректерді түрлендіру. Бейне тізбегінен ақпаратқа бай репрезентация алу үшін екі сатылы өңдеу құбыры қолданылады. Hands and Pose үлгілері негізінде негізгі нүктелердің 3D координаттары шығарылады: әр қол үшін 21 буын және дене үшін 33 буын белгіленеді (3–4-суретте көрсетілген). Осы өңдеу нәтижесінде әр кадр 225 өлшемді векторға кодталып, барлық анықталған негізгі нүктелердің кеңістіктік орналасуын толық сипаттайды.



Сурет 3 – Негізгі нүктелер үшін MediaPipe қолданылған үлгісі



Сурет 4 – Негізгі нүктелер бойынша 3D кеңістік координаттарын алуға MediaPipe пайдаланылған үлгісі

Ең алдымен 50 кадрдан тұратын терезе қалыптастырылады. Бекітілген кадр ұзындығын қолданудың бірнеше артықшылығы бар. Біріншіден, ол пакеттік деректерді тиімді өңдеуге мүмкіндік береді. Екіншіден, кіріс деректер үшін стандартталған форматты қамтамасыз етеді. Сонымен қатар, 225 өлшемді кадр векторы сызықтық қабат арқылы модельге сәйкес 256 өлшемді кеңістікке түрлендіріледі. Түрлендірілген векторлар тізбегі TCN-ге енгізіледі. TCN үш қалдық блок негізінде құрылған, әр блок қабылдау өрісін кеңейтеді. Әр блок екі бірөлшемді конволюциялық қабат, ReLU белсендіру функциясы және оқу барысында тастау (dropout) реттеуін қамтиды. Терең нейрондық желілерді тиімді оқыту үшін қалдық қосылымдарды қолдану маңызды рөл атқарады.

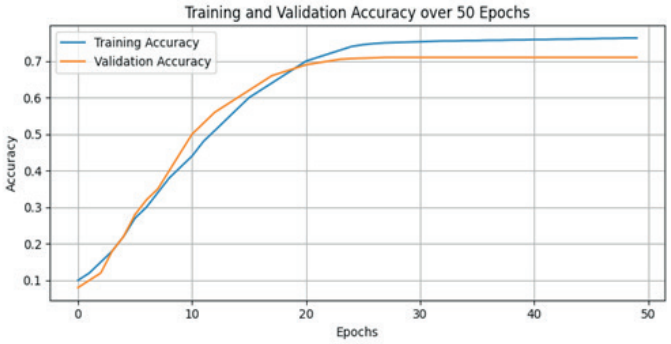
Себеп-салдарлық конволюциялар мен кеңейтілген ядро өлшемін қолдану арқылы TCN параметрлерінің санын едәуір арттырмай, бірнеше кадрдағы қозғалыстың ұзақмерзімді үлгілерін тиімді үйренуге мүмкіндік береді. Осылайша, бейне → функция өңдеу құбыры буын позицияларының күрделі кеңістік сипаттамаларын және қозғалыстың уақыттық динамикасын ескеретін ықшамдалған, бірақ сенімді «жадыны» қалыптастырады.

Нәтижелер. Негізгі оқыту кезеңінде KRSL-2025 деректер жиынының толық көлемі қолданылды. Бұл деректер жинағы 50 қатысушының орындауындағы 173 сөйлемді қамтиды, әрбір сөйлем қатысушылар тарапынан бес рет қайталанған. Оқыту үшін деректердің 80%-ы, ал валидацияға 20%-ы кездейсоқ таңдалып алынды. Деректерді бөлу кезінде дәйектілікті қамтамасыз ету мақсатында сөйлем құрылымы мен орындаушылар сипаттамалары ескерілген. Тәжірибе баптауларында пакет өлшемі 4-ке, уақытша терезе ұзындығы 50 кадрға, ал шығыс тізбектің максималды ұзындығы 20 таңбалауышқа теңестірілді. Модельдің жасырын өлшемділік көрсеткіші (hidden size) 256 мәніне орнатылды.

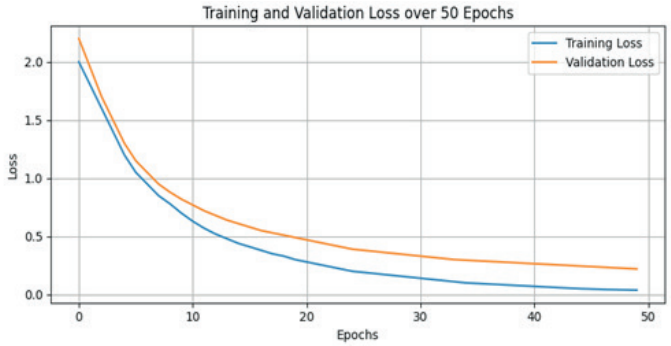
Модель архитектурасы ретінде үш қалдық блокты қамтитын TCN таңдалды. Әр блоктың конфигурациясы: кеңейту коэффициенттері [1, 2, 4], ядро өлшемі 3 және dropout мөлшері 0,1 алынды. Transformer декодерінің архитектурасы келесі параметрлермен сипатталады: 2 қабат, 4 назар аудару басы, 1024 өлшемді алға бағытталған желі (FFN) және 0,1 dropout коэффициенті. Параметрлерді жаңарту үшін 3×10^{-4} оқыту жылдамдығымен жұмыс істейтін AdamW оңтайландырғышы қолданылды. Айқаспалы энтропия негізіндегі шығын функциясы (CrossEntropyLoss) модельді оқыту барысында <pad> белгілерімен толтырылған позицияларды елемеу үшін конфигурациялау жасалынды.

Жаттығудың әрбір кезеңінде декодер болжамдары алынып отырды, кросс-энтропия шығыны есептелді, қателік кері таралып, салмақтар жаңартылды және градиент нормасы тіркелді. Модель салмақтарын оңтайландыруға шығын көрсеткіші кері тарату әдісімен пайдаланылды, градиент нормасы сақталды. Оқыту кезеңі аяқталғаннан кейін валидациялық жиынтықта тестілеу жүргізілді. Бұл әрекет орташа шығын мәнін және токен деңгейіндегі болжам дәлдігін бағалау мақсатында жүзеге асырылды. Оқыту және валидация кезеңдеріндегі шығын мен дәлдік көрсеткіштері, сонымен қатар негізгі өнімділік метрикалары толықтай history_full.json файлына жазылып сақталған. Оқыту процесінің 50 кезеңдік динамикасы 4 және 5-суреттерде ұсынылған. Аталған деректер TCN моделінің оқыту процесіндегі

тұрақтылығын, модельдің шамадан тыс үйрену (overfitting) қаупін және қолданылған архитектуралық шешімдердің тиімділігін бағалауға мүмкіндік берді.



Сурет 5 – Деректер жиынын оқыту және тексеру барысындағы өнімділік көрсеткіштері



Сурет 6 – Деректер жиынын оқыту және тексеру барысындағы қателік мәндері

Модельдің дәлдік көрсеткіштері TCN архитектурасына негізделген KRSL тану жүйесінің өнімділігін бағалау критерийі ретінде пайдаланылды. Оқыту деректерінде модель жаттығу кезеңінен кейін шамамен 75,31% дәлдікке қол жеткізді (толығырақ ақпарат 1-кестеде көрсетілген).

Кесте 1 – Әртүрлі әдістемелердің өнімділік метрикалары бойынша салыстырмалы талдауы

Модель	**Деректер жиыны**	**Метрика (Дәлдік)**	**Шектеулері**
(Li et al., 2020)	WLASL2000	62,63%	Жаңа пайдаланушыларға бейімделу барысында қиындықтар кездеседі.
Біздің ұсынған	KRSL-2025	75,31%	Деректер жиынтығы мен сөздік қордың шектеулілігі

Кестедегі мәліметтерге сүйене отырып, Li et al. (2020) зерттеуінде сипатталған тәсіл WLASL2000 деректер жиынында 62,63% дәлдік көрсеткен. Аталған әдістемелердің негізгі шектеуі ретінде жаңа пайдаланушылар үшін оны меңгерудің күрделілігі аталады. Біздің TCN архитектурасына негізделген

моделіміз KRSL-2025 деректер жиынында 75,31% дәлдік көрсеткішімен өнімділігін көрсетті, бұл әдістемелік тұрғыдан алдыңғы шешімдерден жоғары нәтиже болып есептеледі. Модельдің деректерді тиімді өңдеуі және маңызды үлгілерді анықтау қабілеті жоғары дәлдікпен дәлелденді. Зерттеу барысында деректер жиыны мен сөздік қордың шектеулі екені атап өтілгенімен, модель уақытша тәуелділіктерді тиімді түрде модельдей алады. Алынған нәтижелер кестедегі деректермен салыстырғанда біздің моделіміздің алдыңғы әдістерге қарағанда жоғары өнімділік көрсететінін және KRSL-2025 деректер жиынында айтарлықтай сапалы нәтижелерге қол жеткізетінін растайды.

TCN моделінің KRSL-2025 деректер жиынындағы өнімділік көрсеткіштерін кешенді талдау нәтижелері 2-кестеде ұсынылған. Зерттеу нәтижелері көрсеткендей, көп жағдайда (ID0000, ID0015, ID0035, ID0051) TCN моделінің болжамдары эталондық синтаксиске толық сәйкес келіп, тек мағыналық тұрғыдан бейтарап синонимдік алмастыруларды қамтиды. Алайда, кейбір жағдайларда (ID0037, ID0054, ID0084, ID0099, ID0144) модель шығаратын сөз тізбегінде өзгерістер немесе лексикалық алмастырулар байқалады, бұл сөйлем құрылымының трансформацияланғанын көрсетеді. Ал бастапқы мағына ішінара өзгерген жағдай (ID0062) негізгі лексикалық элементтердің семантикалық баламалармен ауыстырылуынан туындаған. Дегенмен, модель лексикалық және синтаксистік вариацияларға қарамастан негізгі ақпаратты сақтап, дәл жеткізе алу қабілетін жалпы нәтиже көрсеткіштері дәлелдейді.

Кесте 2 – TCN-Transformer моделінің KRSL-2025 деректеріндегі болжам үлгілері

№	ID0000	ID0015	ID0035	ID0037	ID0051	ID0062	ID0054	ID0084	ID0099	ID0144
Түпнұсқа (эталондық) мәтін	Здравствуйте	Что случай	Вызов врач	Помогите боль голова сердце и трудно воздух	Спасибо большое	Мой возраст	До следующей встречи	Я рад тебя	Где ты работа	Я хочу практика мой жестовый язык
Болжамның үлгілері	Здравствуйте	Что происходит	Вызов доктора	Помогите боль голова и сердце трудно дышать	Большое спасибо	Мой год рождения	До скорой встречи	Я очень рад	Где ты работаешь	Я хочу практиковать язык жестов

Ұсынылған жүйеде MediaPipe пен TCN үлгілерінің үйлесімді қолданылуы деректерді өңдеу жылдамдығын секундына 40 кадр (40 FPS) өңдеу жылдамдығын қамтамасыз етілді. Өнімділікті бағалау кезінде RTX 3090 GPU пайдаланып, әр операцияның орташа кідіріс уақыты 20–25 миллисекунд аралығында тіркелді.

Талқылау. TCN моделінің ым тілін мәтінге аудару үшін күрделі жүйелерді әзірлеудегі практикалық маңызы алынған нәтижелер арқылы дәлелденді. Зерттеу барысында ресурстары шектеулі ым-ишара тілдері үшін MediaPipe

арқылы алынған кілттік нүктелерді TCN архитектурасының мүмкіндіктерімен үйлестіретін жаңа әдістеме ұсынылды.

Осы зерттеудегі KRSL-2025 деректер жиыны 173 сөйлемнен тұратындығынан, оның сөздік қорының шектеулілігі модельдің тану сапасына әсерін тигізді. Бұл мәселені шешу мақсатында болашақ зерттеулерде деректер көлемін ұлғайту және лексикалық құрамды байыту жоспарланып отыр. Модельдің тану қабілетін арттыру және сөздік қордың көптүрлілігін қамтамасыз ету үшін деректер корпусын кеңейту қажет деп саналады. Сонымен қатар, модельдің тұрақтылығын және тану сапасын жетілдіру үшін арнайы оқыту стратегиялары, ерте тоқтату механизмдері және деректерді бөлуге арналған жаңа тәсілдер енгізілетін болады.

Қорытынды. Осы жұмыста TCN архитектурасына негізделген ым тілін тану және оны мәтінге аудару моделі құрылды. KRSL-2025 деректер жиынындағы 50 қатысушы орындаған 173 сөйлемнің барлық жазба үлгілері модельді оқыту кезінде толық көлемде пайдаланылды. Деректер жиыны әртүрлі орындаушыларды қамти отырып, репрезентативтілікті қамтамасыз етті және ым-ишара тілінің көпқырлы сипаттамасын ұсынды. Модельдің тану сапасын бағалау мақсатында деректер 80/20 қатынасында оқыту және валидация жиындарына бөлінді, бұл сөйлемдердің синтаксистік құрылымының иерархиялық үлестірімін сақтауға мүмкіндік берді. Эксперимент нәтижелері TCN моделінің ым тілінің уақыттық құрылымдарын өңдеудегі тиімділігін ұсынып, модель оқыту жиынында 75,31% дәлдікке қол жеткізді. Алынған нәтижелер ұсынылған тәсілдің тиімділігін дәлелдеп, TCN моделінің есептеу қуаты шектеулі және нақты уақыт режимінде жұмыс істеуді қажет ететін жүйелерде ым тілін тану салаларында кеңінен қолданылу мүмкіндігін көрсетеді. Болашақ зерттеулер көрнекі-кеңістіктік ақпаратты кодтау архитектураларын зерттеуге, мультимодальды деректерді (мысалы, бейне және аудио ағындарын) интеграциялауға және ауқымды деректер жиынтықтарында алдын ала оқыту әдістерін қолдануға басымдық берілуі мүмкін.

References

Sharma A., Guo D., Parmar A., Ge J., & Li H. (2024) Promoting Sign Language Awareness: A Deep Learning Web Application for Sign Language Recognition. In Proceedings of the 2024 8th International Conference on Deep Learning Technologies (ICDLT). — P. 22-28. https://doi.org/10.1145/3695719.369572 (in Eng.)

Mejía-Peréz K., Córdova-Esparza D.M., Terven J., Herrera-Navarro A.M., García-Ramírez T., & Ramírez-Pedraza A. (2022) Automatic recognition of Mexican sign language using a depth camera and recurrent neural networks. Applied Sciences, 12(11), 5523. https://doi.org/10.3390/app12115523 (in Eng.)

Bedaso M., Hussein A., & Diriba C. (2024) Developing Sign Language Recognition Model for Afaan Oromoo Words Using a Deep Learning Techniques. https://doi.org/10.21203/rs.3.rs-4218445/v1 (in Eng.)

Shetty S., et al. (2024). Gesture-to-Text: A Real-Time Indian Sign Language Translator with Pose Estimation and LSTMs. Procedia Computer Science. 235. — P. 2684–2692. https://doi.org/10.1016/j.procs.2024.04.253 (in Eng.)

Tunga A., Nuthalapati S.V., & Wachs J. (2021) Pose-based sign language recognition using GCN

and BERT. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 31-40). https://doi.org/10.48550/arXiv.2012.00781 (in Eng.)

Hosain A.A., Santhalingam P.S., Pathak P., Rangwala H., Kosecka J. (2021) Hand Pose Guided 3D Pooling for Word-Level Sign Language Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. — P. 3429–3439. https://doi.org/10.1109/WACV48630.2021.00347 (in Eng.)

Sumon M.S.I., Ali M.B., Bari S., Ohi I.R., Islam M., & Rahman S.M. (2024) Sign Language Word Detection Using LRCN. In IOP Conference Series: Materials Science and Engineering. — Vol. 1305, No.1. — P. 012023. IOP Publishing. doi:10.1088/1757-899X/1305/1/012023 (in Eng.)

Tur A.O., & Keles H.Y. (2021) Evaluation of hidden markov models using deep cnn features in isolated sign recognition. Multimedia tools and applications, 80(13). — P. 19137-19155. https://doi.org/10.1007/s11042-021-10593-w (in Eng.)

Xu X., Meng K., Chen C., & Lu L. (2023) Isolated Word Sign Language Recognition Based on Improved SKResNet-TCN Network. Journal of Sensors, 2023(1), 9503961. https://doi.org/10.1155/2023/9503961 (in Eng.)

Sincan O.M., Junior J., Jacques C.S., Escalera S., & Keles H.Y. (2021) Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — P. 3472-3481. DOI: 10.1109/CVPRW53098.2021.00386 (in Eng.)

Özdemir O., Kindiroglu A.A., Camgöz N.C., Akarun L. (2020) BosphorusSign22k Sign Language Recognition Dataset. arXiv preprint arXiv:2004.01283. https://doi.org/10.48550/arXiv.2004.01283 (in Eng.)

Chen, Y., Li, J., Lin, S., Xu, Y., & Yang, C. (2024). A BiLSTM and CTC based Multi-Sensor information fusion frame for continuous sign language recognition. In 2024 10th International Conference on Electrical Engineering, Control and Robotics (EECR). — P. 310-315. IEEE. DOI: 10.1109/EECR60807.2024.10607314 (in Eng.)

Kapitanov A., Karina K., Nagaev A., & Elizaveta P. (2023) Slovo: Russian sign language dataset. In International Conference on Computer Vision Systems (pp. 63-73). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44137-0_6 (in Eng.)

Buribayev Z., Aouani M., Zhangabay Z., Yerkos A., Abdirazak Z., Zhassuzak M. (2025) Enhancing Kazakh Sign Language Recognition with BiLSTM Using YOLO Keypoints and Optical Flow. Applied Sciences, 15(10), 5685. https://doi.org/10.3390/app15105685 (in Eng.)

Yerimbetova A., Sakenov B., Sambetbayeva M., Daiyrbayeva E., Berzhanova U., & Othman M. (2025) Creating a Parallel Corpus for the Kazakh Sign Language and Learning. Applied Sciences, 15(5), 2808. https://doi.org/10.3390/app15052808 (in Eng.)

Yerimbetova A., Sambetbayeva M., Daiyrbayeva E., Sakenov B., Berzhanova U. (2025). Teren oqytu adisin qoldanu arqyly qazaq ım tilin tanuga arnalgan model qury [Creating a Model for Recognizing the Kazakh Sign Language Using the Deep Learning Method]. Academic Scientific Journal of Computer Science. No.1. — P. 108–123. https://doi.org/10.32014/2025.2518-1726.328 (in Kaz.)

Lea C., Vidal R., Reiter A., & Hager G.D. (2016) Temporal convolutional networks: A unified approach to action segmentation. In European conference on computer vision (pp. 47-54). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-49409-8_7 (in Eng.)

Sharma V., Gupta A.K., Sharma A., et al. (2025) A Unified Approach for Continuous Sign Language Recognition and Translation. International Journal of Data Science and Analytics. 20. — P. 1547–1561. https://doi.org/10.1007/s41060-024-00549-2 (in Eng.)

Qiu H., Niu Y., Shang J., Gao L., & Xu D. (2023) A piecewise method for bearing remaining useful life estimation using temporal convolutional networks. Journal of Manufacturing Systems, 68. — P. 227-241. https://doi.org/10.1016/j.jmsy.2023.04.002 (in Eng.)

Mukushev M., Ubingazhibov A., Kydyrbekova A., Imashev A., Kimmelman V., & Sandygulova A. (2022) FluentSigners-50: A signer independent benchmark dataset for sign language processing. Plos one, 17(9), https://doi.org/10.1371/journal.pone.0273649 (in Eng.)

Li D., Rodriguez C., Yu X., & Li H. (2020) Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF winter conference on applications of computer vision. — P. 1459-1469. doi: 10.1109/WACV45572.2020.9093512 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 461–475

https://doi.org/10.32014/2026.2518-1726.423

IRSTI 50.41.25
UDC 004.942

Zhukabayeva T.K.[1], Benkhelifa E.[2], Mardenov Y.M.[1], Baumuratova D.[1], Karabayev N.[1], 2026.
[1]L.N. Gumilyov Eurasian National University, Kazakhstan;
[2]Staffordshire University, UK.
E-mail: emardenov@gmail.com

DECISION SUPPORT FOR RESPONDING TO ATTACKS IN CYBER-PHYSICAL INDUSTRIAL INTERNET-OF-THINGS SYSTEMS

Zhukabayeva Tamara — Professor, Faculty of Information Technology, L.N. Gumilyov Eurasian National University, Kazakhstan,
E-mail: t.zhukabayeva@astanait.edu.kz, https://orcid.org/0000-0001-6345-5211;
Benkhelifa Elhadj — Professor, Digital, Tech, Innovation & Business, Staffordshire University, UK,
E-mail: e.benkhelifa@staffs.ac.uk, https://orcid.org/0000-0001-6168-2664;
Mardenov Yerik — Researcher, L.N. Gumilyov Eurasian National University, Astana International University, Kazakhstan,
E-mail: emardenov@gmail.com and https://orcid.org/0000-0001-9284-9797;
Baumuratova Dilaram — Researcher, PhD, L.N. Gumilyov Eurasian National University, Kazakhstan,
E-mail: baumuratova.d@gmail.com and https://orcid.org/0000-0001-6429-6435;
Karabayev Nurdaulet — Researcher, L.N. Gumilyov Eurasian National University, Kazakhstan,
E-mail: 020419501012@enu.kz and https://orcid.org/0000-0001-9284-9797.

Abstract. Industrial Internet-of-Things (IIoT) systems have emerged as a foundational technology underpinning a wide range of modern applications, including environmental monitoring, healthcare systems, industrial automation, and smart infrastructures. Their inherent advantages incorporate scalability, low energy consumption, intelligence and programmability, rapid data acquisition, reliability, low cost, and the absence of maintenance requirements. At the same time, they attract the attention of attackers, making the task of ensuring the security of such networks crucial. This study introduces a comprehensive decision support methodology for responding to cyberattacks in IIoT systems, including systems using edge computing principles, based on analytical modeling for the analysis of networks, attacks, and countermeasures. The methodology also employs machine learning, rules, and multi-criteria optimization for attack detection and countermeasure selection. It consists of three stages: data collection

on events occurring in the system, detection of attacks and anomalies, and selection of countermeasures for the identified attacks. Within the framework of the methodology, models of networks, attacks, and countermeasures, specific for IIoT systems, are presented. The main components and topology of IIoT systems are considered within the network model. The main types of attacks and possible countermeasures for IIoT systems are identified to specify the models of attacks and countermeasures. The algorithm for countermeasure selection within the proposed methodology based on rules and multi-criteria optimization and using the specified models is described. The application of the methodology is demonstrated through an example. The advantages and limitations of the proposed methodology are analyzed, and directions for future research are outlined.

Key words: IIoT systems, cyberattacks, countermeasures, decision support, response

For citations: Zhukabayeva T.K., Benkhelifa E., Mardenov Y.M., Baumuratova D., Karabayev N. Decision support for responding to attacks in cyber-physical industrial internet-of-things systems. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 461–475. DOI: https://doi.org/10.32014/2026.2518-1726.423

Financing. *This work by the staff of L.N. Gumilyov Eurasian National University is carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23489127).*

Жукабаева Т.К.[1], Бенхелифа Э.[2], Марденов Е.М.[1], Баумуратова Д.[1], Карабаев Н.[1], 2026.

[1]Л.Н. Гумилёв атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]Стаффордшир университеті, Ұлыбритания.
E-mail: emardenov@gmail.com

КИБЕРФИЗИКАЛЫҚ ӨНЕРКӘСІПТІК ИНТЕРНЕТ ЗАТТАРЫ ЖҮЙЕЛЕРІНДЕГІ ШАБУЫЛДАРҒА ӘРЕКЕТ ЕТУ КЕЗІНДЕ ШЕШІМ ҚАБЫЛДАУДЫ ҚОЛДАУ

Жукабаева Тамара — профессор, ақпараттық технологиялар факультеті, Л.Н. Гумилёв атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail**:** t.zhukabayeva@astanait.edu, https://orcid.org/0000-0001-6345-5211;
Бенхелифа Эльхадж — профессор, Digital, Tech, Innovation & Business мектебі, Стаффордшир университеті, Ұлыбритания,
E-mail: e.benkhelifa@staffs.ac.uk, https://orcid.org/0000-0001-6168-2664;
Ерик Марденов — зерттеуші, Л.Н. Гумилёв атындағы Еуразия ұлттық университеті, Астана халықаралық университеті, Астана, Қазақстан,
E-mail: emardenov@gmail.com, https://orcid.org/0000-0001-9284-9797;

Диларам Баумуратова —PhD, Л.Н. Гумилёв атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: baumuratova.d@gmail.com, https://orcid.org/0000-0001-6429-6435;
Нурдаулет Карабаев – зерттеуші, Л. Н. Гумилёв атындағы Еуразия ұлттық университеті, Астана, Казахстан,
Электронная почта: E-mail: 020419501012@enu.kz and https://orcid.org/0000-0001-9284-9797.

Аннотация. Өнеркәсіптік интернет заттары жүйелері (Industrial Internet of Things, IIoT) қоршаған ортаны мониторингтеу, денсаулық сақтау жүйелері, өнеркәсіптік автоматтандыру және ақылды инфрақұрылымдар сияқты заманауи қолданбалардың кең ауқымының негізін құрайтын базалық технологияға айналды. Олардың негізгі артықшылықтарына масштабталуы, төмен энергия тұтынуы, интеллектуалдылығы мен бағдарламалануы, деректерді жинаудың жоғары жылдамдығы, сенімділігі, төмен құны, сондай-ақ техникалық қызмет көрсетуді қажет етпеуі жатады. Сонымен қатар, мұндай жүйелер қаскөйлердің назарын өзіне аударады, бұл олардың қауіпсіздігін қамтамасыз ету мәселесінің өзектілігін арттырады. Осы зерттеуде IIoT жүйелеріндегі кибершабуылдарға әрекет ету кезінде шешім қабылдауды қолдауға арналған кешенді әдістеме ұсынылады, оның ішінде шеткі есептеу (edge computing) қағидаттарын қолданатын жүйелер де қамтылған. Әдістеме желілерді, шабуылдарды және қарсы шараларды талдауға арналған аналитикалық модельдеуге негізделген, сондай-ақ шабуылдарды анықтау және қарсы шараларды таңдау үшін машиналық оқыту әдістерін, ережелерді және көпкритерийлі оңтайландыруды пайдаланады. Ұсынылған әдістеме үш кезеңнен тұрады: жүйеде орын алатын оқиғалар туралы деректерді жинау; шабуылдар мен аномалияларды анықтау; анықталған шабуылдар үшін қарсы шараларды таңдау. Әдістеме аясында IIoT жүйелеріне тән желілердің, шабуылдардың және қарсы шаралардың модельдері ұсынылған. Желілік модельде IIoT жүйелерінің негізгі компоненттері мен топологиясы қарастырылады. Шабуылдар мен қарсы шаралар модельдерін нақтылау үшін IIoT ортасына тән негізгі шабуыл түрлері мен ықтимал қарсы әрекеттер айқындалған. Әзірленген модельдерді қолдана отырып, ережелер мен көпкритерийлі оңтайландыруға негізделген қарсы шараларды таңдау алгоритмі сипатталған. Әдістеменің қолданылуы практикалық мысал арқылы көрсетілген. Ұсынылған тәсілдің артықшылықтары мен шектеулері талданып, болашақ зерттеулердің негізгі бағыттары айқындалған.

Түйін сөздер: IIoT жүйелері, кибершабуылдар, қарсы шаралар, шешім қабылдауды қолдау, әрекет ету

Жукабаева Т.К.[1], Бенхелифа Э.[2], Марденов Е.М.[1], Баумуратова Д.[1], Карабаев Н.[1], 2026.
[1]Евразийский национальный университет имени Л.Н. Гумилёва,
Астана, Казахстан;
[2]Стаффордширский университет, Великобритания.
E-mail: emardenov@gmail.com

ПОДДЕРЖКА ПРИНЯТИЯ РЕШЕНИЙ ПРИ РЕАГИРОВАНИИ НА АТАКИ В КИБЕРФИЗИЧЕСКИХ ПРОМЫШЛЕННЫХ СИСТЕМАХ ИНТЕРНЕТА ВЕЩЕЙ

Жукабаева Тамара — профессор, факультет информационных технологий, Евразийский национальный университет имени Л. Н. Гумилёва, Казахстан,
E-mail**:** t.zhukabayeva@astanait.edu, https://orcid.org/0000-0001-6345-5211;
Эльхадж Бенхелифа — профессор, Digital, Tech, Innovation & Business, Стаффордширский университет, Великобритания
E-mail: e.benkhelifa@staffs.ac.uk, https://orcid.org/0000-0001-6168-2664;
Ерик Марденов — исследователь, Евразийский национальный университет имени Л.Н. Гумилёва, Международный университет Астана, Казахстан,
E-mail: emardenov@gmail.com, https://orcid.org/0000-0001-9284-9797;
Диларам Баумуратова — исследователь, PhD, Евразийский национальный университет имени Л.Н. Гумилёва, Казахстан
E-mail: baumuratova.d@gmail.com, https://orcid.org/0000-0001-6429-6435;
Нурдаулет Карабаев — исследователь, Евразийский национальный университет имени Л.Н. Гумилёва, Казахстан,
Электронная почта: E-mail: 020419501012@enu.kz and https://orcid.org/0000-0001-9284-9797.

Аннотация. Промышленные системы интернета вещей (Industrial Internet of Things, IIoT) являются базовой технологией, лежащей в основе широкого спектра современных приложений, включая промышленную автоматизацию, системы мониторинга, здравоохранение и интеллектуальную инфраструктуру. Ключевыми преимуществами IIoT являются масштабируемость, энергоэффективность, высокая скорость обработки данных, программируемость и надёжность. Вместе с тем такие системы представляют собой привлекательную цель для кибератак, что обусловливает необходимость разработки эффективных методов обеспечения их безопасности. В исследовании предложена комплексная методология поддержки принятия решений при реагировании на кибератаки в системах IIoT, включая среды с использованием пограничных вычислений (edge computing). Методология основана на аналитическом моделировании сетей, атак и контрмер, а также включает методы машинного обучения, экспертные правила и подходы многокритериальной оптимизации для обнаружения атак и выбора оптимальных мер реагирования. Предложенная методология включает три основных этапа: сбор и анализ данных о событиях в системе, обнаружение атак и аномалий, а также выбор контрмер для нейтрализации выявленных угроз. В рамках исследования разработаны модели сетевой

архитектуры, типов атак и соответствующих контрмер, характерных для IIoT-среды. Описан алгоритм выбора контрмер, основанный на правилах и многокритериальной оптимизации с использованием предложенных моделей. Практическое применение методологии продемонстрировано на прикладном примере. Проведён анализ преимуществ и ограничений предложенного подхода, а также определены направления для дальнейших исследований.

Ключевые слова: системы IIoT, кибератаки, контрмеры, поддержка принятия решений, реагирование

Introduction and Related Work. IIoT systems are spatially distributed sensor nodes capable of collecting data, processing it, and interacting via wireless channels (Tossa et al., 2025). At present, they are applied in smart home management, transportation and logistics, industry, agriculture, environmental monitoring, urban surveillance, entertainment, healthcare, and energy, as well as in security and battlefield monitoring (Kardi et al., 2019).

«Ideal» IIoT systems are scalable, consume little energy, are intelligent and programmable, provide rapid data acquisition, are reliable, low-cost, and require no maintenance. However, they also attract the attention of attackers. In particular, passive attacks those aimed at stealing information on IIoT systems are discussed in (Keerthika et al., 2021) (eavesdropping and traffic analysis). Active attacks those aimed at damaging the network by altering or replacing data are considered in (Verma et al., 2022) (denial of service), (Sánchez et al., 2021) (masquerade attack), (Hu et al., 2022) (replay attack), (Hu et al., 2014) (selective forwarding attack), (Sujihelen et al., 2022) (node replication), (Al-Ahmadi et al., 2022) (wormhole attack), (Faris et al., 2023) (Sybil attack), (Teng et al., 2023) (sinkhole attack), (Allimuthu et al., 2022) (rushing attack), and (Beddoe et al., 2013) (message modification). The types of attacks on IIoT systems are further classified in (Keerthika et al., 2021) and (Verma et al., 2020).

One of the key challenges in deploying IIoT systems is ensuring cybersecurity (Jadhav et al., 2017). Possible countermeasures are outlined in (Turakulovich et al., 2019) and (Mamdouh et al., 2018). Among them are security measures that should be incorporated during the system design stage, such as the use of secure data transmission protocols (Turakulovich et al., 2019) (such measures for all identified attacks are discussed in (Keerthika et al., 2021)). However, these alone cannot completely eliminate the possibility of a successful attack. Therefore, it is also necessary to integrate protection mechanisms into the system that enable the detection of cyberattacks and the implementation of countermeasures, for example, through the use of machine learning (Mamdouh et al., 2018). This study focuses on decision support issues based on such protection mechanisms.

This paper proposes a decision support methodology for responding to cyberattacks in IIoT systems, based on analytical modeling for simulating the system, attacks, and countermeasures; machine learning for attack detection; rules

and multi-criteria optimization for countermeasure selection. The application of the methodology is demonstrated through an example. The advantages and limitations of the proposed methodology are analyzed, and directions for future research are outlined. Thus, the main contributions of this research are as follows: (1) a decision support methodology for responding to cyberattacks in IIoT systems, based on analytical modeling, machine learning, and rules and multi-criteria optimization; (2) analytical models of system, attacks, and countermeasures considering specificity of IIoT systems, and considering the types of attacks and appropriate countermeasures; (3) algorithm for countermeasure selection based on rules and multi-criteria optimization and using the specified models; (4) demonstration of applicability of the proposed methodology on the toy example.

The paper is organized as follows. Relevant studies are briefly reviewed in the introduction. Section 2 presents the key definitions and problem statement. Section 3 describes the proposed decision support methodology for responding to cyberattacks in IIoT systems. Section 4 provides an example of the methodology's application and a discussion.

Materials and Methods

IIoT systems. A IIoT system can be formalized as follows:

$$IS = <N, C>, \quad (1)$$

where

N – nodes,

C – wireless links between them.

A node consists of core and additional components (Tossa et al., 2025). The core components include:

Data acquisition component (consisting of a physical sensing device that collects data from the local environment and an Analog-to-Digital Converter, ADC) collects analog data from the local environment and converts it into a digital format understandable by the processor,

Data processing component, including ones using edge computing (a processor complemented with Random Access Memory) manages the operation of other components and analyzes/processes the data obtained from the data acquisition component,

Communication component (which can be optical or radio-frequency) transmits and receives data in the wireless medium,

Power supply component (a small, limited-capacity battery).

The additional components may include GPS (Global Positioning System), a mobility system, and an energy generator.

According to (Tossa et al., 2025), depending on the application domain and architecture, the following types of nodes are distinguished:

Standard node includes a data transmission component and a data processing component;

Sensor node (source) a standard node with a data acquisition or detection component;

Actuator node a standard node with a component for performing specific mechanical tasks;

Gateway a standard node that relays traffic into the network;

Base station a standard node with a serial converter connected to a second communication component for relaying data from nodes to users or other networks

Various architectures of IIoT systems are distinguished, including:

Flat architecture all nodes have equal energy, computing resources, and memory, perform identical roles, and each node can interact with any other,

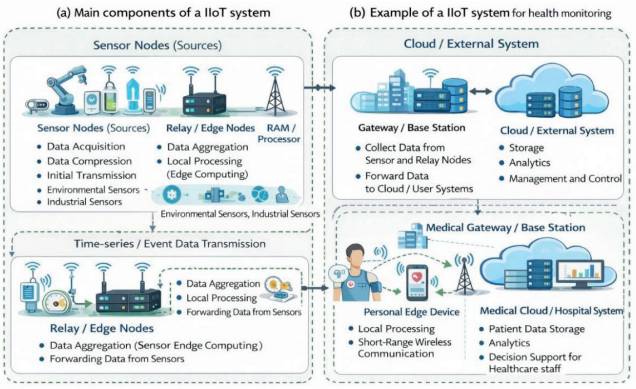
Hierarchical architecture nodes possess different resources and perform different roles, and Multitier architecture a hierarchical architecture in which nodes are grouped into clusters according to their functionality.

Depending on the architecture, the role of each network node is defined: source, relay, or data collection node. Typically, a IIoT system consists of a base station (data collection node) interacting with multiple sensors (sources and relays) (hierarchical architecture). A sensor (source) collects data, compresses it, and transmits it to the base station (data collection node) either directly or through other sensors (relays). The base station forwards the data to the system. An example of such an architecture is shown in Figure 1 (a - main components of a IIoT system and the connections between them, b - an example of a IIoT system for health monitoring).

The following features of IIoT systems are significant for developing a decision support methodology for responding to cyberattacks in such systems:

Heterogeneity in terms of energy sources, data processing capabilities, communication ranges, and sensing modalities, which complicates the deployment process of IIoT systems;

Constraints related to network lifetime, coverage, energy/memory/ computational resources, and data transmission time.

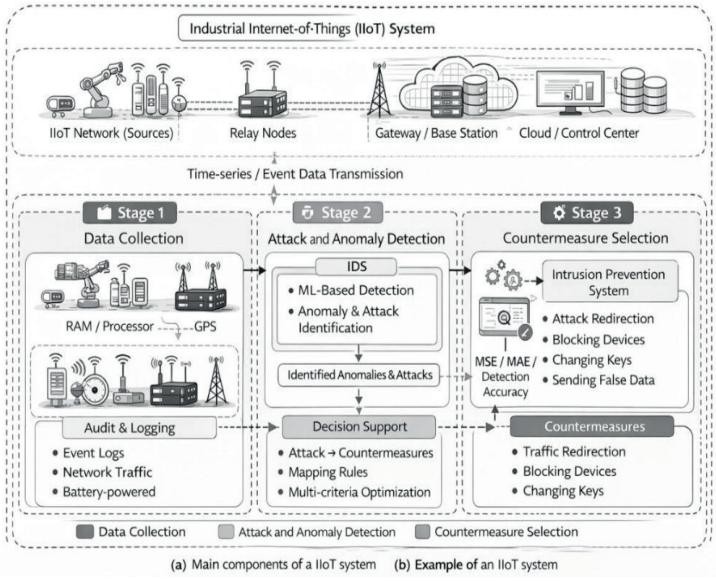


a) Main components of a IIoT system b) Example of a IIoT system for health monitoring

Figure 1 – Hierarchical architecture of a IIoT system.

Problem Statement of the Study. Given: a IIoT system (IS), a set of events – E, a set of attacks on the IS – A, a set of security mechanisms – SMP, and a set of countermeasures – CM. It is necessary to develop a decision support methodology for responding to cyberattacks in IIoT systems that enables the mapping of the set of attacks A onto the set of countermeasures CM: F: A → CM, taking into account resource and privilege constraints.

Decision Support for Responding to Cyberattacks in IIoT systems. The proposed decision support methodology for responding to cyberattacks in IIoT systems is based on analytical modeling for simulating the system, attacks, and countermeasures; machine learning for attack detection; rules; and multi-criteria optimization for countermeasure selection. The methodology consists of three stages: data collection on events occurring in the system, detection of attacks and anomalies, and selection of countermeasures for the identified attacks (Figure 2).



a) Main components of a IIoT system b) Example of an IIoT system

Figure 2 – Decision support methodology for responding to cyberattacks.

For data collection, it is proposed to implement auditing and logging; for detecting attacks and anomalies an intrusion detection system based on machine learning; and for countermeasure selection an intrusion prevention system based on rules and multi-criteria optimization.

Models. First, we define a set of models based on the relationships between the objects and subjects involved in the decision-making process, including network objects (nodes and links between them), their vulnerabilities and weaknesses, known exploits and attacks, incidents, as well as countermeasures (Figure 3). In the event of a security incident, protection mechanisms which may be network objects

or components of such objects are employed to implement countermeasures (for example, a firewall can be used to block a connection).

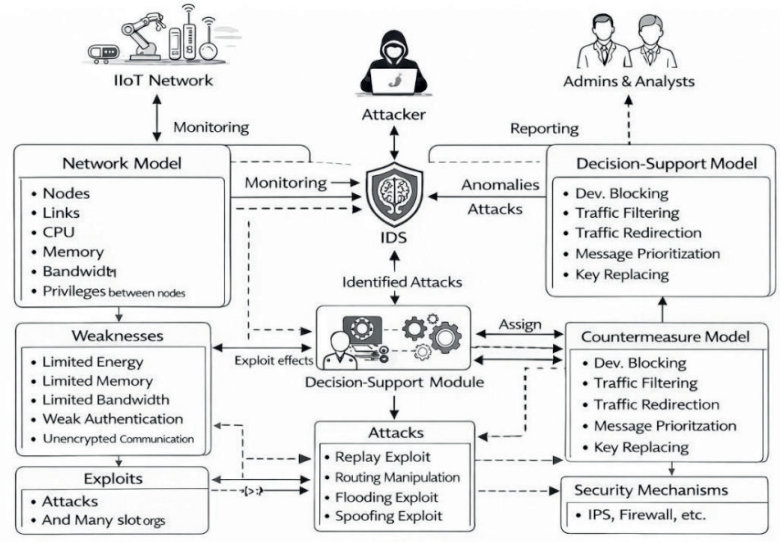


Figure 3 – Relationships between objects and subjects involved in the decision-making process for response.

Thus, it is necessary to extend the set of possible node roles in the network with security mechanisms, such as intrusion detection and prevention systems (Figure 4).

To account for the possible roles of network nodes as well as embedded security mechanisms, we define the elements of the network model from Equation 1 ($IS = <N, C>$) as follows:

N – network nodes, $N = \{\cup_i n_i\}$, where i – number of system node, $n_i = <name, role, SM>$, where name is the node identifier, role is the node's role determining available resources (energy, memory, computing capacity) and privileges (user/administrator/owner), and SM denotes the security mechanisms available for this node, $SM = \{\cup_i sm_i\}$, where i – number of security mechanism, $sm_i = <name, CM>$, where CM – the set of countermeasures related to the security mechanism sm_i;

C – wireless and wired links between them, $C = <prot, trCh>$, where *prot* is the transmission protocol, and *trCh* is the bandwidth.

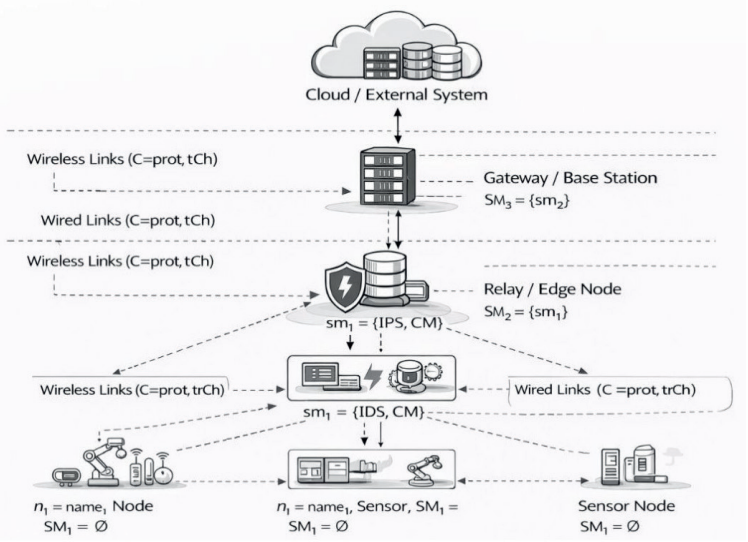


Figure 4 – Hierarchical architecture of a IIoT system with security mechanisms.

To identify possible security mechanisms and countermeasures, it is necessary to define the potential attacks. In (Shahzad et al., 2017), the authors highlight the following threats to IIoT systems and their corresponding attacks:

Confidentiality threat (unauthorized access to information; in IIoT system terms a message transmitted over the wireless network can be read and understood by an attacker): Eavesdropping (information gathering from carriers, wiretapping, interception of electromagnetic/radio frequency signals, traffic analysis, staff negligence), Interaction (data theft, replay attack, masquerade attack, integrity violation, man-in-the-middle attack, bypassing security mechanisms, physical intrusion, authorization violation), System compromise (trapdoor attack, Trojan attack, service spoofing, viruses/worms).

Integrity threat (unauthorized modification or theft of information; in IIoT system terms a message transmitted over the network is altered or forged): Modification (denial, interception/modification), Interaction (see above), System compromise (see above).

Availability threat (denial of service or prevention of authorized access; in IIoT system terms – inability to transmit a message or for a node to use resources): Denial of Service (integrity violation, resource exhaustion), Interaction (see above), System compromise (see above).

Non-repudiation threat (denial of an action that has occurred or claiming that an action has not occurred; in IIoT system terms providing information to an unauthorized/unauthenticated or fake sensor): Reposting, Theft/modification, Interaction (see above), System compromise (see above).

In (Shahzad et al., 2017), the authors highlight the following possible security measures and countermeasures according to the types of threats:

Confidentiality (counteracting unauthorized access to information): passwords and certificates (key and certificate management must be ensured), firewalls with access control lists, auditing and logging, intrusion detection systems, antivirus/spyware software.

Integrity (counteracting unauthorized modification or theft of information): passwords and certificates (key and certificate management must be ensured; compared to confidentiality measures, digital signatures and checksums are additionally introduced), firewalls with access control lists, auditing and logging, intrusion detection systems, antivirus/spyware software.

Availability (counteracting denial of service or prevention of authorized access): backup during and recovery after an attack, security and vulnerability incident reporting, bandwidth limitation, access management, firewalls with access control lists, auditing and logging, backups and recovery, intrusion detection systems, antivirus/spyware software, network and system management.

Non-repudiation (counteracting denial that an action took place or claims that it did not): passwords and certificates (key and certificate management must be ensured), auditing and logging (for legal proceedings).

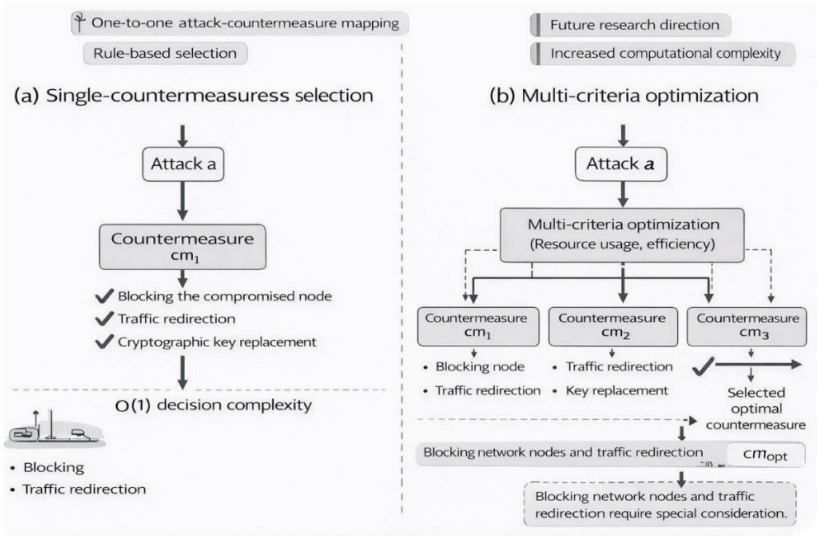


Figure 5 – Comparison of single-countermeasure selection and multi-criteria optimization in the proposed decision-support methodology.

The security mechanisms applied during the operational phase include:

• Firewalls with access control lists (measures: blocking access to a specific node, blacklisting, message prioritization, bandwidth limitation, and traffic redirection),

• Audit and logging (a passive measure, i.e., used for data collection but not involved in supporting response decision-making),

• Intrusion detection systems (a passive measure, i.e., used for detecting attacks but not involved in supporting response decision-making, performing monitoring to identify attacks and anomalies),

• Antivirus/anti-spyware software (blocking malicious software),

• Network and system management (access control, bandwidth limitation).

Within the intrusion prevention subsystem of the security system (Figure 4), it is proposed to implement the selection of countermeasures for different types of attacks based on rules, using Table 1 (Keerthika et al., 2021).

Table 1 – Countermeasures in IIoT systems

Attack	**Threat to**	**Countermeasures**
Denial of Service (Verma et al., 2022)	availability	message prioritization
Jamming	availability	Spread spectrum, region mapping and traffic redirection
Physical Attack	availability	traffic redirection
Spoofing, Routing Information Forgery	availability, integrity, confidentiality, non-repudiation	Block compromised node and reconnect, change key
Selective Forwarding (Hu et al., 2014)	availability	traffic redirection
Sink Hole Attack (Teng et al., 2023)	availability	Block compromised node and redirect traffic
Black Hole Attack	availability, integrity, confidentiality, non-repudiation	traffic redirection
Sybil Attack (Faris et al., 2023)	availability, integrity, confidentiality, non-repudiation	Block nodes and redirect traffic
Wormhole Attack (Al-Ahmadi et al., 2022)	availability, integrity, confidentiality, non-repudiation	traffic redirection
HELLO Flood Attack	availability	message prioritization, traffic redirection
Acknowledgment Spoofing (Verma et al., 2020)	integrity, confidentiality, non-repudiation	Block fake node, reconnect with legitimate node, change key
Neglect and Greed Attack (Verma et al., 2020)	availability	Recover message transmission algorithm
Homing Attack (Verma et al., 2020)	confidentiality	Collect information, encryption
Node Takeover Attack	availability, integrity, confidentiality, non-repudiation	Block fake node, reconnect with legitimate node, change key
Node Failure Attack	availability	traffic redirection, restore node
Monitoring and Eavesdropping (Keerthika et al., 2021)	confidentiality	Collect information, encryption
Traffic Analysis (Keerthika et al., 2021)	confidentiality	Collect information, encryption
Attacker Masking	availability, integrity, confidentiality, non-repudiation	Collect information, encryption, restore message transmission algorithm

Node Replication (Sujihelen et al., 2022)	integrity	Block fake node, reconnect with legitimate node, change key
False Knot	availability	Block fake node, reconnect with legitimate node, change key
Replay Attack (Hu et al., 2022)	confidentiality, non-repudiation	Blocking fake node, reconnect with legitimate node, change key
Rushing Attack (Allimuthu et al., 2022)	availability	Blocking fake node, restoring message transfer algorithm, change key
Vampire Attack	Threat to	Replacing nodes, redirecting traffic

Considering Table 1, let us introduce two additional models: the attack model and the countermeasure model. The attack model is defined as:

$A = \{\cup_i a_i\}$, where i – number of attack, $a_i = <type, n_j>$, where type – type of attack, n_j – attacked network objects.

The countermeasure model is defined as:

$CM = \{\cup_i cm_i\}$, where i – number of countermeasure, $cm_i = <name, Pr, R>$, where *name* – name of the countermeasure, *Pr* – required privileges, *R* – required resources.

Algorithm. Since the number of countermeasure options for each type of attack is relatively small, the following rule-based algorithm is proposed for the countermeasure selection stage of the decision-support methodology:

Input data: *IS* system, attack *a*.

1. Map *a* to the *IS* using *n*.

2. Select all *cm* for *a* : *CM* <- *cm* #using rules defined on the basis of Table 1.

3. Map *CM* to countermeasures in *SM* : *CM_lim* #to define available countermeasures.

4. From *CM_lim*, select the optimal *CM_opt* considering *Pr* and *R*.

Output data: countermeasures *CM_opt*.

Results. Let us consider the application of the methodology using a simple example. Suppose that the network shown in Figure 1(b) is subjected to a spoofing attack or a routing information spoofing attack. A security system is deployed in the network (as shown in Figure 4). The implementation of the methodology proceeds as follows:

Stage 1. Based on the information collected by the auditing and logging component, the intrusion detection system uses machine learning to identify an attack or anomaly, namely, a routing information spoofing attack. It reveals that one of the sensor nodes *n* is malicious.

Stage 2. Using Algorithm 1 (*IIoT* system and attack *a* are input data). On the 1st step map *a* to the IIoT system using *n*. The output of this step is malicious node $n = <name, role, SM>$.

On the second step the countermeasures *cm* against *a* are determined using rules defined on the basis of Table 1: IF "a routing information spoofing attack" THEN action "Blocking the compromised node and redirecting traffic, key replacement.". There is only one response option *cm* "Blocking the compromised node and redirecting traffic, key replacement.". Step 3 of the algorithm consists in mapping *cm* to countermeasures in *SM* and obtaining *CM_lim = cm*. Step 4 is skipped as soon as there is only on countermeasure cm in the *CM lim* list. Implementation of this measure *cm* only requires administrator rights and user interaction for key replacement. Thus, final list of countermeasures *CM_opt* contains one countermeasure "Blocking the compromised node and redirecting traffic, key replacement." For implementation against the detected routing information spoofing attack.

Discussion. For the system to operate effectively, it is necessary to implement data collection and processing, as well as methodologies for detecting attacks and anomalies. Machine learning–based methods, such as those described in (Ahmad et al., 20122), have demonstrated good performance. This constitutes one of the directions for future research.

At present, most attacks in Table 1 have only one corresponding countermeasure, which does not require additional optimization in terms of resource usage and efficiency, and ensures low algorithmic complexity of O (1). It should be noted that the methodology can be easily extended with new countermeasures. In such cases, a multi-criteria optimization algorithm will be required to select the most suitable countermeasure, which also represents a future research direction.

In addition, algorithms for blocking network nodes and redirecting traffic deserve special attention.

Conclusion.

This study addressed the problem of decision support for responding to cyberattacks in IIoT systems. The architecture and key features of IIoT systems were analyzed. The main types of attacks and possible countermeasures were identified. A decision support methodology for responding to cyberattacks in IIoT systems was proposed, based on analytical modeling of the system, attacks, and countermeasures; machine learning for attack detection; rules; and multi-criteria optimization for countermeasure selection. Within the methodology, models of the system, attacks, and countermeasures, as well as an algorithm for countermeasure selection, were developed. The application of the methodology was demonstrated on a test example. Future work will focus on implementing attack and anomaly detection based on machine learning, analyzing the specifics of multi-criteria optimization in countermeasure selection, and examining in greater detail measures and algorithms related to traffic redirection.

References

Ahmad R., Wazirali R., & Abu-Ain T. (2022) Machine Learning for Wireless Sensor Networks Security: An Overview of Challenges and Issues. Sensors, 22(13), — 4730 p. https://doi.org/10.3390/s22134730(in Eng.)

Al-Ahmadi S. (2022) A Novel Energy-efficient Wormhole Attack Prevention Protocol for WSN based on Trust and Reputation Factors. Proceedings of the 11th International Conference on Sensor Networks. — P. 191–201. https://doi.org/10.5220/0010951400003118 (in Eng.)

Allimuthu U., & Mahalakshmi K. (2022) Efficient Mobile Ad Hoc Route Maintenance Against Social Distances Using Attacker Detection Automation. Mobile Networks and Applications, 28(1). — P. 128–159. https://doi.org/10.1007/s11036-022-02040-3 (in Eng.)

Beddoe M.A., & Guruswamy K. (2013) *U.S. Patent No. 8,601,585*. Washington, DC: U.S. Patent and Trademark Office.

Faris M., Mahmud M.N., Salleh, M.F.M., & Alnoor A. (2023). Wireless sensor network security: A recent review based on state-of-the-art works. International Journal of Engineering Business Management, 15. https://doi.org/10.1177/18479790231157220 (in Eng.)

Hu B., Tang W., & Xie Q. (2022) A two-factor security authentication scheme for wireless sensor networks in IoT environments. Neurocomputing, 500, 741–749. https://doi.org/10.1016/j.neucom.2022.05.099 (in Eng.)

Hu Y., Wu Y., & Wang H. (2014) Detection of Insider Selective Forwarding Attack Based on Monitor Node and Trust Mechanism in WSN. Wireless Sensor Network, 06(11). — P. 237–248. https://doi.org/10.4236/wsn.2014.611023 (in Eng.)

Jadhav R., & V., V. (2017) Security Issues and Solutions in Wireless Sensor Networks. International Journal of Computer Applications, 162(2), 14–19. https://doi.org/10.5120/ijca2017913256 (in Eng.)

Kardi A., & Zagrouba R. (2019) Attacks classification and security mechanisms in Wireless Sensor Networks. Advances in Science, Technology and Engineering Systems Journal, 4(6). — P. 229–243. https://doi.org/10.25046/aj040630 (in Eng.)

Keerthika M., & Shanmugapriya D. (2021) Wireless Sensor Networks: Active and Passive attacks - Vulnerabilities and Countermeasures. Global Transitions Proceedings, 2(2). — P. 362–367. https://doi.org/10.1016/j.gltp.2021.08.045 (in Eng.)

Mamdouh M., I. Elrukhsi M.A., & Khattab A. (2018) Securing the Internet of Things and Wireless Sensor Networks via Machine Learning: A Survey. 2018 International Conference on Computer and Applications (ICCA). — P. 215–218. https://doi.org/10.1109/comapp.2018.8460440 (in Eng.)

Sánchez J.D.V., Urquiza-Aguiar L., Paredes M.C.P., & Osorio D.P.M. (2020) Survey on physical layer security for 5G wireless networks. Annals of Telecommunications, 76(3–4). — P. 155–174. https://doi.org/10.1007/s12243-020-00799-8 (in Eng.)

Shahzad F., Pasha M., & Ahmad A. (2017) A survey of active attacks on wireless sensor networks and their countermeasures. arXiv Preprint. https://doi.org/10.48550/arXiv.1702.07136 (in Eng.)

Sujihelen L., Boddu R., Murugaveni S., Arnika Ms., Haldorai A., Reddy P.C.S., Feng S., & Qin J. (2022) Node Replication Attack Detection in Distributed Wireless Sensor Networks. Wireless Communications and Mobile Computing, 2022(1). Portico. https://doi.org/10.1155/2022/7252791 (in Eng.)

Teng Z., Li M., Yu L., Gu J., & Li M. (2023) Sinkhole Attack Defense Strategy Integrating SPA and Jaya Algorithms in Wireless Sensor Networks. Sensors, 23(24). — 9709 p. https://doi.org/10.3390/s23249709 (in Eng.)

Tossa F., Faga Y., Abdou W., Ezin E.C., & Gouton P. (2025) Wireless Sensor Network Deployment: Architecture, Objectives, and Methodologies. Sensors, 25(11). — 3442 p. https://doi.org/10.3390/s25113442 (in Eng.)

Turakulovich K.Z., & Tokhirovich S.L. (2019) Analysis of Security Protocols in Wireless Sensor Networks. 2019 International Conference on Information Science and Communications Technologies (ICISCT). — P. 1–4. https://doi.org/10.1109/icisct47635.2019.9012015 (in Eng.)

Verma R., & Bharti S. (2020) A Survey of Network Attacks in Wireless Sensor Networks. Information, Communication and Computing Technology. — P. 50–63. https://doi.org/10.1007/978-981-15-9671-1_4 (in Eng.)

Verma V., & Jha V.K. (2022) A Survey of Recent Research Methodologies for the Security Provisioning in Wireless Sensor Networks. International Journal of Information Security and Assurance Engineering, (IJISAE). (in Eng.)

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Ответственный редактор *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Т. Апендиев*
Верстка на компьютере: *Г.Д. Жадырановой*

Подписано в печать 31.03.2026.
Формат 60x881/8.
20,0 п.л. Заказ 1.

«Central Asian Academic Research Center» LLP
Алматы, Шевченко к-сі, 28