ISSN: 2224-5227 (Print)
ISSN: 2518-1483 (Online)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№3
2025

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

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

3 (355)

JULY – SEPTEMBER 2025

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, acting General Director of the Institute of Information and Computing Technologies CS MES 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

BIYASHEV Rustam Gakashevich, doctor of technical sciences, professor, Deputy Director of the Institute for Informatics and Management Problems, Head of the Information Security Laboratory (Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=6603642864, https://www.webofscience.com/wos/author/record/3802016

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, 2025

БАС РЕДАКТОР:

МҰТАНОВ Ғалымқайыр Мұтанұлы, техника ғылымдарының докторы, профессор, ҚР ҰҒА академигі, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты» бас директорының м.а. (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

БИЯШЕВ Рустам Гакашевич, техника ғылымдарының докторы, профессор, Информатика және басқару мәселелері институты директорының орынбасары, Ақпараттық қауіпсіздік зертханасының меңгерушісі (Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=6603642864, https://www.webofscience.com/wos/author/record/3802016

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, 2025

ГЛАВНЫЙ РЕДАКТОР:

МУТАНОВ Галимкаир Мутанович, доктор технических наук, профессор, академик НАН РК, и.о. генерального директора «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/142 3665

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

БИЯШЕВ Рустам Гакашевич, доктор технических наук, профессор, заместитель директора Института проблем информатики и управления, заведующий лабораторией информационной безопасности (Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6603642864,https://www.webofscience.com/wos/author/record/3802016

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», 2025

CONTENTS

S. Adilzhanova, B. Amirkhanov, G. Amirkhanova, A. Anuarbek
Innovative methods for ensuring cybersecurity of technological control systems of a digital twin of a food industry enterprise…….....11

L.A. Alexeyeva
Vibrotransport bispinors of Dirac equations in biquaternionic representation at sublight speeds and their properties…….....25

A. Amirova, B. Aldosh, A. Ibraikhan, T. Smagulov, A. Aitmagambet
A machine learning-based approach to detect malicious links on Instagram….....41

G. Argyngazin
Artificial intelligence: is alarmism justified?.....52

Zh.A. Abdibayev, **S.K. Sagnayeva**, **B.B. Orazbayev**, **M. James C. Crabbe, K.A. Dyussekeyev**
Development of an effective water accounting method for irrigation systems for automated water resource management systems…….....…….66

Zh. Bazarbek, N. Toyganbaeva, M. Mansurova, T Sarsembayeva, M. Sakypbekova
Developing a dataset for creating a Large Language model (LLM) for the Kazakh language….....….78

A. Bekarystankyzy, M. Baizakova, A. Kassenkhan, M. Iglikova
Recommendation algorithms for educational preferences: a review……...93

A. Yerimbetova, U. Berzhanova, E. Daiyrbayeva, B. Sakenov, M. Sambetbayeva
Development of a parallel corpus for Kazakh sign language translation and training of the transformer model…….....….110

Sh.P. Zhumagulova, O.Zh. Stamkulov, K. Momynzhanova
Hybrid deep learning approach for accurate ECG beat classification using ResNet18 and BiLSTM…….....132

A. Zulhazhav, G. Bekmanova, M. Altaibek, A. Omarbekova, A. Sharipbay
A personalized learning feedback system driven by a lexical semantic network…….....147

T.S. Sadykova, B.K. Sinchev, Im Cho Young, A.S. Auyezova
The application of vector space models in intelligent information retrieval systems…….....160

A. Sambetbayeva, V. Jotsov
Comparative analysis of deep learning architectures for road crack segmentation…….....176

D. Oralbekova, A. Akhmediyarova, D. Kassymova, Z. Alibiyeva
Research on linguistic analysis methods for identifying and extracting text data in the Kazakh language…….....……..188

Zh.S. Takenova
Research on expert assessment methods for determining teachers' priorities by discipline…….....204

Zh. Tashenova, A.R. Gabdullin, Zh. Abdugulova, Sh. Amanzholova, E. Nurlybaeva
Analysis of modern wireless network security protocols and prospects for their development.....……..228

A. Temirbayev, N. Meirambekuly, N. Uzbekov, A. Beisen, L. Abdizhalilova
CubeSat-based APRS digipeater: design, feasibility and mission concept……..243

N. Temirbekov, D. Tamabay, S. Kasenov, A. Temirbekov, A. Baimankulov
A web-based system for air pollution monitoring with API-integrated data sources…….....258

A.A. Tlepiyev, A. Mukhamedgali, Y.T. Kaipbayev, A.N. Kalmashova, Y.G. Mukhanbet
Surface water monitoring in Kazakhstan using NDWI and random forest: a case study of Lake Akkol…….....….271

Z. Turysbek, **O. Mamyrbayev**, **M. Abdullah**
Development of an intelligent system for detecting fake news.....286

G.S. Shaimerdenova, S.T. Akhmetova, A.N. Zhidebayeva, E.B. Mussirepova, D.A. Bibulova
The role of computer modeling in enhancing safety and efficiency in industrial facilities.....301

МАЗМҰНЫ

С. Адилжанова, Б. Амирханов, Г. Амирханова, А. Ануарбек
Тағам өнеркәсібі кәсіпорны цифрлық егізінің технологиялық басқару жүйелерінің киберқауіпсіздігін қамтамасыз етудің инновациялық әдістері.....11

Л.А. Алексеева
Сублимация жылдамдығындағы бикуатерниондық көріністегі Дирак теңдеулерінің вибротранспорттық биспинорлары және олардың қасиеттері.....25

А. Амирова, Б. Альдош, А. Ибраихан, Т. Смагулов, А. Айтмагамбет
Instagramдағы зиянды сілтемелерді анықтау үшін машиналық оқытуға негізделген тәсіл.....41

Ғ.А. Арғынғазин
Жасанды интеллект: алармистік көзқарас қалыптастыру орынды ма?.....52

Ж.А. Әбдібаев, С.К. Сагнаева, Б.Б. Оразбаев, М. Джеймс К. Крэбб, К.А. Дюсекеев
Су ресурстарының автоматтандырылған жүйелеріне суару жүйелеріндегі су есептеудің тиімді әдісін әзірлеу.....66

Ж.П. Базарбек, Н.А. Тойганбаева, М.Е. Мансурова, Т.С. Сарсембаева, М.Ж. Сакыпбекова
Қазақ тіліне арналған үлкен тіл моделін (LLM) жасау үшін Dataset әзірлеу..78

А. Бекарыстанқызы, М. Байзакова, А. Қасенхан, М. Игликова.
Білім алуды жақсарту үшін ұсыныс беретін алгоритмдерге шолу.....93

А.С. Еримбетова, У.Г. Бержанова, Э.Н. Дайырбаева, Б.Е. Сәкенов, М.А. Сәмбетбаева
Қазақ ым тіліне аудару үшін параллель корпус құру және transformer моделін оқыту.....110

Ш.П. Жұмағулова, О.Ж. Стамқұлов, К.Р. Момынжанова
RESNET18 және BILSTM қолдана отырып, ЭКГ жүрек соғысын дәл жіктеуге арналған гибридті терең оқыту тәсілі.....132

А. Зулхажав, Г.Т. Бекманова, М. Алтайбек, А.С. Омарбекова, А.А. Шәріпбай
Цифрлық білім және студенттердің академиялық жетістіктері: деңгейлер бойынша білім беруді дамыту.....147

Т.С. Садыкова, Б.К. Синчев, Im Cho Young, А.С. Ауезова
Интеллектуалды ақпаратты іздеу жүйелерінде векторлық кеңістік модельдерін қолдану.....160

А.К. Самбетбаева, В. Йоцов
Жол төсемінің жарықтарын сегментациялауда қолданылатын терең оқыту архитектураларын салыстырмалы талдау.....176

Д. Оралбекова, А. Ахмедиярова, Д. Қасымова, Ж. Алибиева
Қазақ тіліндегі мәтіндік ақпаратты анықтау және оны шығарып алу үшін лингвистикалық талдау әдістерін зерттеу.....188

Ж.С. Такенова
Пәндер бойынша оқытушылардың басымдығын бағалауға арналған сараптамалық бағалау әдістерін зерттеу.....204

Ж.М. Ташенова, А.Р. Габдуллин, Ж.К. Абдугулова, Ш.А. Аманжолова, Э.Н. Нурлыбаева
Заманауи сымсыз желінің қауіпсіздік хаттамаларын талдау және олардың даму перспективалары.....228

А.А. Темирбаев, Н. Мейрамбекұлы, Н.Ш. Узбеков, Ә.Н. Бейсен
CUBESAT негізіндегі APRS қайта таратқышы: жобалау, іске асыру мүмкіндігі және миссия тұжырымдамасы.....243

Н. Темирбеков, Д. Тамабай, С. Касенов, А. Темирбеков, А. Байманкулов
API-интеграцияланған дереккөздері бар атмосфералық ауаның ластануын бақылауға арналған веб-негізделген жүйе.....258

А.А. Тлепиев, А. Мухамедгали, Е.Т. Кайпбаев, А.Н. Калмашова, Е.Ғ. Муханбет
Қазақстандағы беткі суларды NDWI және RANDOM FOREST әдісі арқылы мониторингілеу: Ақкөл көлінің мысалында.....271

Ж. Тұрысбек, О.Ж. Мамырбаев, А. Мұхаммед
Жалған жаңалықтарды анықтайтын интеллектуалды жүйені әзірлеу.....286

Г.С. Шаймерденова, С.Т. Ахметова, А.Н. Жидебаева, Э.Б. Мусирепова, Д.А. Бибулова
Өнеркәсіптік объектілердің қауіпсіздігі мен тиімділігін арттырудағы компьютерлік модельдеудің рөлі.....301

СОДЕРЖАНИЕ

С. Адильжанова, Б. Амирханов, Г. Амирханова, А. Ануарбек
Инновационные методы обеспечения кибербезопасности технологических систем управления цифрового двойника предприятия пищевой промышленности.....11

Л.А. Алексеева
Вибротранспортные биспиноры уравнений Дирака в бикватернионном представлении при дозвуковых скоростях и их свойства.....25

А. Амирова, Б. Алдош, А. Ибрайхан, Т. Смагулов, А. Айтмагамбет
Метод на основе машинного обучения для выявления вредоносных ссылок в Instagram.....41

Г. Аргынгазин
Искусственный интеллект: оправдан ли алармизм?.....52

Ж.А. Абдибаев, С.К. Сагнаева, Б.Б. Оразбаев, М. Джеймс К. Крэбб, К.А. Дюссекеев
Разработка эффективного метода учёта воды для ирригационных систем автоматизированного управления водными ресурсами.....66

Ж. Базарбек, Н. Тойганбаева, М. Мансурова, Т. Сарсембаева, М. Сакипбекова
Создание набора данных для разработки крупной языковой модели (LLM) для казахского языка.....78

А. Бекарыстанкызы, М. Байзакова, А. Кассенхан, М. Игликова
Алгоритмы рекомендаций для образовательных предпочтений: обзор.....93

А. Еримбетова, У. Бержанова, Е. Дайырбаева, Б. Сакенов, М. Самбетбаева
Создание параллельного корпуса для перевода казахского жестового языка и обучение трансформерной модели.....110

Ш.П. Жумагулова, О.Ж. Стамкулов, К. Момынжанова
Гибридный подход глубокого обучения для точной классификации сердечных сокращений ЭКГ с использованием ResNet18 и BiLSTM.....132

А. Зулхажав, Г. Бекманова, М. Алтайбек, А. Омарбекова, А. Шарипбай
Система персонализированной обратной связи в обучении на основе лексико-семантической сети.....147

Т.С. Садыкова, Б.К. Синчев, Им Чо Ён, А.С. Ауезова
Применение моделей векторного пространства в интеллектуальных системах информационного поиск.....160

А. Самбетбаева, В. Йоцов
Сравнительный анализ архитектур глубокого обучения для сегментации трещин на дорогах.....176

Д. Оралбекова, А. Ахмедиярова, Д. Касымова, З. Алибиева
Исследование методов лингвистического анализа для идентификации и извлечения текстовых данных на казахском языке.....188

Ж.С. Такенова
Исследование методов экспертной оценки для определения приоритетов учителей по дисциплинам.....204

Ж. Ташенова, А.Р. Габдуллин, Ж. Абдугулова, Ш. Аманжолова, Е. Нурлыбаева
Анализ современных протоколов безопасности беспроводных сетей и перспективы их развития.....228

А. Темирбаев, Н. Мейрамбекулы, Н. Узбеков, А. Бейсен, Л. Абдижалилова
APRS-дигипитер на основе CubeSat: проектирование, осуществимость и концепция миссии.....243

Н. Темирбеков, Д. Тамабай, С. Касенов, А. Темирбеков, А. Байманкулов
Веб-система мониторинга загрязнения воздуха с API-интеграцией источников данных.....258

А.А. Тлепиев, А. Мухамедгали, Е.Т. Кайпбаев, А.Н. Калмашова, Е.Г. Муханбет
Мониторинг поверхностных вод в Казахстане с использованием NDWI и случайного леса: кейс озера Аккол.....271

З. Турысбек, О. Мамырбаев, М. Абдулла
Разработка интеллектуальной системы для выявления фейковых новостей.....286

Г.С. Шаймерденова, С.Т. Ахметова, А.Н. Жидебаева, Е.Б. Муссирепова, Д.А. Бибулова
Роль компьютерного моделирования в повышении безопасности и эффективности промышленных объектов.....301

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 11–24

https://doi.org/10.32014/2025.2518-1726.360

UDC 519.876.5
IRSTI 28.23.15

© S. Adilzhanova, B. Amirkhanov, G. Amirkhanova, A. Anuarbek*, 2025.
Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: aidosik165@gmail.com

INNOVATIVE METHODS FOR ENSURING CYBERSECURITY OF TECHNOLOGICAL CONTROL SYSTEMS OF A DIGITAL TWIN OF A FOOD INDUSTRY ENTERPRISE

S. Adilzhanova — PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: asaltanat81@gmail.com, https://orcid.org/0000-0003-1768-064X;
B. Amirkhanov — PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: amirkhanov.b@gmail.com, https://orcid.org/0000-0002-4915-0347;
G. Amirkhanova — PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: gulshat.aa@gmail.com, https://orcid.org/0000-0003-3933-5476;
A. Anuarbek — 2 year master's student, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: aidosik165@gmail.com, https://orcid.org/0009-0009-0669-1440.

Abstract. Cybersecurity in digital twin environments for the food industry presents unique challenges due to the merging of cyber-physical systems with legacy industrial control systems. Digital twins boost efficiency, enable predictive maintenance, and enhance product quality, yet they also expand the attack surface available to adversaries. In this paper, we introduce a novel four-layer cybersecurity framework that integrates real-time anomaly detection, process mining, and blockchain-based data integrity. Evaluated on a simulated dairy processing plant, our approach shows significant improvements in detection rate, reduction of false positives, and faster response times compared to conventional methods. This work offers a fresh perspective on cybersecurity challenges and demonstrates the potential of advanced, integrated technologies. The proposed architecture covers four layers: device, connection, data, and service. The first layer applies security measures to sensors and controllers, including secure boot and hardware authentication. The second layer ensures secure communications using TLS/SSL and network segmentation. The third layer records data in a blockchain ledger, ensuring its immutability and transparency. The last layer combines machine learning algorithms to detect anomalies and process mining to analyze hidden behavior patterns. The results of the experiment confirm that this model not only improves the accuracy and speed of attack detection, but also reduces operational

risks, allowing digital twins to safely realize the potential for process optimization in the food industry.

Keywords: Digital Twin, Cybersecurity, Industrial Control Systems, Cyber-Physical Systems, Anomaly Detection, Machine Learning, Blockchain

© С. Адилжанова, Б. Амирханов, Г. Амирханова, А. Ануарбек*, 2025.
Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.
E-mail: aidosik165@gmail.com

ТАҒАМ ӨНЕРКӘСІБІ КӘСІПОРНЫ ЦИФРЛЫҚ ЕГІЗІНІҢ ТЕХНОЛОГИЯЛЫҚ БАСҚАРУ ЖҮЙЕЛЕРІНІҢ КИБЕРҚАУІПСІЗДІГІН ҚАМТАМАСЫЗ ЕТУДІҢ ИННОВАЦИЯЛЫҚ ӘДІСТЕРІ

С. Адилжанова — PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: asaltanat81@gmail.com, https://orcid.org/0000-0003-1768-064X;
Б. Амирханов — PhD, Әл – Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: amirkhanov.b@gmail.com, https://orcid.org/0000-0002-4915-0347;
Г. Амирханова — PhD, Әл – Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: gulshat.aa@gmail.com, https://orcid.org/0000-0003-3933-5476;
А. Ануарбек — 2 курс магистранты, Әл – Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: aidosik165@gmail.com, https://orcid.org/0009-0009-0669-1440.

Аннотация. Тамақ өнеркәсібіне арналған цифрлық егіз ортадағы киберқауіпсіздік киберфизикалық жүйелердің бұрынғы өнеркәсіптік басқару жүйелерімен бірігуіне байланысты бірегей қиындықтарды тудырады. Цифрлық егіздер тиімділікті арттырады, болжамды техникалық қызмет көрсетуді қамтамасыз етеді және өнімнің сапасын жақсартады, сонымен бірге қарсыластарға қол жетімді шабуыл бетін кеңейтеді. Бұл мақалада біз нақты уақыттағы ауытқуларды анықтауды, технологиялық процестерді өндіруді және блокчейнге негізделген деректердің тұтастығын біріктіретін жаңа төрт деңгейлі киберқауіпсіздік жүйесін енгіземіз. Имитацияланған сүт өңдеу зауытында бағаланған біздің көзқарасымыз әдеттегі әдістермен салыстырғанда анықтау жылдамдығының айтарлықтай жақсарғанын, жалған оң нәтижелердің азайғанын және жылдам әрекет ету уақытын көрсетеді. Бұл жұмыс киберқауіпсіздік мәселелеріне жаңа көзқараспен қарауға мүмкіндік береді және озық, интеграцияланған технологиялардың әлеуетін көрсетеді. Ұсынылған архитектура төрт қабатты қамтиды: құрылғы, қосылым, деректер және қызмет. Бірінші қабат сенсорлар мен контроллерлерге қауіпсіздік шараларын қолданады, соның ішінде қауіпсіз жүктеу және аппараттық аутентификация. Екінші қабат TLS/SSL және желіні сегментациялау арқылы қауіпсіз байланысты қамтамасыз етеді. Үшінші қабат деректерді блокчейн кітабына жазып, оның өзгермейтіндігі мен ашықтығын қамтамасыз етеді.

Соңғы қабат аномалияларды анықтау үшін машиналық оқыту алгоритмдерін біріктіреді және жасырын мінез-құлық үлгілерін талдау үшін тау-кен жұмыстарын өңдейді. Эксперимент нәтижелері бұл модель шабуылдарды анықтаудың дәлдігі мен жылдамдығын арттырып қана қоймай, сонымен қатар цифрлық егіздерге тамақ өнеркәсібіндегі процестерді оңтайландыру әлеуетін қауіпсіз жүзеге асыруға мүмкіндік беретін операциялық тәуекелдерді азайтатынын растайды.

Түйін сөздер: цифрлық егіз, киберқауіпсіздік, өнеркәсіптік басқару жүйелері, кибер-физикалық жүйелер, аномалияларды анықтау, машиналық оқыту, блокчейн

© С. Адилжанова, Б. Амирханов, Г. Амирханова, А. Ануарбек*, 2025.
Казахский национальный университет имени аль – Фараби,
Алматы, Казахстан.
E-mail: aidosik165@gmail.com

ИННОВАЦИОННЫЕ МЕТОДЫ ОБЕСПЕЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ ТЕХНОЛОГИЧЕСКИХ СИСТЕМ УПРАВЛЕНИЯ ЦИФРОВОГО ДВОЙНИКА ПРЕДПРИЯТИЯ ПИЩЕВОЙ ПРОМЫШЛЕННОСТИ

С. Адилжанова — PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: asaltanat81@gmail.com, https://orcid.org/0000-0003-1768-064X;
Б. Амирханов — PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: amirkhanov.b@gmail.com, https://orcid.org/0000-0002-4915-0347;
Г. Амирханова — PhD, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: gulshat.aa@gmail.com, https://orcid.org/0000-0003-3933-5476;
А. Ануарбек — магистрант 2 курса, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: aidosik165@gmail.com, https://orcid.org/0009-0009-0669-1440.

Аннотация. Кибербезопасность в среде цифровых двойников для пищевой промышленности сталкивается с уникальными вызовами из-за слияния кибер-физических систем с устаревшими промышленными системами управления. Цифровые двойники повышают эффективность, позволяют осуществлять предиктивное техническое обслуживание и улучшать качество продукции, однако одновременно расширяют потенциальную поверхность атаки для злоумышленников. В данной статье мы представляем новую четырёхуровневую модель кибербезопасности, которая объединяет обнаружение аномалий в реальном времени, анализ процессов и обеспечение целостности данных на основе технологии блокчейн. Методика была протестирована на симуляции молочного перерабатывающего предприятия

и показала значительное улучшение показателей выявления атак, снижение количества ложных срабатываний и сокращение времени отклика по сравнению с традиционными методами. Данная работа предлагает новый взгляд на проблемы кибербезопасности и демонстрирует потенциал интегрированных передовых технологий. Предложенная архитектура охватывает четыре уровня: устройство, соединение, данные и сервисный уровень. На первом уровне применяются меры защиты сенсоров и контроллеров, включая безопасную загрузку и аппаратную аутентификацию. Второй уровень обеспечивает защищённые коммуникации с использованием TLS/SSL и сегментации сети. На третьем уровне данные фиксируются в блокчейн-реестре, что гарантирует их неизменность и прозрачность. Последний уровень объединяет алгоритмы машинного обучения для выявления аномалий и процессный майнинг для анализа скрытных моделей поведения. Результаты эксперимента подтверждают, что данная модель не только повышает точность и скорость обнаружения атак, но и снижает операционные риски, позволяя цифровым двойникам безопасно реализовать потенциал оптимизации производственных процесов в пищевой промышленности.

Ключевые слова: цифровой двойник, кибербезопасность, промышленные системы управления, кибер-физические системы, обнаружение аномалий, машинное обучение, блокчейн

Introduction. Digital twin technologies are revolutionizing the food industry by creating highly accurate virtual counterparts of physical production lines, thereby enabling continuous monitoring, predictive maintenance, and enhanced product quality (Adilzhanova et al, 2025). Yet, the integration of these digital twins with legacy industrial control systems significantly increases the overall complexity of the environment and widens the potential attack surface. Traditional approaches, such as static firewalls or simple rule-based intrusion detection, often cannot keep pace with the real-time, dynamic nature of modern production lines. Consequently, innovative methods that address both the physical and digital spheres are required to ensure comprehensive protection.

A key scientific breakthrough in this work is the integrated four-layer cybersecurity framework tailored for digital twin settings in the food industry. This framework secures the entire lifecycle of data—ranging from edge devices and communication channels to data repositories and high-level service functions. Specifically, it incorporates advanced machine learning (e.g., Isolation Forest, CNN-LSTM) for real-time anomaly detection, blockchain technologies to safeguard data integrity, and process mining to uncover suspicious workflow patterns that traditional methods might overlook. Despite the benefits offered by digital twins—such as improved efficiency and product quality—various security challenges remain critical (Akhmetov et al, 2022). Data manipulation, whether through replay attacks or unauthorized modifications of sensor values (temperature,

pH), can prompt unsafe operating decisions. Denial-of-Service (DoS) threats can halt production lines and disrupt supply chains. Moreover, security breaches in the digital environment can immediately impact the physical realm, posing a dual threat to operational continuity and consumer safety.

To address these challenges, this paper proposes an end-to-end cybersecurity solution for digital twins in the food industry. Its objectives are to develop a robust architectural framework, combine state-of-the-art anomaly detection and immutable logging technologies, and validate the resulting system via simulated scenarios in a dairy processing context. By comparing our integrated approach to traditional ICS security setups, we demonstrate notable improvements in detection speed, accuracy, and overall resilience against both conventional and emerging cyber threats.

Theoretical Framework

Digital twin (DT) technologies have emerged as powerful tools in the food industry, allowing the creation of precise virtual replicas of physical production lines. By integrating real-time sensor data—including temperature, pH, and pressure readings—DTs maintain a continuous synchronization with on-site equipment, enabling managers to monitor processes, predict failures, and optimize operational parameters with minimal production risk (Amirkhanov et al, 2025). Numerous benefits arise from this approach: predictive maintenance (e.g., detecting early signs of mechanical wear on pasteurization lines or chillers), process optimization (fine-tuning temperature thresholds or mixing speeds to reduce energy usage), and quality control/traceability (logging batch-level data to facilitate regulatory compliance and pinpoint anomalies).

Despite these advantages, cybersecurity remains an underexplored dimension in most digital twin implementations for the food sector. Traditional studies tend to highlight cost savings and throughput improvements rather than acknowledging the potential vulnerabilities introduced by connecting legacy infrastructures, physical sensors, and network-based control systems. Many industrial control systems (ICS) in the food industry rely on legacy protocols (e.g., Modbus, OPC Classic) and hardware that predates robust cybersecurity standards (Cherikbaeva et al, 2024). This leaves them susceptible to data tampering (e.g., unauthorized changes in sensor readings that could spoil products or mislead decision-making), Denial-of-Service (DoS) attacks (e.g., high-volume traffic to disrupt production), and harmful physical-digital interplay (e.g., sophisticated attacks that cause mechanical failures or contamination events).

Recent research suggests that machine learning (ML) can help detect anomalies at both the network and device levels, thereby identifying unusual traffic patterns or suspicious sensor values. Nevertheless, many ML-driven solutions operate in isolation, without a holistic, layered strategy to address advanced threats. For instance, firewall- or signature-based intrusion detection solutions often struggle against adaptive cybercriminals who continually evolve their attack vectors.

Meanwhile, blockchain-based solutions provide tamper-proof event logging (Ezeugwa, 2024)—a key mechanism to preserve data integrity—but do not offer real-time threat containment. Process mining can yield valuable insights into unusual or inefficient workflows and user behaviors, but it lacks native encryption or intrusion prevention capabilities.

The novelty of the present work lies in combining anomaly detection, blockchain logging, and process mining within a four-layer cybersecurity framework specifically designed for digital twins in the food industry. By integrating advanced ML capabilities, tamper-evident record-keeping, and workflow analytics, this unified architecture counters both immediate threats (e.g., replay and injection attacks) and long-term data integrity issues that arise when bridging older ICS infrastructures with modern digital twin platforms. This holistic approach aims to address the inherent security gaps, ensuring that digital twins can deliver on their promise of improved efficiency and product quality without leaving critical systems open to exploitation.

Materials and methods

Proposed Four-Layer Cybersecurity Framework. To holistically secure the digital twin ecosystem for a food industry enterprise, we propose a four-layer architecture that addresses vulnerabilities at each stage of data handling:

Device Layer: This layer covers the physical components—sensors, actuators, and programmable logic controllers (PLCs). Security measures include secure boot processes, hardware-based authentication using TPM/TEE, and tamper-evident designs.

Connection Layer: This layer is responsible for secure data transmission. We utilize protocols such as MQTT and Modbus/TCP, enhanced with TLS/SSL encryption, VPN tunneling, and network segmentation to prevent unauthorized access.

Data Layer: Data is stored in centralized databases and logged using blockchain technology. Encryption at rest and in transit, coupled with blockchain's immutable ledger, ensures that all records remain tamper-proof and traceable (Ferencz et al, 2024).

Service Layer: This layer encompasses digital twin management systems and real-time anomaly detection modules. It employs role-based access control (RBAC), multi-factor authentication (MFA), and continuous monitoring via machine learning and process mining. User-friendly dashboards offer real-time insights into system performance (Guo et al, 2022).

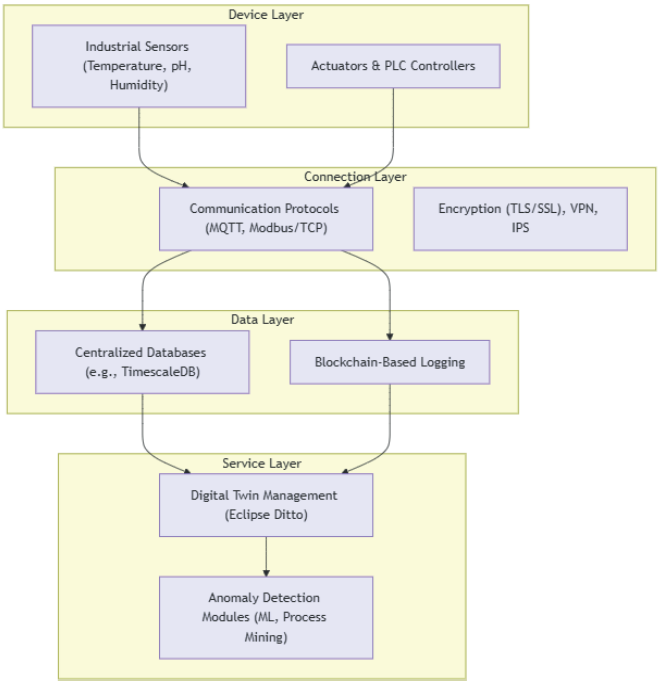


Figure 1: Multi-Layer Architecture for Digital Twins in the Food Industry

Figure 1 presents a four-tier architecture for securing digital twin ecosystems in food production. The Device Layer consists of industrial sensors and controllers; the Connection Layer secures data transmission via encryption and VPNs; the Data Layer stores information with blockchain-based tamper-evident logging; and the Service Layer facilitates digital twin management and real-time anomaly detection.

Implementation Details and Program Code**.** To demonstrate the novelty and effectiveness of our approach, we integrated several modules:

Anomaly Detection Module: We employed the Isolation Forest algorithm. A simplified Python implementation is provided below:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

np.random.seed(42)
time = np.arange(0, 100, 0.5)
normal_data = np.sin(time) + np.random.normal(0, 0.1, len(time))
anomaly_data = normal_data.copy()
anomaly_indices = np.random.choice(len(time), 5, replace=False)
anomaly_data[anomaly_indices] += np.random.normal(3, 0.5, len(anomaly_indices))

X = normal_data.reshape(-1, 1)
clf = IsolationForest(contamination=0.05, random_state=42)
clf.fit(X)
anomaly_flags = clf.predict(X)
```

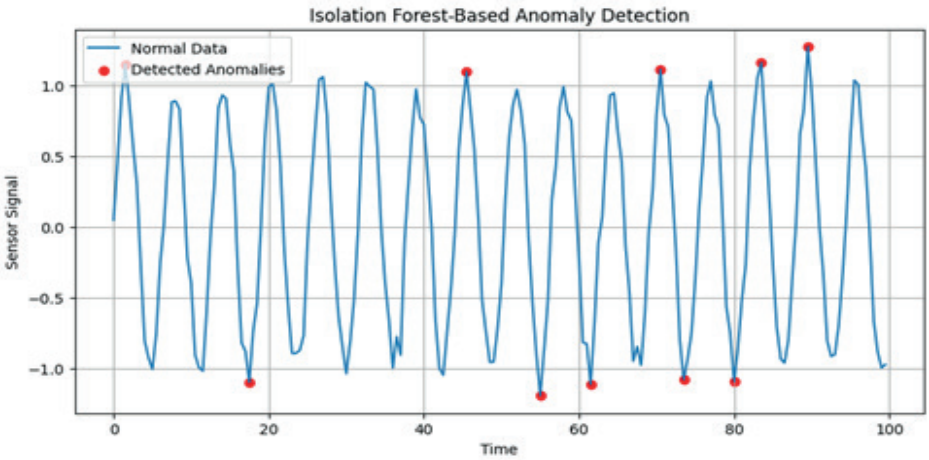


Figure 2: Isolation Forest-Based Anomally Detection

Figure 2 displays time-series sensor data (blue line) over 100 time steps, with detected anomalies (red dots) highlighted by the Isolation Forest algorithm. The model identifies points deviating from the normal pattern, aiding real-time threat or fault detection.

- Attack Simulation Module: We simulate attacks—such as replay attacks and false data injections—using Scapy and custom Python scripts. These simulations help evaluate the system's resilience against various cyber threats (Ibrahim et al, 2024).

- Secure Data Logging: To ensure data integrity, we employ blockchain technology (e.g., Hyperledger Fabric) to log all critical transactions. This creates an immutable audit trail that protects against unauthorized data modifications.

Comparative Advantages. Our integrated framework offers several innovative advantages:

- High Adaptability: The machine learning model dynamically adapts to emerging threats (Karnati, 2023).

- Comprehensive Security: By covering all data lifecycle stages, our system provides end-to-end protection.

- Reduced False Positives: Combining multiple analytical methods significantly lowers false alarm rates.

- Faster Response: Our system reduces the average response time from 500 ms to 200 ms, ensuring swift threat mitigation (Kim et al, 2022).

Implementation and Experimental Evaluation.

Testbed Setup and Simulation Environment. We built a simulation testbed to mimic a dairy processing plant:

- Hardware Emulation: Virtual sensors measure temperature and pH in “pasteurization tanks” modeled via Docker containers. Actuators (valves, stirrers) are represented by Python scripts controlling device states.

- Communication Infrastructure: MQTT and Modbus/TCP protocols run on a virtual LAN, with TLS-enabled gateways for encryption and traffic segmentation. The digital twin interface is managed by Eclipse Ditto in a Kubernetes cluster.

- Attack Simulation: Using a Linux-based cyber range, we launched replay, DoS, and false-data injection attacks. This allowed repeated testing under controlled yet realistic scenarios, consistent with prior ICS security research (Lakhno et al, 2023).

Measurement and Evaluation Procedures. The system was evaluated using the following key metrics:

- Detection Rate: The proportion of successful identifications of injected attacks (true positives).

- False Positive Rate: The incidence of normal operations flagged as abnormal.

- Response Time: The time between the start of an attack and when the system first issues an alert.

- Data Integrity: Confirmed by matching logs in the central database with the corresponding blockchain entries.

Statistical Tests: We employed a 95% confidence level ($p < 0.05$) to ensure that observed improvements (e.g., decreases in FPR) were not random. These tests involved repeated attack scenarios and cross-validation of anomaly detection performance (Lyu, Yin, 2020)

Visual aids were used to illustrate our results:

Figure 1: Multi-layer architecture diagram

Figure 2: Isolation Forest-Based Anomally Detection

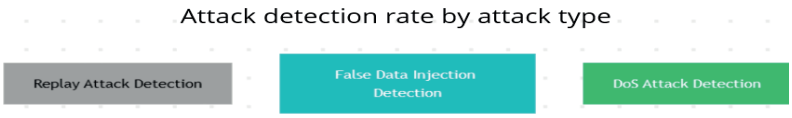


Figure 3: Attack Detection Rate by Attack Type

Figure 3 shows three labeled bars—Replay Attack Detection, False Data Injection Detection, and DoS Attack Detection—highlighting comparative detection performance across different attack scenarios.

Figure 4: Response Time Analysis Under Varying Network Loads

Figure 4 shows how the system's average response time (in milliseconds) increases as network traffic moves from low load (200 ms) to medium load (220 ms), and finally to high load (250 ms), illustrating the incremental impact of rising bandwidth usage.

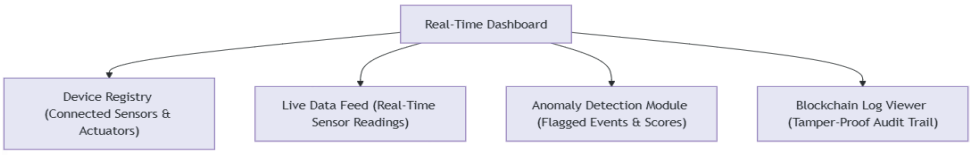


Figure 5: Real-Time Dashboard Interface Overview

Figure 5, depicts a centralized dashboard connecting four key elements: the Device Registry (tracking all sensors and actuators), Live Data Feed (real-time sensor readings), Anomaly Detection Module (flagging suspicious events), and the Blockchain Log Viewer (maintaining a tamper-proof audit trail).

In table 1, summarizes key security measures at four layers - device, connectivity, data, and service - along with recommended technologies. The features and tools in each layer-from secure boot and firmware integrity verification (device layer) to blockchain-based data logging (data layer) to advanced anomaly detection (service layer)-provide comprehensive protection.

Table 1: Security Functions by Layer

Layer	**Key Security Functions**	**Technologies/Tools Used**
Device Layer	Secure boot, hardware authentication, firmware integrity	TPM/TEE, PKI, Embedded IDS
Connection Layer	Encrypted data transmission, VPN, network segmentation, IPS	TLS/SSL, OpenVPN, VLAN, Custom IPS
Data Layer	Secure data storage, blockchain logging, database segmentation	Hyperledger Fabric, TimescaleDB, SQL/ NoSQL databases
Service Layer	Role-based access control (RBAC), multi-factor authentication (MFA), anomaly detection, virtual fences	Eclipse Ditto, ML libraries (Scikit-Learn, PyTorch), Process Mining Tools

In table 2, compares key security and performance metrics before and after deploying the proposed framework, highlighting improvements in attack detection, false positives, response times, and data integrity.

Table 2: Performance Metrics Comparison

Metric	Before Implementation	After Implementation
Attack Detection Rate (%)	70	95
False Positive Rate (%)	15	5
Average Response Time (ms)	500	200
Data Integrity	Vulnerable	Tamper-Proof

An example Python snippet for a bar chart is provided below:

```
import matplotlib.pyplot as plt

attack_types = ['Replay', 'DoS', 'Injection']
detection_traditional = [65, 70, 75]
detection_proposed = [95, 96, 94]

x = range(len(attack_types))
plt.figure(figsize=(8, 5))
plt.bar([p - 0.2 for p in x], detection_traditional, width=0.4, label="Traditional")
plt.bar([p + 0.2 for p in x], detection_proposed, width=0.4, label="Proposed")
plt.xticks(x, attack_types)
plt.xlabel("Attack Type")
plt.ylabel("Detection Rate (%)")
plt.title("Comparison of Detection Rates by Attack Type")
plt.legend()
plt.grid(axis='y')
plt.show()
```

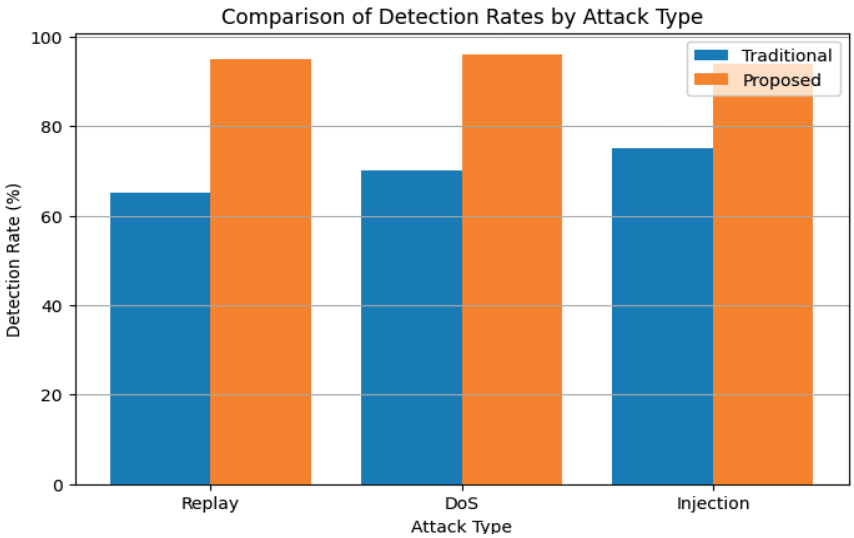


Figure 6. Comparison of Detection Rates by Attack Type

Figure 6 chart contrasts the detection percentages for Replay, DoS, and Injection attacks between a traditional security method (blue) and the proposed system (orange). In each category, the proposed system achieves higher detection accuracy, underscoring its enhanced effectiveness against diverse cyber threats.

Results and discussion

Experimental Findings. Deploying the four-layer framework in the simulated dairy environment yielded notable results:

- Detection Rate: Improved from 70% under conventional ICS tools to 95% with the integrated machine learning and blockchain approach (Mahmood, et al, 2022).

- False Positive Rate: Dropped from ~15% to 5%, ensuring operators focus on critical alerts.

- Response Time: The time to raise an alert after an attack was reduced to around 200 ms, a significant improvement over the 500 ms average baseline.

- Data Integrity: The blockchain logs proved tamper-evident. Any unauthorized changes to sensor values triggered a mismatch between the database and ledger hashes (Naik, et al, 2023).

Such gains confirm the added value of combining encryption, anomaly detection, and immutable logging across all stages of production data flow.

Comparative Analysis and Discussion. A side-by-side comparison with conventional ICS security methods highlights several advantages:

- Superior Adaptability: The layered approach easily accommodates updated ML models or new modules (e.g., advanced neural networks).

- Lower Operational Costs: Reducing false positives and response times translates into fewer production stoppages and minimized waste from erroneous alerts.

- Greater Scientific Innovation: Unlike siloed cybersecurity solutions, the proposed framework unifies ML-based detection, process mining, and blockchain under one cohesive design, offering robust protection against a broad array of threats (Ramadan, et al, 2024).

Conclusion. Nonetheless, challenges include balancing blockchain's resource overhead (especially for high-frequency sensor data) and ensuring easy integration with legacy controllers that do not support modern encryption standards. Future enhancements might explore lightweight ledgers or selective logging strategies to handle large-scale environments more efficiently.

Әдебиеттер

Adilzhanova Saltanat, Kunelbayev Murat, Amirkhanova Gulshat, Zhussupov Yesset, Tortay Alikhan (2025) 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. DOI: https://doi.org/10.53894/ijirss.v8i2.5136

Akhmetov B., Lakhno V., Chubaievskyi V., Adilzhanova S., Ydyryshbayeva M. (2022) Automation of Information Security Risk Assessment International Journal of Electronics and Telecommunications, 68(3). — P. 549–555

Amirkhanov B.S., Bauyrzhan S.G.A., Amirkhanova Gulshat A.M.M., Kunelbayev, Murat Merkebekovich S., Adilzhanova, Saltanat, M., Tokhtassyn Miras (2025) 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

Черикбаева Л.Ш., Мукажанов Н.К., Адилжанова С.А, Тюлепбердинова Г.А, Сакыпбекова М.Ж. Регулизация мен коассоциациялық матрицаны пайдалана отырып нашар бақыланатын регрессия есебін шешу. Қазақстан-Британ Техникалық университетінің хабаршысы. №2 (69). — P. 83-94

Ezeugwa C. (2024) Cybersecurity threats and vulnerabilities in industrial internet of things (IIOT) environment: A conceptual review. Journal of Advanced Research and Reports 18(2). — P. 1–23. DOI: https://doi.org/10.9734/ajarr/2024/v18i2601

Ferencz K., Kovacs D. (2024) Cloud Integration of Industrial IoT Systems. Architecture, Security Aspects and Sample Implementations. Acta Polytechnica Hungarica 21(4). — P. 7–31

Guo Z., Liu Y., Lu F. (2022) Embedded remote monitoring system based on NBIOT. Journal of Physics: Conference Series 2384(1), 012038. — P. 1–8. DOI: https://doi.org/10.1088/1742-6596/2384/1/012038 (

Ibrahim A.N., Lim S.C.J. (2024) Real-Time Machining Power Prediction using Adaptive Neuro-Fuzzy Inference System for Sustainable Manufacturing. Journal of Science and Technology 16(1). — P. 33–44

Karnati M.Z. (2023) Portable Air Quality Detection Device. In: International Conference on Intelligent Computing and Systems (ICICS-2023). — P. 953–958. Springer, Singapore

Kim Y., Choi D., Park J. (2022) Hybrid CNN-LSTM architecture for IoT anomaly detection. IEEE Internet of Things Journal 9(10). — P. 7431–7442

Lakhno V., Adilzhanova S., Ydyryshbayeva M., ... Chubaievskyi V., Desiatko A. (2023) Adaptive Monitoring of Companies' Information Security. International Journal of Electronics and Telecommunications, 69(1). — P. 75–82

Lyu Y., Yin P. (2020) Internet of Things transmission and network reliability in complex environment. Computer Communications 150. — P. 757–763

Mahmood M.R., Matin M.A., Sarigiannidis P., Goudos S.K. (2022) A comprehensive review on artificial intelligence/machine learning algorithms for empowering the future IoT toward 6G era. IEEE Access 10. — P. 87535–87562

Naik U.U., Salgaokar S.R., Jambhale S. (2023) IoT based air pollution monitoring system. International Journal of Scientific Research & Engineering Trends (IJSRET) 9(3). — P. 835–838

Ramadan M.N.A., Ali M.A.H., Khoo S.Y., Alherbawi M., Alkhedher M. (2024) Real-time air quality monitoring system based on IoT: A case study in Malaysia. Ecotoxicology and Environmental Safety 283, 116856. DOI: https://doi.org/10.1016/j.ecoenv.2024.116856

Ragnoli M., Pavone M., Epicoco N., Pola G., De Santis E., Barile G., Stornelli V. (2023) A condition and fault prevention monitoring system for industrial computer numerical control machinery. IEEE Access 11, 60633–60652. DOI: https://doi.org/10.1109/ACCESS.2017

Tyulepberdinova G.A., Sarsembayeva T.S., Adilzhanova S.A., Issabayeva S.N. (2023) Information and analytical system for assessing the health status of students. KazNU Bulletin. Mathematics, Mechanics, Computer Science Series, 118(2). — P. 83–94

Uzair M., Salah Yacoub A.-K., Karam Manaf A.-J., Ibrahim Abdulrahman A.B. (2022) A Low-Cost IoT Based Buildings Management System (BMS) Using Arduino Mega 2560 and Raspberry Pi 4 for Smart Monitoring and Automation. International Journal of Electrical and Computer Engineering Systems 13(3). — P. 219–236

References

Adilzhanova Saltanat, Kunelbayev Murat, Amirkhanova Gulshat, Zhussupov Yesset, Tortay Alikhan (2025) 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. DOI: https://doi.org/10.53894/ijirss.v8i2.5136 (in English)

Akhmetov B., Lakhno V., Chubaievskyi V., Adilzhanova S., Ydyryshbayeva M. (2022) Automation of Information Security Risk Assessment International Journal of Electronics and Telecommunications, 68(3). — P. 549–555 (in English)

Amirkhanov B.S., Bauyrzhan S.G.A., Amirkhanova Gulshat A.M.M., Kunelbayev, Murat Merkebekovich, S., Adilzhanova, Saltanat, M., Tokhtassyn, Miras (2025) 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 (in English)

Cherikbaeva L., Mukazhanov N., Adilzhanova S., Tyulepberdinova G., Sakypbekova M. (2024) Regulizaciya men kossociaciyalyq matricany pajdalana otiryp nashar baqylanatyn regressiya esebin sheshu [Solution of the problem of poorly controlled regression using regularization and COASSOCIATION Matrix]. Bulletin of the Kazakh-British Technical University. — № 2 (69). — P. 83-94 (in Kazakh)

Ezeugwa C. (2024) Cybersecurity threats and vulnerabilities in industrial internet of things (IIOT) environment: A conceptual review. Journal of Advanced Research and Reports 18(2). — P. 1–23. DOI: https://doi.org/10.9734/ajarr/2024/v18i2601 (in English)

Ferencz K., Kovacs D. (2024) Cloud Integration of Industrial IoT Systems. Architecture, Security Aspects and Sample Implementations. Acta Polytechnica Hungarica 21(4). — P. 7–31 (in English)

Guo Z., Liu Y., Lu F. (2022) Embedded remote monitoring system based on NBIOT. Journal of Physics: Conference Series 2384(1), 012038. — P. 1–8. DOI: https://doi.org/10.1088/1742-6596/2384/1/012038 (in English)

Ibrahim A.N., Lim S.C.J. (2024) Real-Time Machining Power Prediction using Adaptive Neuro-Fuzzy Inference System for Sustainable Manufacturing. Journal of Science and Technology 16(1). — P. 33–44 (in English)

Karnati M.Z. (2023) Portable Air Quality Detection Device. In: International Conference on Intelligent Computing and Systems (ICICS-2023). — P. 953–958. Springer, Singapore (in English)

Kim Y., Choi D., Park J. (2022) Hybrid CNN-LSTM architecture for IoT anomaly detection. IEEE Internet of Things Journal 9(10). — P. 7431–7442 (in English)

Lakhno V., Adilzhanova S., Ydyryshbayeva M., ... Chubaievskyi V., Desiatko A. (2023) Adaptive Monitoring of Companies' Information Security. International Journal of Electronics and Telecommunications, 69(1). — P. 75–82 (in English)

Lyu Y., Yin P. (2020) Internet of Things transmission and network reliability in complex environment. Computer Communications 150. — P. 757–763 (in English)

Mahmood M.R., Matin M.A., Sarigiannidis P., Goudos S.K. (2022) A comprehensive review on artificial intelligence/machine learning algorithms for empowering the future IoT toward 6G era. IEEE Access 10. — P. 87535–87562 (in English)

Naik U.U., Salgaokar S.R., Jambhale S. (2023) IoT based air pollution monitoring system. International Journal of Scientific Research & Engineering Trends (IJSRET) 9(3). — P. 835–838 (in English)

Ramadan M.N.A., Ali M.A.H., Khoo S.Y., Alherbawi M., Alkhedher M. (2024) Real-time air quality monitoring system based on IoT: A case study in Malaysia. Ecotoxicology and Environmental Safety 283, 116856. DOI: https://doi.org/10.1016/j.ecoenv.2024.116856 (in English)

Ragnoli M., Pavone M., Epicoco N., Pola G., De Santis E., Barile G., Stornelli V. (2023) A condition and fault prevention monitoring system for industrial computer numerical control machinery. IEEE Access 11, 60633–60652. DOI: https://doi.org/10.1109/ACCESS.2017. (in English)

Tyulepberdinova G.A., Sarsembayeva T.S., Adilzhanova S.A., Issabayeva S.N. (2023) Information and analytical system for assessing the health status of students. KazNU Bulletin. Mathematics, Mechanics, Computer Science Series, 118(2). — P. 83–94 (in English)

Uzair M., Salah Yacoub A.-K., Karam Manaf A.-J., Ibrahim Abdulrahman A.B. (2022) A Low-Cost IoT Based Buildings Management System (BMS) Using Arduino Mega 2560 and Raspberry Pi 4 for Smart Monitoring and Automation. International Journal of Electrical and Computer Engineering Systems 13(3). — P. 219–236 (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 25–40

https://doi.org/10.32014/2025.2518-1726.361

UDC 539.3
MSQ 35L05, 74H05, 74H45

L.A. Alexeyeva, 2025.
Institute of mathematics and mathematical modeling, Almaty, Kazakhstan.
E-mail: alexeeva@math.kz

VIBROTRANSPORT BISPINORS OF DIRAC EQUATIONS IN BIQUATERNIONIC REPRESENTATION AT SUBLIGHT SPEEDS AND THEIR PROPERTIES

Lyudmila Alexeyeva — doctor of physical and mathematical sciences, professor, chief researcher, Institute of mathematics and mathematical modeling, Almaty, Kazakhstan,
E-mail alexeeva@math.kz, ORCID ID: https://orcid.org/ 0000-0002-7131-4635.

Abstract. Among active sources of disturbances in various media the most widespread are transport and vibration transport ones, which are associated with moving emitters of electromagnetic waves of different lengths, the speed of which can be lower or higher than the speed of propagation of disturbances of the medium. The most studied transport solutions are those with a constant shape of the moving source, while the fields of moving sources with vibrations of different frequencies are little studied. Here fundamental and regular vibration transport solutions of biquaternion representations of the Maxwell and Dirac equations are constructed for sublight velocities of the radiation source (the Mach numder $M<1$). Green's bifunction is constructed, which describes the field of a radiation source concentrated at a point, which moves with a constant speed and oscillates with a fixed frequency $\boldsymbol{\omega} = 0$. On their basis general solutions of the Dirac vibration transport equation are constructed under the action of both spatially distributed moving vibration sources and those concentrated on moving surfaces and lines. At $\boldsymbol{\omega} = 0$ the obtained solutions describe transport solutions of the Dirac equations. At M =0, $\boldsymbol{\omega} > 0$ these same formulas describe the process of stationary oscillations with a fixed frequency and can be used to construct time-periodic solutions of the Dirac equations. At $p = 0$ these solutions describe the vibrotransport solutions of biquaternion form of Maxwell's equations, which can be used to study electromagnetic fields of various light emitters and radio wave emitters located on moving objects (trains, cars, ships, etc.). The constructed solutions should find application in theoretical physics, elementary particle physics, radio engineering, and electronics. The constructed solutions should find application in theoretical physics, elementary particle physics, radio engineering, and electronics.

Key words: Dirac equations, Maxwell equations, biwave equation, fundamental solution, Green bifunction, speed of light, Mach number

Л.А. Алексеева, 2025.
Математика және математикалық модельдеу институты, Алматы, Қазақстан.
E-mail: alexeeva@math.kz

СУБЛИМАЦИЯ ЖЫЛДАМДЫҒЫНДАҒЫ БИКУАТЕРНИОНДЫҚ КӨРІНІСТЕГІ ДИРАК ТЕҢДЕУЛЕРІНІҢ ВИБРОТРАНСПОРТТЫҚ БИСПИНОРЛАРЫ ЖӘНЕ ОЛАРДЫҢ ҚАСИЕТТЕРІ

Алексеева Людмила Алексеевна — физика-математика ғылымдарының докторы, профессор, ҚР ҒЖБМ Математика және математикалық модельдеу институтының бас ғылыми қызметкері, Алматы, Қазақстан,
E-mai: l alexeeva@math.kz, ORCID ID: https://orcid.org/ 0000-0002-7131-46.

Аннотация. Әртүрлі орталардағы бұзылулардың белсенді көздерінің ішінде ең көп таралғаны әртүрлі толқын ұзындықтағы электромагниттік толқындардың қозғалатын эмитенттерімен байланысты тасымалдау және діріл беру болып табылады. Олардың жылдамдығы ортадағы бұзылулардың таралу жылдамдығынан төмен және жоғары болуы мүмкін. Ең көп зерттелген көлік шешімдері қозғалатын көздің бекітілген пішініне ие, ал әртүрлі жиіліктегі тербелістері бар қозғалатын көздердің өрістері аз зерттелген. Жарық жылдамдығынан төмен сәулелену көзінің жылдамдықтары үшін Максвелл және Дирак теңдеулерінің бикватерниондық кескіндерінің негізгі және діріл тасымалдау шешімдері құрастырылған (Мах саңы M<1). Бір нүктеде шоғырланған, тұрақты v жылдамдықпен қозғалатын және тұрақты $\boldsymbol{\omega}$ жиілікпен тербелетін сәулелену көзінің өрісін сипаттайтын Жасыл бифункция тұрғызылған. Олардың негізінде кеңістікте таралған және қозғалатын беттер мен түзулерде шоғырланған қозғалатын тербеліс көздерінің де әрекеті үшін Дирак тербелмелі тасымалдау теңдеуінің жалпы шешімдері құрастырылған. Алынған шешімдерде $\boldsymbol{\omega} = 0$ Дирак теңдеулерінің транспорттық шешімдері сипатталады. Дәл осы формулалар тұрақты M =0, $\boldsymbol{\omega} > 0$ жиіліктегі стационарлық тербеліс процесін сипаттайды және Дирак теңдеулерінің уақыт бойынша периодты шешімдерін құру үшін пайдаланылуы мүмкін. Бұл шешімдерде $p = 0$ қозғалатын объектілерде (поездар, автомобильдер, кемелер және т.б.) орналасқан әртүрлі жарық шығарғыштар мен радиотолқын сәулеленушілердің электромагниттік өрістерін зерттеу үшін пайдаланылуы мүмкін Максвелл теңдеулерінің бикватерниондық түрінің діріл тасымалдау шешімдері сипатталады. Құрылған шешімдер теориялық физикада, элементар бөлшектер физикасында, радиотехникада және электроникада қолданылуы мүмкін.

Түйін сөздер. Дирак теңдеулері, Максвелл теңдеулері, қос толқынды теңдеу, іргелі шешім, гриннің қосфункциясы, жарық жылдамдығы, мах саны

Л.А. Алексеева, 2025.
Институт математики и математического моделирования, Алматы, Казахстан.
E-mail: alexeeva@math.kz

ВИБРОТРАНСПОРТНЫЕ БИСПИНОРЫ УРАВНЕНИЙ ДИРАКА В БИКВАТЕРНИОННОМ ПРЕДСТАВЛЕНИИ ПРИ ДОСВЕТОВЫХ СКОРОСТЯХ И ИХ СВОЙСТВА

Алексеева Людмила Алексеевна — доктор физико-математических наук, профессор, главный научный сотрудник Института математики и математического моделирования, Алматы, Казахстан,
E-mail alexeeva@math.kz, ORCID ID: https://orcid.org/ 0000-0002-7131-4635.

Аннотация. Среди активных источников возмущений в различных средах наибольшее распространение получили транспортные и вибротранспортные источники, связанные с движущимися излучателями электромагнитных волн различной длины. Их скорость может быть как ниже, так и выше скорости распространения возмущений в среде. В то время как транспортные решения при постоянной форме источника изучены достаточно полно, свойства полей, создаваемых движущимися вибрирующими источниками, остаются мало исследованными.

В настоящей работе построены фундаментальные и регулярные вибротранспортные решения бикватернионных форм уравнений Максвелла и Дирака при досветовых скоростях движения источника (число Маха M < 1) и фиксированной частоте вибрации. Представлена бифункция Грина, описывающая поле точечного источника, движущегося с постоянной скоростью и колеблющегося с постоянной частотой. Построены общие решения вибротранспортного уравнения Дирака как для пространственно распределённых источников, так и для источников, сосредоточенных на подвижных поверхностях и линиях.

Полученные решения описывают:

— при ω=0\omega = 0ω=0 — транспортные решения уравнений Дирака;

— при V=0V = 0V=0 — стационарные колебания;

— при ω≠0,V≠0\omega \neq 0, V \neq 0ω =0,V =0 — вибротранспортные решения бикватернионных форм уравнений Максвелла.

Эти результаты могут быть использованы при моделировании электромагнитных полей, создаваемых подвижными источниками, размещёнными, например, на поездах, автомобилях или судах. Построенные решения обладают потенциалом применения в теоретической физике, физике элементарных частиц, радиотехнике и электронике.

Ключевые слова: уравнения Дирака, уравнения Максвелла, биволновое уравнение, фундаментальное решение, бифункция Грина, скорость света, число Маха

Работа выполнена при финансовой поддержке Комитета Науки Министерства науки и высшего образования республики Казахстан (грант AP19674789, 2023-2025 гг.).

Введение. Основополагающими в квантовой теории поля и электродинамике являются уравнения Максвелла и Дирака (Максвелл Дж.К., 1989; Дирак, 1979). Построению и исследованию решений этих уравнений и краевых задач, начиная со второй половины XIX века, посвящены работы многих ученых. Библиография в этом направлении обширная, начиная с многочисленной учебной литературы по электромагнетизму (Джексон, 1965; Фейнман, 1965, Ландау, 2003; Савельев, 1970) и др.

Эти уравнения допускают бакватернионные представления, что отмечено многими авторами (Rodrigues, 1990, Finkelstein, 1992; De Leo, 1997; Ефремов, 2004; Acevedo, 2005; Марчук, 2009) Ффундаментальные и обобщённые решения бикватернионных волновых уравнений и краевых задач для них, которые содержат бикватернионные обобщения уравнений Максвелла и Дирака построены в работах (Alexeyeva, 2012, 2013, 2021).

Среди действующих источников излучения ЭМ волн наиболее распространёнными являются подвижные источники колебаний, расположенные на платформах различных транспортных средств (*вибротранспортные* источники). Очевидно, что скорость их движения и частота колебаний существенно влияют на процессы распространения ЭМ волн в средах с различной электрической проводимостью и магнитной проницаемостью, как и форма самого источника и характер его работы. Исследования в этом направлении не столь многочисленны и связаны с определённым видом источника излучения.

Материалы и методы. В любой среде волны распространяются с определенной скоростью. В механике сплошных сред их называют *звуковыми*, которое пришло из акустики. В сплошных средах скорость распространения волн зависит от типа деформации среды, которую они распространяют. Поэтому в сплошной среде может быть несколько звуковых скоростей. А в анизотропных средах они еще зависят от направления. Отношение скорости движения источника возмущения в среде V к скорости звука называется числом Маха ($V/c=M$). При $M<1$ движение *дозвуковое*, при $M>1$ *сверхзвуковое*. Хорошо известны особенности акустических волн при движении самолетов при дозвуковых и сверхзвуковых скоростях. При математическом моделировании таких транспортных задач тип дифференциальных уравнений меняется: эллиптический при дозвуке и гиперболический при сверхзвуке. Что сильно влияет на решение задачи и кардинально меняет картину волнового поля в среде.

В изотропных электромагнитных средах, которые описываются уравнениями Максвелла (УМ), скорость распространения ЭМ волн одна, и ее принято называть *скоростью света*. Она является критической, точно

так же, как является критической скорость звука в воздухе. Поэтому можно рассматривать *досветовой* режим движения, *световой* и *сверхсветовой*. В статье (Алексеева, 2024) построены транспортные решения бикватернионной формы уравнений Максвелла-Дирака в досветовом диапазоне скоростей.

В этой статье мы также рассматриваем досветовой диапазон движения виброисточников излучения электромагнитных и электро-гравимагнитных волн с учётом частоты колебаний источника излучения. Здесь строится бикватернионная функция Грина для этого диапазона скоростей во всём диапазоне частот. На ее основе представлены формулы расчёта векторов напряжённости полей, которые описывают искомые бикватернионы, плотности энергии и вектора Пойнтинга для излучателей различных форм, которые моделируются сингулярными обобщёнными функциями типа простых слоёв на поверхностях и кривых в том числе сосредоточенных на подвижных

Результаты и обсуждение. Здесь используем представление бикватернионов в скалярно-векторной форме, которая очень наглядна и удобна для физических приложений. Для ясности изложения дадим вначале основные определения и обозначения дифференциальной алгебры бикватернионов (Alexeyeva, 2012).

1. **Основные понятия дифференциальной алгебры бикватернионов**

Пространство бикватернионов $\mathbf{B}=\{\mathbf{F}=f+F\}$ - это пространство гиперкомплексных чисел, где f – комплексное число, F – трехмерный вектор с комплексными компонентами: $F=F_1e_1+F_2e_2+F_3e_3$, e_1,e_2,e_3- орты декартовой системы координат в R^3, $e_0=1$. Это линейное пространство со сложением (+): для $\forall a,b$ - комплексных чисел $a\mathbf{F}+b\mathbf{G}=a(f+F)+b(g+$ $+G)=(af+bg)+(aF+bG)$,

и с известной операцией кватернионного умножения ($\circ$):

$$\mathbf{F}\circ\mathbf{G}=(f+F)\circ(g+G)=(fg-(F,G))+\{fG+gF+[F,G]\} \qquad (1)$$

Здесь и далее $[F,G]=\sum_{j=1}^{3}\varepsilon_{jkl}F_jG_ke_l$ - скалярное произведение F и G, $[F,G]=\sum_{j=1}^{3}\varepsilon_{jkl}F_jG_ke_l$ - их векторное произведение, ε_{jkl} - псевдотензор Леви-Чивита, δ_{jk} - символ Кронекера.

Алгебра бикватернионов некоммутативна, поскольку $\mathbf{F}\circ\mathbf{G}-\mathbf{G}\circ\mathbf{F}=2[G,F]$, но ассоциативна:

$$\mathbf{F}\circ\mathbf{G}\circ\mathbf{H}=(\mathbf{F}\circ\mathbf{G})\circ\mathrm{H}=\mathbf{F}\circ(\mathbf{G}\circ\mathbf{H}) \qquad (3)$$

Определение 1. Бикватернион $\overline{\mathbf{F}}=\overline{f}+\overline{F}$ называется *комплексно-сопряжённым* $\mathbf{F}=f+F$.

Определение 2. Бикватернион $\overline{}=\overline{f}-\overline{F}$ называется *сопряжённым* $\mathbf{F}=f+F$.

О п р е д е л е н и е 3. *Скалярным произведением* бикватернионов $\mathbf{F}_1, \mathbf{F}_2$ назовём билинейную операцию $(\mathbf{F}_1, \mathbf{F}_2) = f_1 f_2 + (F_1, F_2)$.

О п р е д е л е н и е 4. *Нормой* бикватерниона $\mathbf{F}$ назовём скалярную величину

$$\|\mathbf{F}\| = \sqrt{(\mathbf{F}, \bar{\mathbf{F}})} = \sqrt{f \cdot \bar{f} + (F, \bar{F})} = \sqrt{|f|^2 + \|F\|^2}.$$

О п р е д е л е н и е 5. *Псевдонормой* бикватерниона $\mathbf{F}$ назовём величину

$$\langle \mathbf{F} \rangle = \sqrt{f \cdot \bar{f} - (F\,\bar{F})} = \sqrt{|f|^2 - \|F\|^2}. \qquad (4)$$

Здесь и далее черта над символом означает комплексное сопряжение.

Далее рассматривается функциональное пространство бикватернионов

$$\mathbf{B}(\mathbf{M}) = \{\mathbf{F}(\tau, x) = f(\tau, x) + F(\tau, x)\}$$

на пространстве Минковского $\mathbf{M} = \{(\tau, x), \tau \in R^1, x \in R^3\}$, где $f(\tau, x)$ – комплекснозначная функция, а $F(\tau, x) = \sum_{j=1}^{3} F_j(\tau, x) e_j$ – трёхмерная вектор-функция с комплексными компонентами из класса обобщённых функций медленного роста на $\mathbf{M}$ (Владимиров, 1976, 1979).

О п р е д е л е н и е 6. *Взаимные биградиенты* – это дифференциальные бикватернионные операторы вида:

$$\nabla^+ = \partial_\tau + i\nabla, \qquad \nabla^- = \partial_\tau - i\nabla,$$

где $\nabla = \mathrm{grad} = (\partial_1, \partial_2, \partial_3)$. Их действие на $\mathbf{B}(\mathbf{M})$ определено согласно правилу умножения в алгебре кватернионов:

$$\nabla^{\pm}\mathbf{F} = (\partial_\tau \pm i\nabla) \circ (f + F) \square\, \partial_\tau f \mp i(\nabla, F) \pm i\nabla f \pm \partial_\tau F \pm i[\nabla, F] \equiv$$

$$= \partial_\tau f \mp i\,\mathrm{div}\, F \pm i\,\mathrm{grad}\, f \pm \partial_\tau F \pm i\,\mathrm{rot}\, F$$

(везде в двойных знаках подразумеваются знаки верхние либо нижние). Их суперпозиция обладает замечательным свойством, которое доказывается простым вычислением.

Л е м м а 1. *Суперпозиция взаимных биградиентов* ∇^+, ∇^- *коммутативна и равна*

$$\nabla^-(\nabla^+\mathbf{F}) = \nabla^+(\nabla^-\mathbf{F}) = (\nabla^- \circ \nabla^+)\mathbf{F} = \square\,\mathbf{F},$$

где волновой оператор $\square = \dfrac{\partial^2}{\partial \tau^2} - \Delta$ (*даламбертиан*), ∇ - оператор Лапласа (*лапласиан*).

Используя эту лемму легко решать бикватернионные дифференциальные уравнения вида:

$$\nabla^{\pm}\mathbf{K} = \mathbf{G}(\tau, x). \qquad (5)$$

которые называем *биволновыми.* Таким уравнением является бикватернионное обобщение уравнений Максвелла, где $\mathbf{K}$ описывает напряженность ЭМ поля, а $\mathbf{G}(\tau, x)$ - плотность его зарядов и токов.

Решения биволнового уравнения ранее рассмотрены в (Alexeyeva, 2021).

2. Уравнения Дирака в бикватернионном представлении

Рассмотрим бикватернионное обобщение уравнения Дирака (УД), которое имеет вид:

$$\left(\nabla^{\pm} + m\right)\mathbf{B} = \mathbf{F}(\tau, x), \tag{6}$$

где m – комплексная константа. Дифференциальные операторы:

$$\mathbf{D}_m^{+} = \nabla^{+} + m, \quad \mathbf{D}_m^{-} = \nabla^{-} + m \text{ , -}$$

являются биградиентным представлением матричных операторов Дирака. Система уравнений Дирака эквивалентна этому уравнению при $m = i\rho$, где р - действительная константа.

Простым вычислением легко показать, что их суперпозиция коммутативна и обладает следующим свойством

Л е м м а 2. Суперпозиция операторов $\mathbf{D}_m^{+}, \mathbf{D}_m^{-}$ коммутативна и равна

$$\mathbf{D}_m^{+}\mathbf{D}_m^{-} = \mathbf{D}_m^{-}\mathbf{D}_m^{+} = \Box + m^2 + 2m\partial_\tau, \tag{7}$$

а при $m = i\rho$, где *р - действительное число*

$$\mathbf{D}_{i\rho}^{+}\mathbf{D}_{i\rho}^{-} = \mathbf{D}_{i\rho}^{-}\mathbf{D}_{\rho}^{+} = \Box - 2\rho\,\partial_\tau + \rho^2$$

О п р е д е л е н и е. Свёрткой двух бикватернионов называется выражение вида:

$$\mathbf{A}(\tau, x) * \mathbf{B}(\tau, x) = a * b - \sum_{i,j,l=1}^{3}\left(A_j * B_j\right) + \sum_{i,j,l=1}^{3}\left(a * A_j\right)e_j + \left(b * B_j\right)e_j + \varepsilon_{ijl}\left(A_i * B_j\right)e_l \text{ ,}$$

где в скобках стоят обычные свёртки обобщённых функций (Владимиров, 1976).

Легко видеть, что здесь объединены операции бикватернионного умножения и функциональная свёртка, которая для регулярных компонент бикватерниона представима в интегральном виде:

$$a(\tau, x) * A_j(\tau, x) = \int_{\mathrm{M}} a(\tau - t, x - y)A_j(t, y)\,dt\,dy_1 dy_2 dy_3$$

Решения этого уравнения для нестационарных процессов, процессов стационарных колебаний и транспортные решения ранее построены и изучены автором в вышеупомянутых работах. Здесь построим вибротранспортные решения БОУД (6).

3. Вибротранспортное уравнение Дирака и его решение

Рассмотрим случай, когда правая часть (6) имеет вид:

$$\mathbf{F}(\tau, \mathbf{x}) = \mathbf{Q}(x, z)e^{i\omega\tau}, \tag{8}$$

$\mathbf{x} = \sum_{j=1}^{3} x_j e_j$, $x = (x_1, x_2)$, $z = x_3 - Vt = x_3 - M\tau$. Здесь введено число Маха $M = V / c$.

Бикватернион $\mathbf{F}(\tau,\mathbf{x})$ описывает движение излучателя в направлении оси X_3 со скоростьюV, которое вибрирует с частотой ω. В этом случае решение (10) будем строить в аналогичном виде $\mathrm{B}(x,z,\tau)=\mathbf{B}(x,z)e^{i\omega\tau}$. Назовем его *вибротранспортным* решением БУД. Тогда

$$\partial_\tau = M\partial_z + i\omega,\ \nabla=(\partial_1,\partial_2,\partial_z)\,, \tag{9}$$

и в подвижной системе координат (x_1,x_2,z) уравнение преобразуется к виду:

$$\begin{aligned}&\mathbf{D}^{\pm}_{M\omega}\mathbf{B}(x,z)=\mathbf{Q}(x,z),\\ &\mathbf{D}^{\pm}_{M\omega}\ \square\ \left\{M\partial_z+\left(m+i\omega\right)\pm i(\partial_1,\partial_2,\partial_z)\right\}=M\partial_z+\left(m+i\omega\right)\pm i\nabla.\end{aligned} \tag{10}$$

Будем называть это уравнение *вибротранспортным уравнением Дирака* (ВТУД).

Л е м м а 3. *Композиция взаимных операторов*

$$\mathbf{D}^{+}_{M\omega}\circ\mathbf{D}^{-}_{M\omega}=\mathbf{D}^{-}_{M\omega}\circ\mathbf{D}^{+}_{M\omega}=-\Delta_2-\mu^2\left(\partial_z\right)^2+2ia\partial_z-b^2,$$

где $\Delta_2=\partial_1^2+\partial_2^2$ *- двумерный лапласиан,* $\mu=\sqrt{\left|1-M^2\right|}$, $a=Mb$, $b=\omega+\rho$.

Д о к а з а т е л ь с т в о: Согласно Лемме 2, получим

$$\mathbf{D}^{+}_{M\omega}\circ\mathbf{D}^{-}_{M\omega}=\mathbf{D}^{-}_{M\omega}\circ\mathbf{D}^{+}_{M\omega}=\left\{M\partial_z+i\left(\rho+\omega\right)+i\nabla\right\}\circ\left\{M\partial_z+i\left(\rho+\omega\right)-i\nabla\right\}$$

$$=-\Delta+\left(M\partial_z\right)^2+2iM\left(\omega+\rho\right)\partial_z-\left(\rho+\omega\right)^2=-\Delta-\mu^2\partial_z^2+2a\partial_z-b^2.$$

Здесь используем введённые обозначения для a,b,μ. В результате получена формула леммы.

Т е о р е м а 1. *Общее решение вибротранспортного уравнения Дирака* (10) *можно представить в виде*

$$\mathbf{B}(x,z)=\mathbf{B}^0(x,z)+\mathbf{D}^{\mp}_{M\omega}\left(\mathbf{Q}*\psi\right), \tag{11}$$

где $\mathrm{B}^0(x,z)$ *решение однородного уравнения* (*при* $\mathrm{F}=0$), $\psi(x,z)$ *- фундаментальное решение уравнения*:

$$\left\{-\Delta_2-\mu^2\left(\partial_z\right)^2+2ia\partial_z-b^2\right\}\psi(x,z)=\delta(z)\delta(x). \tag{12}$$

Д о к а з а т е л ь с т в о. Подставим (11) в (10) и, используя лемму 3, а также свойство ассоциативности кватернионного умножения и свойства свертки с дельта-функцией, получим требуемое:

$$\mathbf{D}^{\pm}_{M\omega}\left\{\mathbf{B}^0+\mathbf{D}^{\mp}_{M\omega}\left(\mathbf{Q}*\psi\right)\right\}=\mathbf{D}^{\pm}_{M\omega}\mathbf{B}^0+\mathbf{D}^{\pm}_{M\omega}\mathbf{D}^{\mp}_{M\omega}\left(\mathbf{Q}*\psi\right)=\left\{-\Delta_2-\mu^2\partial_{zz}+2ia\partial_z-b^2\right\}\left(\mathbf{Q}*\psi\right)=$$

$$=\left\{-\Delta_2-\mu^2\partial_{zz}+2ia\partial_z-b^2\right\}\psi*\mathbf{Q}=\delta(z)\delta(x)*\mathbf{Q}=\mathbf{Q}$$

Осталось вычислить *скалярный потенциал* $\psi(x,z)$. Вид его зависит от знака $\mu^2=1-M^2$.

Скорость распространения волн в среде назовем *световой.* Эта

скорость является критической скоростью движения. Возможны три случая: *досветовая скорость* $M<1 \Rightarrow \mu^2 > 0$, *световая скорость* $M=1 \Rightarrow \mu^2 < 0$, и *сверхсветовая скорость* $M>1 \Rightarrow \mu^2 < 0$. В зависимости от нее меняется тип вибротранспортного уравнения (10): *эллиптический* при досветовой скорости, *параболический* при световой и *строго гиперболический* при сверхсветовой. Рассмотрим здесь досветовой.

Для построения решений используем преобразование Фурье обобщённых функций. Обозначим переменные Фурье (ξ,ζ), соответствующие (x,z). Для регулярных функций, достаточно быстро убывающих на бесконечности, прямое и обратное преобразование Фурье имеет вид:

$$\mathrm{F}^{-}[\overline{f}(\xi,\zeta)] = f(x,z) = \frac{1}{8\pi^3}\int_{R^3} \overline{f}(\xi,\zeta)\exp(-i((x,\xi)+z\zeta))d\xi_1 d\xi_2 d\zeta. \tag{13}$$

$$\mathrm{F}^{-}[\overline{f}(\xi,\zeta)] = f(x,z) = \frac{1}{8\pi^3}\int_{R^3} \overline{f}(\xi,\zeta)\exp(-i((x,\xi)+z\zeta))d\xi_1 d\xi_2 d\zeta. \tag{14}$$

Для сингулярных функций следует использовать определение преобразования Фурье в пространстве обобщённых функций (Владимиров, 1976, 1979).

Для восстановления оригинала используем свойство обратного преобразования Фурье функции $\overline{\varphi}(\zeta) \leftrightarrow \varphi(z)$ при линейном преобразовании координаты:

$$\mathrm{F}^{-1}\left[\overline{\varphi}(\alpha\zeta+\beta)\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{\varphi}(\alpha\zeta+\beta)e^{-iz\zeta}d\zeta = \frac{1}{2\pi\alpha}\int_{-\infty}^{\infty}\overline{\varphi}(\varsigma)e^{-iz\frac{\varsigma-\beta}{\alpha}}d\varsigma =$$

$$= \frac{e^{iz\beta/\alpha}}{2\pi\alpha}\int_{-\infty}^{\infty}\overline{\varphi}(\varsigma)e^{-i\varsigma\frac{z}{\alpha}}d\varsigma = \frac{e^{iz\beta/\alpha}}{2\pi\alpha}\varphi\left(\frac{z}{\alpha}\right)$$

Это свойство позволяет строит оригиналы функции $f(\zeta) = \overline{\varphi}(\alpha\zeta+\beta)$, если известен оригинал $\overline{\varphi}(\zeta)$.

4. Досветовой скалярный потенциал вибротранспортного уравнения Дирака

При $M<1$ скалярный потенциал удовлетворяет уравнению

$$\left\{-\Delta_2 - \mu^2\left(\partial_z\right)^2 + 2ia\partial_z - b^2\right\}\psi(x,z) = \delta(z)\delta(x). \tag{15}$$

а его т*рансформанта Фурье* --

$$\left\{\|\xi\|^2 + \mu^2\zeta^2 + 2a\zeta - b^2\right\}\overline{\psi}(\xi,\zeta) = 1.$$

Следовательно

$$\overline{\psi}(\xi,\zeta) = \frac{1}{\|\xi\|^2 + \mu^2\zeta^2 + 2a\zeta - b^2}. \tag{16}$$

Для построения оригинала преобразуем его к виду:

$$\bar{\psi}(\xi,\zeta)=\frac{1}{\|\xi\|^2+\mu^2\zeta^2+2a\zeta-b^2}=$$

$$=\frac{1}{\xi_1^2+\xi_1^2+(\mu\zeta+a/\mu)^2-b^2-(a/\mu)^2}=\frac{1}{\xi_1^2+\xi_1^2+(\mu\zeta+a/\mu)^2-b^2\left(1+(M/\mu)^2\right)}=$$

$$=\frac{1}{\xi_1^2+\xi_1^2+(\mu\zeta+a/\mu)^2-(b/\mu)^2}=\frac{1}{\xi_1^2+\xi_1^2+(\mu\zeta+a/\mu)^2-(b/\mu)^2}\Rightarrow$$

$$\bar{\psi}(\xi,\zeta)=\frac{1}{\xi_1^2+\xi_1^2+\varsigma^2-k^2},\quad \varsigma=\mu\zeta+a/\mu,\quad k=b/\mu. \tag{17}$$

Для восстановления оригинала используем свойство линейных преобразований в пространстве переменных Фурье:

$$\mathrm{F}\left[\bar{\psi}(\xi,\zeta)\right]=\frac{1}{8\pi^3}\int_{R^2}d\xi_1 d\xi_2\int_{-\infty}^{\infty}\frac{\exp(-i(\xi_1x_1+\xi_1x_2+z\zeta))}{\xi_1^2+\xi_2^2+(\mu\zeta+a/\mu)^2-k^2}d\zeta=$$

$$=\frac{1}{8\mu\pi^3}\int_{R^2}d\xi_1 d\xi_2\int_{-\infty}^{\infty}\frac{\exp(-i(\xi_1x_1+\xi_1x_2+z(\varsigma-a/\mu)/\mu))}{\xi_1^2+\xi_2^2+\varsigma^2-k^2}d\varsigma=$$

$$=\frac{e^{iza/\mu^2}}{8\mu\pi^3}\int_{R^2}d\xi_1 d\xi_2\int_{-\infty}^{\infty}\frac{\exp(-i(\xi_1x_1+\xi_1x_2+\varsigma z/\mu))}{\xi_1^2+\xi_2^2+\varsigma^2-k^2}d\varsigma=\frac{e^{iza/\mu^2}}{\mu}U(x_1,x_2,z/\mu)$$

Здесь $\zeta=(\varsigma-a/\mu)/\mu$. А под знаком интеграла стоит преобразование Фурье фундаментальных решений уравнения Гельмгольца:

$$\Delta U+k^2U+\delta(y)=0,\qquad y=(y_1,y_2,y_3)$$

Среди них, затухающие на бесконечности, имеют следующий вид (Владимиров, 1976, 1979): $U(y)=\dfrac{e^{\pm ik\|y\|}}{4\pi\|y\|}$

Причём условию излучения Зоммерфельда, с учётом временного множителя $e^{i\omega\tau}$, удовлетворяет лишь функция

$$U(y)e^{i\omega\tau}=\frac{e^{-i(k\|y\|-\omega\tau)}}{4\pi\|y\|}, \tag{18}$$

которая описывает сферические фазовые волны, исходящие из источника, сосредоточенного в точке y=0. Следовательно

$$\psi(x,z)=\frac{\mu e^{iza/\mu^2}}{\mu}\frac{e^{-ik\mu^{-1}\sqrt{z^2+(\mu r)^2}}}{4\pi\sqrt{z^2+(\mu r)^2}}=\frac{1}{4\pi\sqrt{z^2+(\mu r)^2}}\exp\left(i\left(zM-\sqrt{z^2+(\mu r)^2}\right)b/\mu^2\right)=$$

$$=\frac{1}{4\pi\sqrt{z^2+(\mu r)^2}}\exp\left(i(\omega+\rho)\alpha(z,r)\right),$$

где $\alpha(z,r)=\left(zM-\sqrt{z^2+(\mu r)^2}\right)/\mu^2$. Следовательно,

$$\psi(x,z)e^{i\omega\tau}=\frac{e^{-i\rho\alpha_-(z,r)}}{4\pi\sqrt{z^2+(\mu r)^2}}\exp\left(i\omega\left(\tau-zM/\mu^2-\sqrt{r^2+(z/\mu)^2}\right)\right)=$$

$$=\frac{e^{i\omega\left(\tau-\sqrt{r^2+(z/\mu)^2}\right)}}{4\pi\sqrt{z^2+(\mu r)^2}}\exp\left(i\left(\rho\sqrt{r^2+(z/\mu)^2}-(\rho+\omega)zM/\mu^2\right)\right). \qquad (19)$$

Определим фазовую поверхность этой волны:

$$\alpha(z,r)(\rho+\omega)=C-\omega\tau,$$

$$zM-\sqrt{z^2+(\mu r)^2}=\mu^2\frac{C-\omega\tau}{\rho+\omega}=f(\tau)\Rightarrow$$

Вычислим

$$z^2+(\mu r)^2=f^2(t)-2zMf(\tau)+(zM)^2,$$

$$(z)^2-(zM)^2+2zMf(\tau)+(\mu r)^2=f^2(\tau),\quad (\mu z)^2+2zMf(\tau)+(\mu r)^2=f^2(\tau),$$

$$(\mu z+Mf(\tau)/\mu)^2+(\mu r)^2=f^2(\tau)+(Mf(\tau)/\mu)^2=\left(\frac{f(\tau)}{\mu}\right)^2,$$

$$\left(z+Mf(\tau)/\mu^2\right)^2+r^2=\frac{f^2(\tau)}{\mu^3}=\mu\left(\frac{C-\omega\tau}{\rho+\omega}\right)^2.$$

В результате получим

$$\left(z+\frac{M}{\mu^2}\left(\frac{C-\omega\tau}{\rho+\omega}\right)\right)^2+r^2=\mu\left(\frac{C-\omega\tau}{\rho+\omega}\right)^2.$$

Это сфера радиуса $\sqrt{\mu}\left|\frac{C-\omega\tau}{\rho+\omega}\right|$ с центром в подвижной точке

$$X^*=\left\{0,0,z=\frac{M(\omega\tau-C)}{\mu^2(\rho+\omega)}\right\}. \qquad (20)$$

Здесь C – произвольная действительная константа.

5. Досветовой волновой потенциал однородного вибротранспортного уравнения Дирака

Построим теперь решения однородного вибротранспортного уравнения Дирака:

$$\mathbf{D}^{\pm}_{M\omega}\mathbf{B}^0(x,z)=0, \qquad (21)$$

Отсюда следует, что

$$\mathbf{D}^{\mp}_{M\omega}\mathbf{D}^{\pm}_{M\omega}\mathbf{B}^0 = \left(-\Delta_2 - \mu^2\left(\partial_z\right)^2 + 2ia\partial_z - b^2\right)\mathbf{B}^0 = 0\,,$$

$\mu = \sqrt{1-M^2}, \quad a = Mb, \quad b = \omega - im$.

Соответствующий скалярный потенциал $\psi_0(x,z)$ и является решением однородного вибротранспортного уравнения:

$$\left(\Delta_2 + \mu^2\partial_z\partial_z - 2ia\partial_z + b^2\right)\psi_0(x,z) = 0\,, \tag{22}$$

преобразование Фурье которого имеет вид:

$$\left(\|\xi\|^2 + \mu^2\zeta^2 - 2a\zeta - b^2\right)\overline{\psi}_0(\xi,\zeta) = 0.$$

Следовательно $\overline{\psi}_0 = \beta(\xi,\zeta)\delta_S(\xi,\zeta)$ - простой слой на поверхности S:

$$S = \left\{(\xi,\zeta) : \|\xi\|^2 + \mu^2\left(\zeta - a/\mu\right)^2 = a^2 + b^2\right\}. \tag{23}$$

Легко видеть, что это осесимметричный эллипс с центром в точке $(\xi,\zeta) = \left(0, 0, \dfrac{a}{\mu}\right)$. Оригинал определяется интегралом по поверхности этого эллипса:

$$\psi_0(x,z) = \int_S \beta(\xi,\zeta)e^{-i(x,\xi)}e^{-iz\zeta}\,dS(\xi,\zeta)\,, \tag{24}$$

где $\beta(\xi,\zeta)$ - произвольная интегрируемая на S функция.

Для построения $\psi_0(x,z)$ можно также использовать решения однородного уравнения Гельмгольца:

$$\Delta u + k^2 u = 0,$$

которые можно разложить в ряды по сферическим гармоникам и сферическим функциям Бесселя:

$$u(y) = \sum_{n,m} a_n J_n(k\|y\|)P_n^m(\cos\theta)e^{im\varphi} = \sum_{n,m} a_n J_n(k\|y\|)P_n^m\left(\frac{y_3}{\|y\|}\right)(\cos\varphi + i\sin\varphi)^m =$$

$$= \sum_{n,m} a_n \frac{J_n(k\|y\|)}{\|y\|_2^m} P_n^m\left(\frac{y_3}{\|y\|}\right)(y_1 + iy_2)^m, \qquad \|y\|_2 = \sqrt{y_1^2 + y_2^2}. \tag{25}$$

Здесь $P_n^m(\cos\theta)$

- присоединенные полиномы Лежандра, θ,φ угловые сферические координаты. Используя свойства сдвига преобразования Фурье и сжатия-растяжения по координатным осям, в результате из (30) получим оригинал этого вибротранспортного волнового потенциала:

$$\psi_0(x,z) = e^{ia/\mu}\sum_{n,m} a_n J_n\left(\frac{b}{\mu}\sqrt{z^2 + \mu^2 r^2}\right)P_n^m\left(\frac{z}{\sqrt{z^2 + \mu^2 r^2}}\right)e^{im\varphi}, \qquad r = \sqrt{x_1^2 + x_2^2}\,. \tag{26}$$

Итак, в формуле (11) теоремы 1 все функции определены и общее решение вибротранспортного уравнения Дирака при досветовых скоростях движения источника построено.

Бикватернион энергии-импульса определяется формулой

$$\mathbf{\Sigma}(x,z)=0,5\mathbf{B}(x,z)\circ\mathbf{B}^{*}(x,z)=$$

$$=0,5\big(b(x,z)+B(x,z)\big)\circ\Big(\bar{b}(x,z)-\bar{B}(x,z)\Big)=w(x,z)+iP(x,z). \qquad (27)$$

Здесь $w(x,z), P(x,z)$ - плотность энергии и аналог вектора Пойнтинга, который показывает направление ее распространения.

9. **Бифункция Грина вибротранспортного уравнения Дирака**

Общее решение ВТУД можно записать в более удобном для вычисления виде, если использовать бикватернионную функцию Грина.

О п р е д е л е н и е. *Бифункцией Грина* вибротранспортного уравнения Дирака $\mathbf{U}^{\pm}(x,z)$ называется решение уравнения (10) при $\mathbf{Q}(x,z)=\delta(x)\delta(z)$:
$\mathbf{D}_{\mathrm{v}}^{\pm}\mathbf{U}^{\pm}(x,z)=\delta(x)\delta(z)$,

удовлетворяющее условиям затухания

$\mathbf{U}^{\pm}(x,z)\to 0$ при $\|(x,z)\|\to\infty$

и условиям излучения Зоммерфельда на бесконечности.

Из теоремы 1 следует, при $\mathbf{F}=\delta(x)\delta(z)$

$$\mathbf{U}^{\pm}(x,z)=\mathbf{D}_{M\omega}^{\mp}\psi(x,z)=$$

$$=\left\{M\partial_z+i\left(\rho+\omega\right)\pm i\nabla\right\}\psi=i\left(\rho+\omega\right)\psi+M\partial_z\psi\pm i\,\mathrm{grad}\,\psi. \qquad (28)$$

Очевидно, что бифункция Грина удовлетворяет условию затухания (33), в силу свойств потенциала $\psi(x,z)$ и его производных.

Т е о р е м а 2. *Частное решение вибротранспортного уравнения Дирака при досветовых скоростях движения* (M<1) *можно представить в виде бикватернионной свертки*

$$\mathbf{B}(x,z)=\mathbf{U}^{\pm}(x,z)*\mathbf{Q}(x,z), \qquad (29)$$

которую для регулярных $\mathbf{Q}(x,z)$ *можно представить в интегральном виде*

$$\mathbf{B}(x,z)=\int_{-\infty}^{\infty}d\zeta\iint_{R^2}\mathbf{U}^{\pm}(x-y,z-\zeta)\circ\mathbf{Q}(y,\zeta)dy_1dy_2.$$

Решение существует, если $\mathbf{Q}(x,z)\in L_1(R^3)$.

Если правая часть (10) *сингулярный бикватернион с носителем на поверхности S или кривой l, например*

$$\mathbf{Q}(x,z)=\mathbf{A}_S(x,z)\delta_S(x,z)\mathbf{Q}(x,z)=\mathbf{A}_L(x,z)\delta_L(x,z),$$

то $\mathbf{B}(x,z)=\int_S\mathbf{U}^{\pm}(x-y,z-\zeta)\circ\mathbf{Q}(y,\zeta)dS(y_1,y_2,\zeta),$

$$\mathbf{B}(x,z)=\int_l \mathbf{U}^{\pm}(x-y,z-\zeta)\circ\mathbf{Q}(y,\zeta)dl(y_1,y_2,\zeta),$$

где интегралы берутся по носителю бикватерниона, т.е. поверхностный и криволинейный.

Заключение. Решения уравнений Дирака в теоретической физике принято называть *спинорами*. Соответственно назвать *биспинорами* их бикватернионное представление.

Формула теоремы 2 даёт решение ВТУД при любых $\mathbf{Q}(x,z)$ из класса сингулярных бикватернионов, в том числе описываемых сингулярными обобщёнными функциями, дельта – функциями и их производными, которые используют для описания движущихся зарядов, диполей, мультиполей и элементарных частиц. Для таких источников излучения следует использовать правила вычисления свёрток обобщённых функций (Владимиров, 1979).

При $\omega=0$ полученные решения описывают транспортные решения уравнений Дирака. При $M=0,\ \ \omega>0$ эти же формулы описывают процесс стационарных колебаний с фиксированной частотой и могут быть использованы для построения периодических по времени решений уравнений Дирака. При $\rho=0$ эти решения описывают решения бикватернионной формы уравнений Максвелла, которые можно использовать для для исследования электромагнитных полей различных световых излучателей и излучателей радиоволн, расположенных на подвижных объектах (поездах, машинах, кораблях и т.п.).

Отметим, что построенная здесь бифункция Грина необходима для решения вибротранспортных краевых задач в областях, ограниченных цилиндрическими поверхностями, по которым движутся излучатели волн в направлении их образующих.

Литература

Максвелл Дж. К. (1989) Трактат об электричестве и магнетизме. — Т. 1, 2. Москва: Наука.

Дирак П.А.М.(1979) Принципы квантовой механики. — Москва: Наука.

Джексон Дж. (1965) Классическая электродинамика. — Москва: Мир.

Фейнман Р., Лейтон Р., Сэндс М. (1965) Фейнмановские лекции по физике. — Т. 5. Электричество и магнетизм. - Москва: Мир.

Ландау Л.Д., Лифшиц Е.М. (2003) Теория поля. Теоретическая физика. —Т. 2. — Москва: Физматлит.

Савельев И.В. (1970) Курс общей физики. Т. 2. Электричество. — Москва: Наука.

Rodrigues W. A . , Capelas de Oliviera E. (1990) Dirac and Maxwell equations in the Clifford and spin-Clifford bundles. Int. Journal of Theoretical Physics. — V.-29. — P. 397–412.

Finkelstein D., Jauch J. M., Schiminovich S., Speiser D. (1992) Foundations of quaternion quantum mechanics. J. Math. Phys., 3. — P. 207–220.

Adler S. L.(1995) Quaternionic quantum mechanics and quantum fields. — New York: Oxford University Press.

De Leo S., Rodrigues Jr.W.A. (1997) Quaternionic quantum mechanics: from complex to complexified quaternions. Int. J. Theor. Phys. — V. 36. — P. 2725–2757.

Ефремов А.П. (2004) Кватерноны: алгебра, геометрия и физические теории. Гиперкомплексные числа в геометрии и физике. — Т. 1. — №1. — С. 111-127.

Acevedo M., Lopez-Bonilla J., Sanchez-Meraz M. (2005) Quaternions, Maxwell Equations and Lorentz Transformations. Apeiron. — V.12. — No. 4. — P. 371-376.

Марчук Н.Г. (2009) Уравнения теории поля и алгебры Клиффорда. — Москва-Ижевск.

Alexeyeva L.A. (2012) Biquaternions algebra and its applications by solving of some theoretical physics equations. Clifford Analysis, Clifford Algebras and their Applications. — V. 7. — No 1. —P. 19-39.

Alexeyeva L.A. (2021) Biquaternionic Wave Equations and the Properties of Their Generalized Solutions Differential Equations. —V. 57.-No 5. — P.594-604. Doi: 10.1134/S0012266121050049

Алексеева Л.А., Азиз Г.Н. (2024) Транспортные решения уравнений Максвелла в бикватернионном представлении при досветовых скоростях. Журнал проблем эволюции открытых систем. — Т. 26. — №.1. — С. 64-73. doi.org/10.26577/JPEOS.2024.v.26.il-i6

Владимиров В.С. (1976) Уравнения математической физики. — Москва: Наука.

Владимиров В.С. (1979) Обобщенные функции в математической физике. — Москва: Наука.

References

Maksvell Dzh. K. (1989) Traktat ob elektrichestve i magnetizme [Treatise on Electricity and Magnetism]. — V.1, 2. Moskva: Nauka (in Russ).

Dirak P.A.M. (1979) Printsipy kvantovoy mekhaniki [Principles of Quantum Mechanics]. — Moskva: Nauka (in Russ).

Dzhekson Dzh. (1965) Klassicheskaya elektrodinamika [Classical Electrodynamics]. — Moskva: Mir (in Russ).

Feynman R., Leyton R., Sends M. (1965) Feynmanovskiye lektsii po fizike [The Feynman Lectures on Physics]. —T.5. Moskva: Mir (in Russ).

Landau L.D., Lifshits Ye.M. (2003) Teoriya polya. Teoreticheskaya fizika [Field Theory. Theoretical Physics]. —T.2. Moskva: Fizmatlit (in Russ).

Savel'yev I.V. (1970) Kurs obshchey fiziki [Course of General Physics]. V.2. Elektrichestvo. — Moskva: Nauka (in Russ).

Rodrigues W.A. , Capelas de Oliviera E. (1990) Dirac and Maxwell equations in the Clifford and spin-Clifford bundles. Int. Journal of Theoretical Physics, 29. — P. 397–412 (in Engl).

Finkelstein D., Jauch J. M., Schiminovich S., Speiser D. (1992) Foundations of quaternion quantum mechanic. J. Math. Phys., 3. — P. 207–220 (in Engl).

Adler S. L. (1995) Quaternionic quantum mechanics and quantum fields. — New York: Oxford University Press (in Engl).

De Leo S., Rodrigues Jr. W.A. (1997) Quaternionic quantum mechanics: from complex to complexified quaternions. Int. J. Theor. Phys. 36. — P. 2725–2757 (in Engl).

Yefremov A.P. (2004) Kvaternony: algebra, geometriya i fizicheskiye teorii [Quaternions: Algebra, Geometry, and Physical Theories]. Giperkompleksnyye chisla v geometrii i fizike. Hypercomplex Numbers in Geometry and Physics, 1, No. 1. — P. 111–127 (in Russ).

Acevedo M., Lopez-Bonilla J., Sanchez-Meraz M. (2005) Quaternions, Maxwell Equations and Lorentz Transformations. Apeiron, 12, no. 4. — P.371–376 (in Engl).

Marchuk N.G. (2009) Uravneniya teorii polya i algebry Klifforda [Field Theory Equations and Clifford Algebras]. Moskva–Izhevsk (in Russ).

Alexeyeva L.A. (2012) Biquaternions algebra and its applications by solving of some theoretical physics equations.Clifford Analysis, Clifford Algebras and their Applications, v. 7, no. 1. — P. 19-39 (in Engl).

Alexeyeva L.A. (2021) Biquaternionic Wave Equations and the Properties of Their Generalized Solutions Differential Equations, 57, no 5. — P. 594-604. Doi: 10.1134/S0012266121050049 (in Engl).

Alekseyeva L.A., Aziz G.N. (2024) Transportnyye resheniya uravneniy Maksvella v bikvaternionnom predstavlenii pri dosvetovykh skorostyakh [Transport solutions of Maxwell's

equations in the biquaternion representation at sublight speeds]. Zhurnal problem evolyutsii otkrytykh sistem (in Russ). doi.org/10.26577/JPEOS.2024.v.26.il-i6 (in Russ).

Vladimirov V.S. (1976) Uravneniya matematicheskoy fiziki [Equations of Mathematical Physics]. — Moskva: Nauka, (in Russ).

Vladimirov V.S. (1979) Obobshchennyye funktsii v matematicheskoy fizike [Generalized Functions in Mathematical Physics]. — Moskva: Nauka (in Russ).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 41–51

https://doi.org/10.32014/2025.2518-1726.362

ҒТМР 81.93.29
ӘОЖ 004.056.5

A. Amirova*, B. Aldosh, A. Ibraikhan, T. Smagulov, A. Aitmagambet, 2025.
Astana IT University, Astana, Kazakhstan.
E-mail: Akzhibek.amirova@astanait.edu.kz

A MACHINE LEARNING-BASED APPROACH TO DETECT MALICIOUS LINKS ON INSTAGRAM

Amirova Akzhibek — PhD, Assistant Professor, Astana IT University, Astana, Kazakhstan, E-mail: akzhibek.amirova@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-5715-4954;
Aldosh Balziya — MSc, Senior Lecturer, Astana IT University, Astana, Kazakhstan, E-mail: b.aldosh@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-2531-9718;
Alinur Ibraikhan — Student, Astana IT University, Astana, Kazakhstan, E-mail: 221596@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0009-3929-7378;
Temirlan Smagulov — Student, Astana IT University, Astana, Kazakhstan, E-mail: 221278@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0001-9039-3594;
Aysultan Aitmagambet — Student, Astana IT University, Astana, Kazakhstan, E-mail: 220920@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0008-2158-4234.

Abstract. With the rapid development of social networks and their integration into everyday life, platforms such as Instagram are becoming increasingly vulnerable to cyberattacks. One of the most common and dangerous vectors is the spread of malicious links. This study focuses on identifying and mitigating threats associated with the placement of harmful URLs in Instagram users' biographies, direct messages, and comments. The authors emphasize that traditional filtering methods, such as blocking or URL matching, are insufficient, since attackers actively use social engineering to disguise the true purpose of their links. To address this issue, a hybrid detection system is proposed that combines machine learning methods (Random Forest, LightGBM, XGBoost) with heuristic analysis. This approach enables a more comprehensive evaluation of suspicious content and significantly improves detection performance. Experimental results showed that the system reached 98% accuracy in classifying suspicious links. It was implemented as a browser extension, allowing users to promptly identify and flag potential threats, which demonstrates its practical value. Although the current version requires local installation, future work will focus on integrating deep learning techniques and incorporating contextual information to further increase automation and precision.

The proposed approach thus makes an important contribution to the development of cybersecurity methods for social networks and can serve as a foundation for scalable threat monitoring systems.

Keywords: Malicious link detection, machine learning, cybersecurity, real-time analytics, Instagram security, URL classification

А. Амирова*, Б. Альдош, А. Ибраихан, Т. Смагулов, А. Айтмагамбет, 2025.
Astana IT University, Астана, Қазақстан.
E-mail: Akzhibek.amirova@astanait.edu.kz

INSTAGRAMДАҒЫ ЗИЯНДЫ СІЛТЕМЕЛЕРДІ АНЫҚТАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУҒА НЕГІЗДЕЛГЕН ТӘСІЛ

Амирова Акжибек — PhD, ассистент профессор, Astana IT University, Астана, Қазақстан,
E-mail: akzhibek.amirova@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-5715-4954;
Альдош Балзия — магистр, аға оқытушы, Astana IT University, Астана, Қазақстан,
E-mail: b.aldosh@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-2531-9718;
Алинур Ибраихан — студент, Astana IT University, Астана, Қазақстан,
E-mail: 221596@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0009-3929-7378;
Темирлан Смагулов — студент, Astana IT University, Астана, Қазақстан,
E-mail: 221278@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0001-9039-3594;
Айсултан Айтмагамбет — студент, Astana IT University, Астана, Қазақстан,
E-mail: 220920@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0008-2158-4234.

Аннотация. Әлеуметтік желілердің қарқынды дамуымен және олардың күнделікті өмірге енуімен Instagram сияқты платформалар кибершабуылдарға осал бола бастады. Ең көп таралған және қауіпті векторлардың бірі – зиянды сілтемелердің таралуы. Бұл зерттеу Instagram қолданушыларының өмірбаянында, тікелей хабарламаларында және түсініктемелерінде зиянды Url Мекенжайларын орналастырумен байланысты қауіптерді анықтауға және азайтуға бағытталған. Авторлар URL мекен-жайларын бұғаттау немесе сәйкестендіру сияқты дәстүрлі сүзгілеу әдістері жеткіліксіз екенін атап көрсетеді, өйткені шабуылдаушылар өздерінің сілтемелерінің шынайы мақсатын жасыру үшін әлеуметтік инженерияны белсенді қолданады. Бұл мәселені шешу үшін машиналық оқыту әдістерін (Random Forest, LightGBM, XGBoost) эвристикалық талдаумен біріктіретін гибридті анықтау жүйесі ұсынылады. Бұл тәсіл күдікті мазмұнды жан-жақты бағалауға мүмкіндік береді және анықтау өнімділігін айтарлықтай жақсартады. Эксперименттік нәтижелер жүйенің күдікті сілтемелерді жіктеуде 98% дәлдікке жеткенін көрсетті. Ол пайдаланушыларға ықтимал қауіптерді дереу анықтауға және белгілеуге мүмкіндік беретін шолғыш кеңейтімі ретінде енгізілді, бұл оның

практикалық құндылығын көрсетеді. Ағымдағы нұсқа жергілікті орнатуды қажет етсе де, болашақ жұмыс автоматтандыру мен дәлдікті одан әрі арттыру үшін терең оқыту әдістерін біріктіруге және контекстік ақпаратты енгізуге бағытталады. Осылайша, ұсынылған тәсіл әлеуметтік желілер үшін киберқауіпсіздік әдістерін дамытуға маңызды үлес қосады және қауіптерді бақылаудың ауқымды жүйелерінің негізі бола алады.

Түйін сөздер. Зиянды сілтемелерді анықтау, машиналық оқыту, киберқауіпсіздік, нақты уақыттағы талдау, Instagram қауіпсіздігі, URL классификациясы

А. Амирова*, Б. Альдош, А. Ибраихан, Т. Смагулов, А. Айтмагамбет, 2025.
Astana IT University, Астана, Казахстан.
E-mail: Akzhibek.amirova@astanait.edu.kz

ПОДХОД НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ ВРЕДОНОСНЫХ ССЫЛОК В INSTAGRAM

Амирова Акжибек — PhD, ассистент профессор, Astana IT University, Астана, Казахстан,
E-mail: akzhibek.amirova@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-5715-4954.
Альдош Балзия — магистр, старший преподаватель, Astana IT University, Астана, Казахстан,
E-mail: b.aldosh@astanait.edu.kz, ORCID ID: https://orcid.org/0000-0002-2531-9718;
Алинур Ибраихан — студент, Astana IT University, Астана, Казахстан,
E-mail: 221596@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0009-3929-7378;
Темирлан Смагулов — студент, Astana IT University, Астана, Казахстан,
E-mail: 221278@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0001-9039-3594;
Айсултан Айтмагамбет — студент, Astana IT University, Астана, Казахстан,
E-mail: 220920@astanait.edu.kz, ORCID ID: https://orcid.org/0009-0008-2158-4234.

Аннотация. С развитием социальных сетей и их глубокой интеграцией в повседневную жизнь такие широко используемые платформы, как Instagram, становятся все более уязвимыми для кибератак. Особенно в последние годы использование вредоносных ссылок превратилось в один из наиболее распространённых и опасных векторов атак, создавая серьёзные угрозы личным данным пользователей, их финансовой безопасности и целостности информационных систем. Подобные атаки часто направлены на обман пользователей с целью получения доступа к их аккаунтам, хищения конфиденциальной информации или распространения вредоносного программного обеспечения. Данное исследование ориентировано на выявление и снижение рисков, связанных с размещением вредоносных URL-адресов в биографиях, личных сообщениях и комментариях пользователей Instagram. Авторы подчёркивают, что традиционные методы фильтрации, такие

как блокировка или сопоставление URL, оказываются недостаточными. Это связано с тем, что злоумышленники активно используют методы социальной инженерии, маскируя истинное назначение ссылок и вводя пользователей в заблуждение. Для решения этой проблемы предлагается гибридная система обнаружения, объединяющая методы машинного обучения (Random Forest, LightGBM, XGBoost) с эвристическим анализом. Система выполняет комплексный анализ различных параметров, классифицирует подозрительные ссылки и достигает высокой эффективности. Экспериментальные результаты показали, что точность классификации достигает 98%. Реализация в формате расширения для браузера позволяет пользователям быстро выявлять и отмечать потенциальные угрозы. В настоящее время система требует локальной установки, однако в дальнейшем планируется внедрение методов глубокого обучения, использование контекстной информации и полная автоматизация процессов. Таким образом, предложенный подход вносит значительный вклад в обеспечение кибербезопасности в социальных сетях и может стать прочной основой для построения масштабируемых систем мониторинга угроз.

Ключевые слова: обнаружение вредоносных ссылок, машинное обучение, кибербезопасность, аналитика в реальном времени, безопасность Instagram, классификация URL

Кіріспе. Заманауи коммуникациялық технологиялар адамдар арасындағы қарым-қатынастарды, сондай-ақ таратылатын ақпарат пен коммерциялық әрекеттерді түбегейлі өзгертті. Қазіргі қоғам әлеуметтік медиа платформаларына қатты тәуелді, онда Instagram өзін олардың арасында әлемдік көшбасшы ретінде көрсетеді. Платформа 2 миллиардтан астам белсенді пайдаланушыларды қабылдайтындықтан, ол пайдаланушыларға іскерлік және жеке қажеттіліктерді қанағаттандыра отырып, нақты уақытта алмасуға және мазмұнды және тікелей хабарламаны бөлісуге мүмкіндік беретін қуатты платформа ретінде жұмыс істейді (Statista, 2024). Әлеуметтік медианың кеңеюі киберқауіптердің пропорционалды өсуіне әкелді, өйткені қылмыскерлер алаяқтық схемаларды жүргізу үшін платформаларды пайдаланады (Alharbi et al., 2024; Sheikhi, 2020). Instagram желісінде таратылатын зиянды сілтемелер пайдаланушылар жиі кездесетін және қауіпті киберқауіптердің бірі болып табылады. Бұл гиперсілтемелер түсініктемелер, DM және био және ақылы жарнамалар арқылы фишингтік алаяқтық, зиянды бағдарламаларды жіберу және қаржылық алаяқтық үшін кіру нүктесі ретінде әрекет етеді (Meshram et al., 2021).

Сандық қауіптер стандартты қауіпсіздік жүйелерін күн сайын тиімділігін төмендететін деңгейге дейін дамыды. Қара тізімдер және ережеге негізделген сүзгілеу сияқты зиянды сілтемелерді анықтау және блоктаудың дәстүрлі әдістері дамып келе жатқан шабуыл стратегияларына ілесу үшін күреседі. Киберқылмыскерлер өздерінің сілтемелерін әртүрлі әдістер арқылы анықтауды қиындатады, соның ішінде URL қысқартқыштары доменді өзгерту және

динамикалық қайта бағыттау тізбегі (Pradeep et al., 2023; Mughaid et al., 2023). Зиянды бағдарламалық қамтамасыз етуді таратушылар жаңылыстыратын байланыс және жалған жүлде жарнамалары немесе жеке басын куәландыратын жалған алаяқтық сияқты алдау әдістері арқылы пайдаланушыларды алдау үшін әлеуметтік инженерия әдістерін пайдаланады (Aljabri et al., 2023; Caruccio et al., 2023). Жағдай анықтаудан жалтаруға тырысатын зиянды сілтемелерді анықтау арқылы қорғаныс қызметін атқаратын жетілдірілген интеллектуалды жүйені талап етеді.

Инстаграмдағы зиянды сілтемелерді тарату жеке пайдаланушыларға қарағанда көбірек әсер ететін әсерлер жасайды. Инстаграмды маркетинг мақсаттары үшін, сонымен қатар брендинг мақсаттары мен тұтынушылармен өзара әрекеттесу үшін қолданатын ықпал етушілермен және ұйымдармен бірге компаниялар әлеуетті киберқауіптерге тап болады (Raja et al., 2021; Kaushik et al., 2022). Жалған өнімді жылжыту және фишингтік шабуылдармен бірге брендке еліктеу схемасы шынайы бизнеске елеулі бедел мен қаржылық зиян келтіреді (Durga et al., 2023). Платформа деңгейіндегі осалдықтарды азайта отырып, қауіптерді анықтау үшін толық қауіпсіздік жүйесін қажет етеді. Зерттеулер қауіпті азайту шешімдерін әзірлеу кезінде анықталған зиянды сілтеме қауіптерін Instagram үшін сенімді анықтау жүйесін құру үшін осы жұмысты жүзеге асырады. Зерттеу миллиондаған әлеуметтік медиа қолданушыларын бүкіл әлем бойынша қорғауды қамтамасыз ету үшін жаңа кибершабуыл жүйелеріне белсенді түрде бейімделетін қауіпсіздік негізін жасау үшін жетілдірілген алгоритмдерді енгізеді (Durga et al., 2023).

Ұсынылған зерттеу Instagram-дағы URL мекенжайларын анықтау үшін статикалық қара тізімдердің немесе ережеге негізделген анықтау әдістерінің орнына оқытуға негізделген тәсілді қолданады (Salamh et al., 2021; Nobili et al., 2023). Бірнеше киберқауіпсіздік зерттеулері қара тізімге негізделген URL сүзгісінің негізгі әлсіз жақтарын сипаттайды, себебі шабуылдаушылар анықтау шараларын айналып өту үшін URL қысқартқыштары мен динамикалық домен жасау әдістерін пайдаланады.

Материалдар мен әдістер. Инстаграмдағы зиянды сілтемелерді анықтауға көмектесу үшін машиналық оқыту және эвристикалық әдістер біріктірілді. Бұл әдістеме белгілі шабуыл үлгілерін пайдалана отырып, жаңа қауіптерді анықтауға қабілетті сенімді және бейімделгіш жүйені қамтамасыз етеді. Әзірленген алгоритм Random Forest, XGBoost және LightGBM көмегімен URL мекенжайларын зиянсыз, бүліну, фишинг және зиянды бағдарламалар санаттарына жіктеді. Жалпы фишинг сипаттамаларына негізделген URL мекенжайларының күдікті сипатын бағалау үшін қолмен жасалған ережелер жиынтығы қолданылды (Alsharida et al., 2023; Prince et al., 2024). Эвристикалық ережелерді машиналық оқытуға қосымша қадамдар ретінде қолдану дәлдікті және жаңа деректердегі үлгілерді анықтау мүмкіндігін жақсартты.

Функция таңдау

Машинамен оқытудың тиімді болуы үшін URL мекенжайларынан пайдалы

мүмкіндіктерді табуға болады (сурет). Бұл ерекшеліктерді келесідей жіктеуге болады:

Лексикалық мүмкіндіктер (URL құрылымы)

− URL ұзындығы: Ұзын URL мекенжайлары күдіктірек болады.

− Арнайы таңбалар саны: @, %, -, _, = және? нормадан тыс мөлшерде.

− Ішкі домендер саны: ішкі домендер тым көп (example.phishing.attack.com).

− Сандық қатынас: домендегі цифрлардың жоғары саны (мысалы, paypal123.com).

− IP мекенжайының болуы: өңделмеген IP мекенжайлары бар URL мекенжайлары (http://192.168.1.1).

− TLD талдауы: Сирек емес TLD (мысалы, .xyz, .tk) зиянды болуы мүмкін. Хост негізіндегі мүмкіндіктер (домен талдауы)

− WHOIS деректері: доменді тіркеу мәліметтерін (жасы, тіркеуші, жасалған күні) тексереді.

− Домен жасы: Жақында тіркелген домендер жиі зиянды.

− Танымалдық: Alexa, Majestic немесе Google Safe Browsing көмегімен домен рейтингін тексереді.

− SSL сертификатының болуы: зиянды домендерде HTTPS шифрлауы жиі болмайды.

Мазмұнға негізделген мүмкіндіктер

− HTML және JavaScript мүмкіндіктері: түсініксіз JavaScript, iframes және күдікті қайта бағыттауларды іздейді. 36

− Қайта бағыттаулардың болуы: тым көп қайта бағыттау фишинг немесе зиянды бағдарлама сайттарын көрсетуі мүмкін.

− Енгізілген сілтемелер: веб-беттегі сілтемелерді тексереді. Желіге негізделген мүмкіндіктер

− DNS ақпараты: доменнің IP мекенжайларын жиі өзгертетінін тексереді (жылдам ағын).

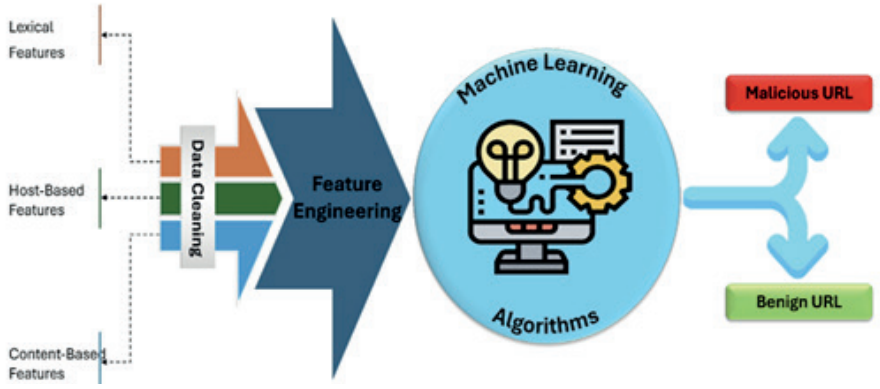
− PTR жазбалары: заңдылықты тексеру үшін кері DNS іздеулері.

− Хостинг туралы ақпарат: зиянды мазмұнмен белгілі хостинг провайдерлерін анықтайды.

Анықтау алгоритмінің жұмыс процесі

Машиналық оқыту алгоритмдерінің кірістері сандар болғандықтан, лексикалық сандық мүмкіндіктер файлдардағы URL мекенжайларынан жасалады (Сурет 1). Осылайша, машиналық оқыту алгоритмдеріне кіріс нақты өңделмеген URL мекенжайларынан гөрі сандық лексикалық мүмкіндіктер болады.

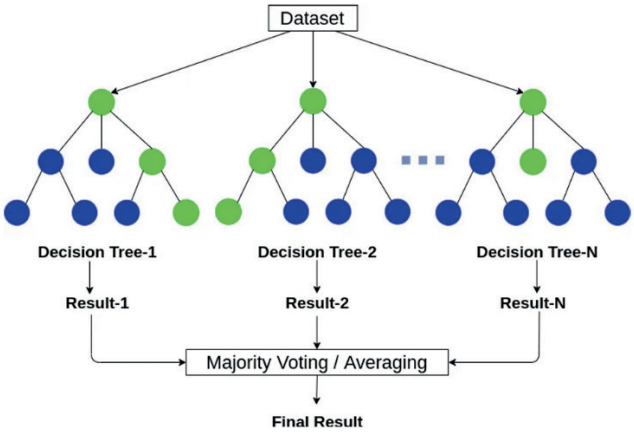
1-сурет. Машиналық оқыту моделі



Сәйкес келетін алгоритмді таңдау

Бұл жұмыс дәстүрлі машиналық оқыту үлгілерімен жақсы орындалады. Ең күшті алгоритмдерді табу үшін талдау дәлдік, дәлдік, еске түсіру және F1 балл негізінде жүргізіледі. Кездейсоқ орман, LightGBM және XGBoost сияқты машиналық оқытудың үш моделі олардың қалай жұмыс істейтінін білу үшін қарастырылады (сурет 2).

2-сурет. Кездейсоқ орман үлгісі



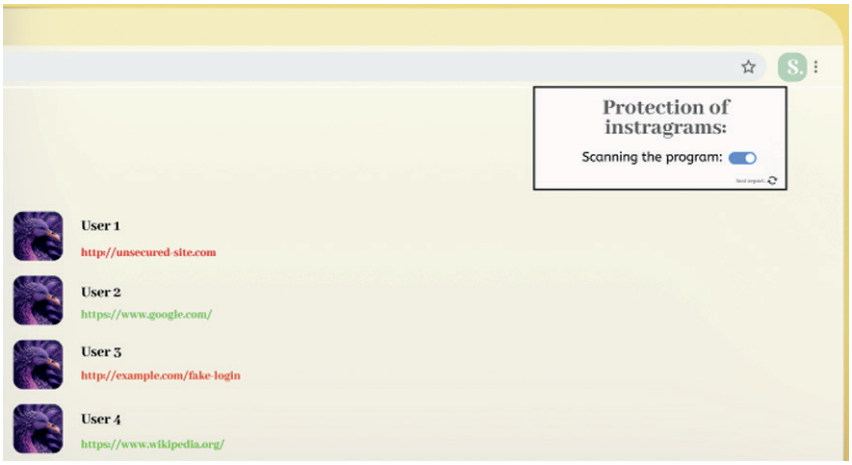
– Деректерді жинау және алдын ала өңдеу 660000-нан астам URL мекенжайлары бар деректер жиынын Firefox қамтамасыз етті. Деректер жинағы жүктелетін, талданатын және стратификацияланған іріктеу арқылы оқу және сынақ жиындарына бөлінген белгіленген URL мекенжайларын қамтиды.

– Функция инженериясы Әр түрлі URL негізіндегі мүмкіндіктер Python көмегімен бағдарламалық түрде шығарылды. Нүктелер саны, арнайы таңбалар

және домен құрылымы сияқты күдікті сипаттамалар талданады. Таңдауды оңтайландыру үшін оқытылған үлгілер арқылы мүмкіндіктің маңыздылығы анықталды.

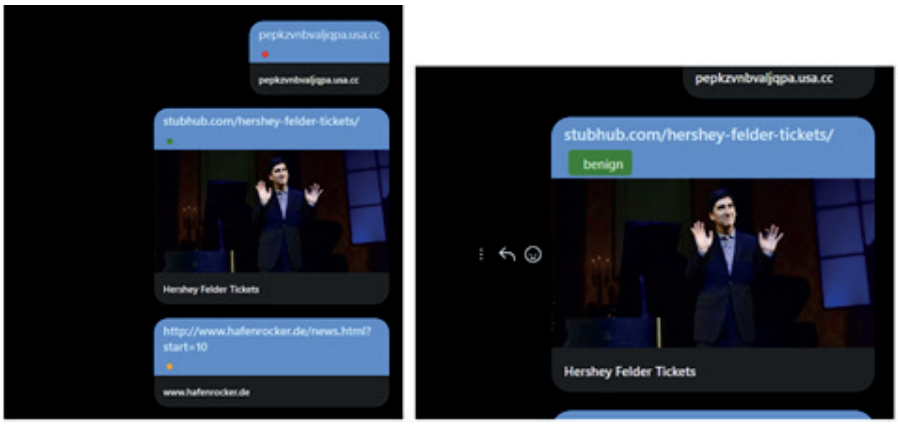
Нәтижелер мен талқылау. Интерфейс макети браузер кеңейтімі пайдаланушыларды Instagram сілтемелерінің қауіпсіздігі туралы ескертетінін көрсету үшін жасалған. 3-суретте интерфейстегі барлық пайдаланушылар және олардың әрқайсысы хабарламаға кіретін URL мекенжайлары көрсетілген. Барлық сілтемелер автоматты түрде сканерленеді және оларға сәйкес түс беріледі: қауіпті немесе жалған сілтемелер үшін қызыл және қауіпсіз сілтемелер үшін жасыл. «Instagram Protection: сканерлеу бағдарламасы» деп белгіленген Instagram қорғау кеңейтімі жоғарғы оң жақ бұрышта орналасқан. Әзірлеудің кейінгі кезеңдерінде мүмкіндіктер қолданбаны пайдалануды жеңілдету және кеңістікті толтырмау үшін жаңартылды. Сілтемелерде қауіп деңгейін көрсету үшін түстерді пайдаланбау үшін жаңа дизайн оны меңзерді апарған кезде көрсететін көпіршікті индикаторлармен ауыстырды. Әрбір сілтеменің жанында белгіше пайда болады және меңзерді оның үстіне апарған кезде, «фишинг», «зиянды бағдарлама» немесе «қатерсіз» сияқты қауіп түрін көресіз. Нәтижесінде дизайн пайдаланушыға ыңғайлы және негізгі қауіпсіздік фактілері әлі де қолжетімді.

3-сурет. Mock-up



Идея іске асырылды және MVP ретінде қарастырылды, бұл оның Instagram веб-сайтымен оңай жұмыс істеуін қамтамасыз етеді және нақты уақытта URL мониторингін ұсынады. Ағымдағы беттегі сілтемелер дереу сканерленеді және сервердегі ақпаратты пайдаланып кеңейтіммен бөлектеледі (Сурет 4).

4-сурет. Instagram Direct Messages ішіндегі бірнеше сілтемелерді тікелей таңбалау



Барлық сілтемелер олардың қандай сілтеме екенін көрсететін шағын белгілермен белгіленеді. «Фишинг» немесе «жақсы» сияқты сөздер кез келген хабар түріндегі немесе пайдаланушы биосындағы сілтемелердің жанында пайда болады, бұл қандай сілтемелер зиянды болуы мүмкін екенін тез көрсетеді. Жүйе тәуекел деңгейін көрсету үшін түсті индикаторлары бар (қызыл, жасыл, қызғылт сары сияқты) бірнеше тегтерді пайдаланады.

Тестілеу кезінде кеңейтім қысқартылған сілтемелер, қайта бағытталған домендер және тұрақты веб-сайттар сияқты әртүрлі URL мекенжайларын өңдеді. Нәтижесінде жүйе деректердің кең ауқымын жоғары дәлдікпен өңдей алады (Сурет 5).

5-сурет. Аралас жағдайдағы мысалдар

Осылайша, бұл тәсіл кәдімгі нақты уақыттағы қауіпсіздік қол жетімді болмаған кезде зиянды сілтемелерден қарапайым шолғышты қорғауды орнатуға болатынын көрсетеді.

Жүйе кейбір эвристикалық ережелермен қатар Random Forest, LightGBM және XGBoost машиналық оқыту үлгілерінің қоспасына негізделген. Сондықтан жасырын шабуылдар ертерек ашылады және проблемаларды танудағы қателер саны азаяды, бұл қажет, өйткені адамдар күнделікті әлеуметтік медианы көп пайдаланады. Модельді жасау кезінде төрт мүмкіндік

мұқият таңдалып, келесі санаттарға топтастырылды: лексикалық, желілік және хост. Нәтижесінде бұл командаға URL мекенжайларындағы өзгерістер, күдікті домендер және веб-сайттардың сілтемелерді қысқарту тәсілдері сияқты зиянды бағдарламаның әдеттен тыс шағын белгілерін табуға мүмкіндік берді. Осы механизмдердің арқасында зиянды сілтемелер тоқтатылады және адамның араласуынсыз басқаларға таралуына жол бермейді.

Зиянды URL мекенжайларын жылдам көруге ғана емес, зерттеу олардың әрекеттерін байқауға және оларды қауіпсіз сайттардан ажыратуға көмектесетін белгілерді таңдауға көмектесті. Нәтижесінде киберқауіпсіздік жүйелері күшейіп, жақсырақ қорғауға ие бола алады, сонымен қатар олар кездесетін қауіптер туралы көбірек түсінеді. Нәтижелерді ескере отырып, жүйені жетілдіретін және оны жағдайлардың кең ауқымында қолдануға жарамды ететін жаңа технологияларды қолдану арқылы одан әрі жұмыс істеуге болады. Конволюционды нейрондық желілер (CNN), қайталанатын нейрондық желілер (RNN) және трансформаторлар сияқты терең оқыту әдістері бүгінгі АИ дамуының негізгі бағыттары болып табылады. Бұл үлгілер жалған қорытындыларды азайту және қауіпті анықтауды жақсарту үшін URL мекенжайындағы таңбаларды, сондай-ақ олардың айналасындағы мәтінмәнді анықтайды.

Нәтижелерді ескере отырып, жүйені жетілдіретін және оны жағдайлардың кең ауқымында қолдануға жарамды ететін жаңа технологияларды қолдану арқылы одан әрі жұмыс істеуге болады. Конволюционды нейрондық желілер (CNN), қайталанатын нейрондық желілер (RNN) және трансформаторлар сияқты терең оқыту әдістері бүгінгі жасанжы интеллект дамуының негізгі бағыттары болып табылады. Бұл үлгілер жалған қорытындыларды азайту және қауіпті анықтауды жақсарту үшін URL мекенжайындағы таңбаларды, сондай-ақ олардың айналасындағы мәтінмәнді анықтайды. Жүйе киберәлемдегі жаңа және жылдам қозғалатын қауіптерге ілесе алуы маңызды. Модельдер нақты уақытта жаңартылуы және олардың нәтижелері өзгеріссіз қалуы үшін ағынды деректерді өңдеуді және бейімделген оқытуды қосу қажет. Пайдаланушы әрекетін талдау және контекстті зерделеу арқылы NLP зиянды сілтемелерді анықтауға көмектесе алады, сондай-ақ адаптивті аутентификация сияқты нәрселерді пайдалана отырып, жеке қорғанысты қолдайды.

Қорытынды. Зерттеулер гибридті машиналық оқыту қаупін бағалау шешімін әзірлеу арқылы Instagram сілтеме қатерін анықтауға айтарлықтай мән берді. Ұсынылған жүйе киберқауіпсіздік саласындағы елеулі прогресті көрсетеді, себебі ол жоғары дәлдік пен сенімділікті және нақты уақыттағы қауіптерді анықтау мүмкіндігін қамтамасыз етеді. Бұл жүйенің болашақ әзірлемелері терең оқыту интеграциясын, сондай-ақ бейімделгіш өңдеуді және контекстті ескеретін талдау мүмкіндіктерін, сонымен қатар қарсыластық сенімділігі мен құпиялылықты қорғау шараларын қажет етеді. Бұл жақсартулар цифрлық платформаларды әзірлеуде жеке пайдаланушыларды да, ұйымдарды да қорғайтын қауіпсіз цифрлық құрылымды жасайды.

References

Statista (2024) Most used social networks 2024, by number of users. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (in Eng.)

Alharbi N., Alkalifah B., Alqarawi G., & Rassam M.A. (2024) Countering social media cybercrime using deep learning: Instagram fake accounts detection. Future Internet, 16(10). — P.367–383. https://doi.org/10.3390/fi16100367 (in Eng.)

Sheikhi S. (2020). An efficient method for detection of fake accounts on the Instagram platform. Revue d'Intelligence Artificielle, 34(4). — P. 429–436. https://doi.org/10.18280/ria.340407 (in Eng.)

Meshram P., Bhambulkar R., Pokale P., Kharbikar K. & Awachat A. (2021, May) Automatic detection of fake profile using machine learning on Instagram. International Journal of Scientific Research in Science and Technology. — P. 117–127. https://doi.org/10.32628/ijsrst218330 (in Eng.)

Pradeep V. & Vaidehi V. (2023) Detection of malicious social bots using Instagram hashtags. International Journal of Advanced Engineering and Management, 8(2). — P.100–112. https://doi.org/10.35629/5252-06048794 (in Eng.)

Mughaid A., et al. (2023) A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. Multimedia Tools and Applications, 82(17). — P.26353–26378. https://doi.org/10.1007/s11042-023-14347-8 (in Eng.)

Aljabri M., Zagrouba R., Shaahid A., Alnasser F., Saleh A., & Alomari D. M. (2023) Machine learning-based social media bot detection: A comprehensive literature review. Social Network Analysis and Mining, 13(1). https://doi.org/10.1007/s13278-022-01020-5 (in Eng.)

Caruccio L., Cimino G., Cirillo S., Desiato D., Polese G. & Tortora G. (2023) Malicious account identification in social network platforms. ACM Transactions on Internet Technology, 23(4). — P.1–25. https://doi.org/10.1145/3625097 (in Eng.)

Raja M.S., & Raj L.A. (2021) Detection of malicious profiles and protecting users in online social networks. Wireless Personal Communications, 127(1). — P.107–124. https://doi.org/10.1007/s11277-021-08095-x (in Eng.)

Kaushik K., Bhardwaj A., Kumar M., Gupta S.K., & Gupta A. (2022) A novel machine learning-based framework for detecting fake Instagram profiles. Concurrency and Computation: Practice and Experience, 34(28). https://doi.org/10.1002/cpe.7349 (in Eng.)

Durga P., & Sudhakar D.T. (2023, January) The use of supervised machine learning classifiers for the detection of fake Instagram accounts. Journal of Pharmaceutical Negative Results. — P.267–279. https://doi.org/10.47750/pnr.2023.14.03.36 (in Eng.)

Salamh F.E., Mirza M.M., Hutchinson S., Yoon Y.H., & Karabiyik U. (2021) What's on the horizon? An in-depth forensic analysis of Android and iOS applications. IEEE Access, 9. — P.99421–99454. https://doi.org/10.1109/ACCESS.2021.3095562 (in Eng.)

Nobili M. (2023) Review OSINT tool for social engineering. Frontiers in Big Data, 6. https://doi.org/10.3389/fdata.2023.1169636 (in Eng.)

Alsharida R.A., Al-Rimy B.A., Al-Emran M., & Zainal A. (2023) A systematic review of multi perspectives on human cybersecurity behavior. Technology in Society, 73. — P.102–118. https://doi.org/10.1016/j.techsoc.2023 (in Eng.)

Prince N.U., et al. (2024, August). AI-powered data-driven cybersecurity techniques: Boosting threat identification and reaction. Nano-NTP, 20(S10). — P.1804–1815. https://doi.org/10.62441/nano-ntp.v20iS10.1804 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 52–65

https://doi.org/10.32014/2025.2518-1726.363

МРНТИ 28.23.01
УДК 004.8

G. Argyngazin, 2025.
National Defense University of the Republic of Kazakhstan,
Astana, Kazakhstan.
E-mail: argyngazin@mail.ru

ARTIFICIAL INTELLIGENCE: IS ALARMISM JUSTIFIED?

Argyngazin Galym — Doctoral student at the National Defense University of the Republic of Kazakhstan, Astana, Kazakhstan,
E-mail: argyngazin@mail.ru, ORCID ID: https://orcid.org/0009-0003-5651-8984.

Abstract. This article explores the scientific validity of alarmist sentiments and predictions regarding the future development of artificial intelligence. It analyzes the evolution of global risks and threats, proposing that the growth of technological progress correlates with a rise in existential risks. The historical context, in which the term «artificial intelligence» emerged, is outlined, along with various definitions of the concept. The paper highlights the key differences between AI-related alarmism and other global threats such as nuclear, environmental, and biotechnological risks. The main arguments of alarmist thinkers are examined, including scientific and expert forecasts that often verge on apocalyptic scenarios. Some of these concerns have already materialized, causing harm to government institutions, the corporate sector, and individuals. The paper also presents counterarguments from opponents of alarmism, who believe that with proper regulation and ethical alignment, artificial intelligence does not pose an existential threat. Based on the analysis of both perspectives, the article concludes that each side offers rational arguments that merit serious consideration by the global scientific community and governments. In this context, the article cites examples of growing efforts to establish ethical standards and regulatory frameworks for AI. While alarmism may be justified in certain cases and should inform public policy, excessive emphasis on worst-case scenarios appears less scientifically grounded at the current stage of AI development and leans toward speculative or science fiction narratives.

Keywords: Artificial intelligence, technological development, global threats, existential risks, alarmism

Ғ.А. Арғынғазин, 2025.
Қазақстан Республикасының Ұлттық қорғаныс университеті,
Астана, Қазақстан.
E-mail: argyngazin@mail.ru

ЖАСАНДЫ ИНТЕЛЛЕКТ: АЛАРМИСТІК КӨЗҚАРАС ҚАЛЫПТАСТЫРУ ОРЫНДЫ МА?

Арғынғазин Ғалым Арғынғазыұлы — Қазақстан Республикасының Ұлттық қорғаныс университетінің докторанты, Астана, Қазақстан,
E-mail: argyngazin@mail.ru, ORCID ID: https://orcid.org/0009-0003-5651-8984.

Аннотация. Мақалада жасанды интеллекттің одан әрі дамуына қатысты алармистік сезімдер мен болжамдардың ғылыми орындылығын анықтаудың әрекеті жасалады. Жаһандық тәуекелдер мен қауіптердің эволюциясы талданады, соның негізінде технологиялық дамудың өсуі экзистенциалды тәуекелдердің өсуіне тікелей пропорционалды деген болжам жасалады. «Жасанды интеллект» терминінің пайда болуының тарихи контексті сипатталады және оның анықтамалары қарастырылады. Жасанды интеллектке қатысты алармизмнің ядролық, экологиялық және биотехнологиялық қауіптер сияқты жаһандық тәуекелдердің басқа мысалдарынан негізгі айырмашылықтары көрсетіледі. Алармистік ғылыми және сараптамалық болжамдарды алға тартатын бағыт өкілдерінің негізгі дәлелдері талданады. Мемлекеттік органдарға, корпоративтік секторға және қарапайым адамдарға зиян келтірген кейбір мысалдар келтіріледі. Сондай-ақ, «ақылды» машиналардың жұмыс принциптерін адами құндылықтармен сәйкестендіру және үйлестіру арқылы жасанды интеллекттің қаупін төмендету мүмкін деген ойды алға тартатын антиаларистік көзқарас өкілдерінің ұстанымдары қарастырылады. Екі тәсілді талдау негізінде екеуінде де ғылыми қоғамдастық пен әлем үкіметтерінің назарын талап ететін ұтымды дәлелдер бар деген қорытынды жасалады. Бұл тұрғыда «этикалық мінез-құлық» мәдениетін қалыптастыруға және жасанды интеллектті реттеуге бағытталған шаралардың жанданғанын растайтын нақты деректер келтірілген. Мақаланы қорытындылай келе, жекелеген ғалымдар мен сарапшылардың алармистік көзқарастарды қалыптастыруы орынды болуы мүмкіндігі және ұлттық мемлекеттер мен трансұлттық құрылымдардың назарында болуы туралы ой айтылады. Алайда жасанды интеллект дамуының қазіргі кезеңінде қауіптер мен тәуекелдерді асыра көрсету, оларды шектен тыс бағалау ғылыми тұрғыдан негізсіз және көбіне қатаң футуристік немесе фантастикалық идеялармен ұштасатыны ескертіледі.

Түйін сөздер: жасанды интеллект, технологиялық даму, жаһандық қауіптер, экзистенциалды тәуекелдер, алармизм

Ғ.А. Арғынғазин, 2025.
Национальный университет обороны Республики Казахстан,
Астана, Казахстан.
E-mail: argyngazin@mail.ru

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ОПРАВДАН ЛИ АЛАРМИЗМ?

Арғынғазин Ғалым Арғынғазыұлы — докторант Национального университета обороны Республики Казахстан, Астана, Казахстан,
E-mail: argyngazin@mail.ru, ORCID ID: https://orcid.org/0009-0003-5651-8984.

Аннотация. В статье осуществлена попытка определения научной оправданности алармистских настроений и прогнозов в отношении дальнейшего развития искусственного интеллекта. Проанализирована эволюция глобальных рисков и угроз, на основании чего, сделано предположение о том, что рост технологического развития прямо пропорционален росту экзистенциональных рисков. Описан исторический контекст появления термина «искусственный интеллект» и рассмотрены его определения. Изложены основные отличия алармизма в отношении искусственного интеллекта от иных примеров глобальных рисков, таких как ядерная, экологическая и биотехнологическая угрозы. Проанализированы основные доводы представителей алармистского подхода, которыми высказываются научные и экспертные предположения, граничащие с апокалиптическими прогнозами. Изложены примеры отдельных подтвердившихся опасений, которые нанесли вред государственным учреждениям, корпоративному сектору и простым людям. Также рассмотрены позиции представителей антиалармистского подхода, доводы которых сводятся к тому, что при правильном регулировании и согласовании принципов работы «умных» машин с человеческим ценностями, искусственный интеллект не является экзистенциальной угрозой. На основании анализа двух подходов, делается вывод о том, что обе точки зрения имеют под собой ряд достаточно рациональных аргументов, требующих внимания научного сообщества и правительств мира. В этом контексте приводятся данные, подтверждающие активизацию мер, направленных на формирование культуры «этического поведения» и регулирование искусственного интеллекта. В заключении делается вывод о том, что алармизм отдельных ученых и экспертов может быть оправдан, а также должен быть взят на вооружение национальными государствами и транснациональными структурами, однако абсолютизация рисков и угроз, создание апокалиптических настроений, на данном этапе развития искусственного интеллекта, выглядят менее научно обоснованными и граничат с строго футуристскими или фантастическими представлениями.

Ключевые слова: Искусственный интеллект, технологическое развитие, глобальные угрозы, экзистенциональные риски, алармизм

Введение. Резкая интенсификация трансформации материального мира в XX веке, связанная с превращением науки в основной фактор технологического прогресса и инноваций, способствовала появлению различных алармистских настроений и прогнозов о рисках глобального масштаба. В 1970 году один из основателей концепции постиндустриального общества Элвин Тоффлер, описывая тенденцию ускорения темпов технологического и социального прогресса, отмечал, «что если последние 50 000 лет существования человека разделить на отрезки жизни приблизительно в 62 года каждый, то окажется около 800 таких отрезков жизни… И подавляющее большинство всех материальных благ, которыми мы пользуемся в повседневной жизни в настоящее время, были придуманы в течение настоящего, 800-го отрезка жизни» (Тоффлер, 2002).

Бурный рост технологичности и масштабы его проникновения в повседневную жизнь не просто привели к углублению индустриализации и автоматизации, но и породили ситуацию, когда результаты деятельности одного или нескольких человек могут повлиять на все человечество, а совершенные ошибки – привести к катастрофическим последствиям. Как отмечал Норберт Винер, раньше абсурдные человеческие действия были относительно безвредными, потому что их ограничивал невысокий уровень технологического развития (Винер, 1966). Однако этот период защищенности быстро подошел к концу, что способствует появлению различных экзистенциональных рисков (Рассел, 2021). Подобное положение послужило активизации научно-теоретических изысканий в данном направлений, однако научное сообщество все еще не поставило точку в вопросе об оправданности распространяемого алармизма, связанного с развитием искусственного интеллекта. Данная статья является попыткой поиска ответа на данный вопрос.

Материалы и методы. На основании изучения научно-теоретических работ проведены диахроническое исследование эволюции глобальных экзистенциональных рисков и сравнительный анализ позиций представителей алармисткого и антиалармистского подходов в отношении дальнейшего развития искусственного интеллекта.

Результаты и обсуждение. В последние десятилетия актуализировались алармистские настроения, связанные с искусственным интеллектом. При этом чем более технологичным и информационным становится общество, тем более растут масштабы обозначаемых угроз.

Появление и интенсификация тревожности, связанной с развитием технологий, имеет свои исторические предпосылки. При этом наиболее явные риски и угрозы глобального масштаба связаны с результатами человеческой деятельности последних 100 лет.

40-е годы прошлого столетия появилось ядерное оружие, что породило ряд серьезных рисков мирового масштаба как во время его разработки, так и после его распространения. За шесть месяцев до первого ядерного испытания ученые, задействованные в Манхэттенском проекте, подготовили отчет

под названием LA-602. В этом документе были представлены результаты исследований, касающихся возможных последствий ядерного взрыва, включая вероятность уничтожения Земли из-за возможного возгорания атмосферы. Этот отчёт можно считать, одним из первых научных исследований, посвященных глобальным угрозам, ставящим под угрозу само существование человечества (SIPRI, 2024). Первый ядерный удар в истории был совершен 6 августа 1945 года Соединёнными Штатами Америки по японскому городу Хиросима. Атомная бомба, названная «Малыш» (Little Boy), была сброшена с бомбардировщика B-29 «Энола Гей». По разным оценкам, около 140 тысяч человек погибло во время взрыва и умерло в течение последующих месяцев.

В начале 2024 года девять стран – Соединенные Штаты, Россия, Великобритания, Франция, Китай, Индия, Пакистан, Корейская Народно-Демократическая Республика (Северная Корея) и Израиль – в совокупности владели 12 121 ядерным оружием, из которых 9585 считались потенциально боеготовыми (SIPRI, 2024). Накопленный арсенал способен многократно уничтожить всю планету и делает ядерную угрозу одной из наиболее серьезных факторов, влияющих на глобальные политические и военные процессы. Геополитическая напряженность и непредсказуемость XXI века, которую метафорично стали называть новой константой, только усугубляют ядерную угрозу.

Следующими в хронологическом порядке глобальными рисками можно назвать получившие наибольшую актуальность в 1960-х и 1970-х годах экологические проблемы, связанные с последствиями индустриализации и негативным эффектом от антропогенного фактора. В 1980-х годах ученые начали всерьез предупреждать о рисках глобального изменения климата из-за выбросов парниковых газов.

К XXI веку развитие генной инженерии и биотехнологий привели к опасениям о создании патогенов или организмов, которые могут представлять угрозу для человечества. Вторжение человека в ареалы животного мира активизировали риск распространения зоонозных вирусов. COVID-19 подтвердил подобные опасения и способствовал появлению новых прогнозов о появлении в будущем «болезни X», ориентируя правительства передовых стран на совершенствование своих политик в области биологической безопасности.

Появление в течение XX и начале XXI века подобных крупных рисков о возможности исчезновения человечества, чего раньше не было за всю историю человечества, имеет закономерный характер. Проследив эволюцию глобальных рисков и угроз, можно вывести следующую закономерность – рост технологического развития прямо пропорционален росту экзистенциональных рисков. Прежде чем перейти к описанию алармистских настроений по поводу появления и развития элементов искусственного интеллекта, предлагается коротко остановиться на историческом контексте появления

термина и данных ему определениях. Впервые термин «искусственный интеллект» был предложен американским информатиком Джоном Маккарти на инициированном им научном семинаре в Дартмутском колледже в 1956 году (Арғынғазин, 2024).

Принятие термина научным сообществом имело больше компромиссный, а не консенсусный характер, поскольку отдельные ученые имели и имеют свое отличное мнение о том, как лучше было бы назвать эту научную область. В качестве альтернативных вариантов предлагались такие словосочетания как кибернетика, исследование автоматов, комплексная обработка информации и машинный интеллект. Основная часть несогласных с выбранной формулировкой термина аргументировали свою позицию возникающим морально-этическим диссонансом, выражающим при применении понятия «интеллект», традиционно свойственного только разумным существам, в отношении машины. Однако предложенный термин, отличавшийся от других своей оригинальностью и привлекательностью, а также подчеркивавший амбициозность стремлений ученых к созданию машин, имитирующих отдельные аспекты человеческого мышления, все же прижился в научном сообществе. Среди ученых имеется множество вариаций определений, данных термину. К примеру, Дж. Маккарти определял его как науку и технику создания интеллектуальных машин», Марвин Мински – как «науку о том, как заставить делать машины то, что потребовало бы интеллекта, если бы это делали люди», Алекс Эндрю предполагал, что искусственный интеллект представляет собой «способы создания вычислительных машин, обладающих интеллектуальным поведением» (Арғынғазин, 2024).

Мировому научному сообществу еще предстоит выработать универсальное и обоснованное определение термина «искусственный интеллект», которое способствовало бы снижению разночтений, недоразумений и двусмысленностей среди исследователей, направляя их к согласованным и скоординированным усилиям в области дальнейшего развития данной области науки и технологии. Тогда как нам в рамках данной статьи предлагается придерживаться достаточно широко распространенного определения, согласно которому искусственный интеллект – свойство машин получать результаты, схожие с отдельными элементами человеческого интеллекта.

Теперь, когда мы определились с категориальным аппаратом, предлагается более глубоко погрузиться в вопросы тревожности и опасений, связанных с искусственным интеллектом. Прежде всего, следует отметить, что алармизм в отношении искусственного интеллекта отличается от предыдущих примеров глобальных рисков. Поскольку в случае с данной технологией происходит не просто очередная смена научно-технологической парадигмы мира, когда новации создают дополнительные возможности, приводящие к новым рискам. Сегодня машина стала приобретать отдельные элементы когнитивных характеристик человека, что порождает среди научной общественности и

экспертов алармистские настроения совершенно иного характера, а также потребность в регулировании этических аспектов использования «умных» машин. Как отмечал по этому поводу израильский математик Моше Варди: «В XIX веке машины конкурировали с мышцами человека. Теперь машины соревнуются с человеческим мозгом. Роботы сочетают в себе интеллект и физическую силу. Мы стоим перед перспективой быть полностью превзойденными своим собственным созданием» (Vardi, 2013).

Важно отметить, что с момента своего появления большая часть продуктов искусственного интеллекта носила строго утилитаристский характер и рассматривалась как очередной инструмент, облегчающий жизнь человека, не предполагающий каких-либо рисков, тем более экзистенционального характера.

Для такого прагматического подхода были соответствующие теоретические основания. В 1944 году известный венгерский математик Джон фон Нейман и американский экономист Оскар Моргенштерн в фундаментальном труде «Теория игр и экономическое поведение» обосновали математические основы теории полезности, которые сводятся к тому, что рациональный агент ориентирован на максимизацию ожидаемой полезности (Нейман, Моргенштерн, 1970). Рациональный агент – это гипотетическая модель, действующая в соответствии с принципами рациональности. Ученые и эксперты с целью максимизации пользы для человека и общества были ориентированы создавать подобные системы, что способствовало расширению масштабов распространения искусственного интеллекта в социальные, экономические, политические и военные сферы. Однако подобные утилитаристские представления о создании «разумных» машин, способных стать человеческими помощниками, по мере экспоненциализации роста вычислительных технологий, все больше стали рождать алармистские настроения в научной и экспертной среде. Если в начале подобные представления больше находили отражение в научно-фантастических произведениях и футуристических представлениях писателей, то уже к 1960 годам стали появляться научные труды, обосновывающие искусственный интеллект как новую угрозу глобальной безопасности.

Способствовало этой тенденции появление все более сложных агентов, ориентированных на самообучение и самосовершенствование через такие технологии как машинное и глубокое обучение, что усилило страх и тревогу общественности, связанные с возможностью потери контроля над этим процессом. В связи с этим алармизм стал новой повесткой для обсуждения рядом ученых и экспертов в области высоких технологий и искусственного интеллекта, которые продолжаются до сегодняшнего дня. В этом контексте отдельные ученые ставят под сомнение саму цель развития искусственного интеллекта, отмечая, что человечество может добиться слишком высоких результатов в этой области, породив большое количество угроз (Рассел, 2021).

В начале XXI века алармистских позиций по отношению к искусственному интеллекту придерживались Стюарт Рассел, Питер Норвиг, Илон Маск, Стивен Хокинг, Билл Гейтс, Ник Бостром, Роберт Джераси, Моше Варди и многие другие. Представители такого подхода, учитывая, что в последние 20-30 лет наблюдается бурный рост искусственного интеллекта и отсутствие реальных механизмов его контроля, высказывали ряд предположений, граничащих с апокалиптическими прогнозами.

В открытом письме, подписанным многими из вышеперечисленных лиц, опубликованным 28 октября 2015 года, отмечаются предостережения, общий смысл которых заключатся в следующем: «Потенциал искусственного интеллекта колоссален, и поэтому важно понять, как воспользоваться его преимуществами и не угодить в опасную ловушку» (Future of Life Institute, 2025). Шведский философ и профессор Оксфордского университета Ник Бостром в своей книге «Искусственный интеллект. Этапы. Угрозы. Стратегии.» указывает, что технологический процесс идет в сторону разработки машинного интеллекта человеческого уровня, который будет значительно превосходить когнитивные способности человека. Таким сверхразумом станет трудно управлять, поскольку он будет иметь возможность самосохранения и сопротивления любым попыткам приостановления его деятельности по реализации сгенерированной им задачи, которая может быть катастрофичной для человечества (Бостром, 2016). Доводы представителей подобного подхода тесно связаны с понятием технологической «сингулярности» – гипотетическим моментом в будущем, когда искусственный интеллект станет настолько продвинутым и самосовершенствующимся, что превзойдет человеческий интеллект и способность контролировать его развитие. Впервые этот термин, в контексте роста технологичности, был использован Джоном фон Нейманом (Ulam, 1958).

Американский изобретатель Рэй Курцвейл, известный своими футурологическими работами в области искусственного интеллекта, спрогнозировал, что технологическая сингулярность, наступит в 2045 году, когда технологии станут настолько мощными, что они смогут создать улучшенные версии самих себя без помощи человека (Kurzweil, 2024).

Опасения отдельных ученых носят менее эмоционально окрашенный характер и связаны с несовершенством алгоритмов или непреднамеренными действиями программистов. Часть подобных опасений уже подтверждается, создавая определенные угрозы для государственных учреждений, корпоративного сектора и простых людей. К примеру, в августе 2012 года американская финансовая компания Knight Capital запустила новую автоматизированную программу для торговли акциями, содержащую ошибку в коде. За 45 минут она совершила более 4 миллионов иррациональных сделок на сумму 7 миллиардов долларов, что едва не привело к банкротству компании (SIPRI, 2024).

В феврале 2024 года в Гонконге произошел инцидент, связанный с мошенничеством, в ходе которого злоумышленники использовали технологию искусственного интеллекта deepfake для хищения 26 миллионов долларов у транснациональной корпорации. Преступники, применяя синтезированные аудио- и видеоматериалы, имитировали голоса и образы руководителей компании во время видеоконференции. В ходе взаимодействий они убеждали сотрудников переводить финансовые средства на определенные банковские счета, выдавая себя за старших руководителей организации (CNN, 2025).

Несмотря на масштабы движения алармистов и указанные примеры использования искусственного интеллекта во вред человеку, имеется немалое количество ученых и экспертов, а также большое количество обывателей, не воспринимающих данные опасения всерьез. Их можно назвать представителями антиалармистской позиции или, как отмечал российский философ и специалист по информационным технологиям Анатолий Ракитов, компьютерного агностицизма и пессимизма (Ракитов, 1991). Они, поддерживая теорию о рациональном агенте, считают, что искусственный интеллект следует рассматривать как инструмент, способствующий улучшению качества жизни человека, отмечая важные утилитаристские и альтруистские характеристики «умных машин», а также их неспособность навредить человечеству. К примеру, известный психолог Рой Баумайстер отмечал: «Так называемые думающие машины – это расширение человеческого разума. Они не существуют в природе. Они не созданы эволюцией, они созданы людьми из чертежей и теорий. Человеческий разум учится создавать инструменты, позволяющие ему лучше работать. Компьютер – один из лучших инструментов» (Брокман, 2017). Психолог Арнольд Трехуб высказывает аналогичную мысль, утверждая, что машины являются лишь «сконструированными человеком артефактами», которые не имеют «своеобразного взгляда на мирские референции» и не могут думать (Брокман, 2017).

Общая тональность их доводов сводится к тому, что при правильном регулировании и согласовании принципов работы «умных» машин с человеческим ценностями, искусственный интеллект не является экзистенциальной угрозой, скорее представляя технологичную возможность для улучшения жизни человека.

Отдельные из них обосновывают мысль, что алармистские представления и так называемая «робофобия» связаны с искаженными представлениями человека о машине и приписывании ей мотивов и характеристик людей. Французский ученый Ян Лекун в этом контексте отмечает: «Наш страх перед роботом, желающим захватить власть, – это проецирование особенностей человеческой природы на машины» (Лекун, 2021). В данном случае под особенностями человеческой природы понимаются идеи об естественном стремлении человека к власти, как центральной движущей силе человеческого существенная, которые восходят к работам Гоббса и Ницше.

Обе точки зрения на будущее человечества, детерминированное развитием искусственного интеллекта, имеют под собой ряд достаточно рациональных аргументов, требующих внимания научного сообщества и правительств мира. В качестве ответа на алармизм и апокалиптические настроения первой группы ученых, в последние годы все активнее стали приниматься меры, направленные на формирование культуры «этического поведения» и регулирование искусственного интеллекта.

Согласно последнему отчету исследовательского института по искусственному интеллекту Стэндфордского университета, в мире наблюдается устойчивая динамика увеличения количества правовых актов, регулирующих искусственный интеллект. Как указывается в отчете, если в США в 2016 году действовал лишь один нормативный акт, регулирующий сферу, то в 2023 году их количество увеличилось до 25 документов. Только в 2023 году общее количество нормативных актов, связанных с искусственным интеллектом, выросло на 56,3% (The Institute for Human-Centered Artificial Intelligence, 2024).

21 мая 2024 года Совет Евросоюза утвердил первый в мире Закон об искусственном интеллекте. Данный закон устанавливает ряд запретов, классифицирует искусственный интеллект в соответствии с уровнем риска от него, возлагая ответственность за высокорисковые системы на их разработчиков.

К примеру, статья 5 Закона предполагает запрет на использование манипуляторных, обманных и иных методов с использованием искусственного интеллекта, способных причинить вред человеку; создание и расширение базы данных распознавания лиц посредством нецелевого извлечения изображений лиц людей из Интернета; определение эмоций физического лица на рабочих местах и в образовательных учреждениях; классификации физических лиц на основе их биометрических данных для определения их расы, политических взглядов, членства в профсоюзах, религиозных или философских убеждений, а также многие другие регуляторные положения (Regulation (EU) 2024/1689 of the European Parliament and of the Council, 2024).

Несмотря на то, что закон ориентирован на минимизацию рисков и угроз современного характера, его идея во многом восходит к трем законам робототехники Айзека Азимова, сформулированным в 1940-е годы, которые предполагают, что робот не может причинить вред человеку (первый закон), должен повиноваться всем его приказам (второй закон) и заботиться о своей безопасности, если это не противоречит первому или второму законам (третий закон) (Пиковер, 2023).

Дальнейшее регулирование искусственного интеллекта, а также следование ученых и разработчиков определенным рекомендациям международного и национального масштабов, вероятнее всего, снизят отдельные риски и угрозы от неконтролируемого развития «умных» машин. Однако полное нивелирование

рисков, абсолютный контроль за технологией и ее разработчиками, а также обеспечение лишь высокогуманистического характера развития искусственного интеллекта, с большей вероятностью, невозможно. Поэтому опасения, связанные с искусственном интеллектом и попытки научного, экспертного сообщества, правительств стран и транснациональных компаний по обеспечению безопасности от искусственного интеллекта становятся новой реальностью, с которой мы будем жить также как с постоянной ядерной угрозой и экологическими, биотехнологическими рисками антропогенного характера.

При этом современная наука еще не подобралась к полному пониманию функционирования человеческого мозга, способности которого планирует эмулировать. Поэтому искусственной машине далеко до человеческого интеллекта, рациональности, а тем более такого понятия как сознание, хотя она уже превосходит его в отдельных когнитивных способностях.

К примеру, Стюарт Рассел, рассуждая на эту тему, отмечает, что машины еще нельзя называть полностью разумными и обладающими достаточными когнитивными способностями: «Обращать внимание исключительно на вычислительную мощность – значит очень сильно заблуждаться. Одна лишь скорость не подарит нам искусственный интеллект» (Рассел, 2021). Или как отмечал математик Кит Девлин: «Если нечто ходит как утка и крякает как утка, это еще не делает его уткой. И если машина демонстрирует некоторые черты мышления (например, способность принимать решения), это еще не делает ее мыслящей» (SIPRI, 2024).

Более того, отдельные ученые считают некорректным сам термин «мыслящая машина». К примеру, нейробиолог Лео Чалуна в этом контексте пишет следующее: «… термин «мыслящая машина» употребляется неправильно. Ни одна машина не задается вечными вопросами: «Откуда я взялась? Зачем я здесь? Куда я иду?» Машины не думают о своем будущем, о своем неизбежном конце или о своем наследии. Чтобы размышлять над такими вопросами, требуется сознание и самосознание. У мыслящих машин их нет, и, учитывая наш нынешний уровень знаний, они вряд ли это получат в обозримом будущем» (SIPRI, 2024).

Американский философ Джон Сирл еще более углубляется в эту тему и отмечает, что машины не могут обрести сознание, поскольку для этого требуется воспроизведение сложных биологических процессов, характерных для человеческого мозга. Он развивает концепцию биологического натурализма, согласно которой создание сознательного существа возможно лишь при воспроизведении физических и химических процессов, происходящих в человеческом мозге, что невозможно в условиях компьютерного интеллекта (Guryanova, Shestakov, Noskov, 2019).

Сегодня ученые не могут создать одну универсальную систему, аналогичную человеку, которая могла бы делать все, что умеют люди. Они

ограничены созданием разных утилитаристских программ с разными задачами, которые зачастую превосходят отдельные человеческие возможности. К примеру, если человек может прочесть одну книгу за 5-10 дней, машину, при соответствующем уровне вычислительной мощности, теоретически можно настроить параллельно прочитать и понять за несколько часов все 150 миллионов когда-либо написанных книг (Рассел, 2021). Однако большая часть экспертов сходится во мнении, что превзойдя человека в отдельных аспектах интеллектуальности, машина не в состоянии сделать это в остальных, к примеру, таких, как эмоциональный интеллект, мышление, эмпатия, креативность, интуиция и сознание.

Мы разделяем такую точку зрения и полагаем, что проводить прямую связь между искусственным и человеческим интеллектом, а тем более прогнозировать превосходство первого над последним на текущем историческом этапе рано. Более того, нам видится, что полная эмуляция машиной человеческого интеллекта невозможна и в долгосрочной перспективе.

Заключение. Учитывая изложенное, мы полагаем, что на данном этапе развития искусственного интеллекта, создание алармистких и апокалиптических настроений в отношении дальнейшего развития «умных машин» выглядит менее научно обоснованным и граничит с футуристскими или фантастическими представлениями. Подобное допущение основано на том, что искусственной интеллект еще не достиг уровня человеческого разума. Более того, мы предполагаем, что реализация подобной перспективы, в условиях текущего научно-технологического развития, имеет крайне невысокую вероятность.

Также представляется недостаточно обоснованным применение принципов антропоморфизма в отношении действий и мотивов машин, то есть восприятие искусственного интеллекта как существа с собственными целями, эмоциями и намерениями. При этом обозначаемые отдельными учеными и экспертами риски и угрозы от неконтролируемого развития искусственного интеллекта с теоретической и практической точки зрения вполне оправданы, а также должны быть взяты на вооружение национальными государствами и транснациональными структурами, в том числе представителями государственных структур Республики Казахстан.

Литература

Тоффлер Э. (2002) Шок будущего: Пер. с англ. – М.: ООО «Издательство АСТ». — 557 с.

Винер Н. (1966) Творец и робот: Обсуждение некоторых проблем, в которых кибернетика сталкивается с религией. — М.: Прогресс. — 104 с.

Рассел С. (2021) Совместимость. Как контролировать искусственный интеллект: Пер. с англ. — М.: Альпина нон-фикшн. — 438 с.

SIPRI Yearbook 2024: Armaments, Disarmament and International Security. (2024) — Oxford University Press. — 682 p.

Арғынғазин Ғ.А. (2024) К вопросу о термине «искусственный интеллект». Вестник Национального университета обороны Республики Казахстан. — № 2. — С. 174–177.

Vardi M.Y. (2013) If machines are capable of doing almost any work humans can do, what will humans do? — 8 p.

Нейман Дж. фон, Моргенштерн О. (1970) Теория игр и экономическое поведение: Пер. с англ. —— М.: Издательство «Наука». — 707 с.

Open letter on artificial intelligence. Future of Life Institute. Электронный ресурс. Режим доступа: *https://futureoflife.org/open-letter/ai-open-letter/* (Дата обращения: 20.03.2025).

Бостром Н. (2016) Искусственный интеллект. Этапы. Угрозы. Стратегии: Пер. с англ. С. Филина. — М.: Манн, Иванов и Фербер. — 528 с.

Ulam S. (1958) Tribute to John von Neumann // Bulletin of the American Mathematical Society. — 49 p.

Kurzweil R. (2024) The Singularity Is Nearer: When We Merge with AI. — Random House. — 320 p.

Deepfake CFO scam rocks Hong Kong. CNN — Электронный ресурс. – Режим доступа: *https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html*. (Дата обращения: 21.03.2025).

Ракитов А.И. (1991) Философия компьютерной революции. — М.: Политиздат. — 287 с.

Брокман Дж. (2017) Что мы думаем о машинах, которые думают: Ведущие мировые ученые об искусственном интеллекте: Пер. с англ. — М.: Альпина нон-фикшн. — 549 с.

Лекун Я. (2021) Как учится машина: Революция в области нейронных сетей и глубокого обучения: Пер. с франц. — М.: Альпина ПРО. — 335 с.

Пиковер К. (2023) Искусственный интеллект: Пер. с англ. А. Ефимовой. — М.: Синдбад. — 224 с.

Artificial Intelligence Index Report 2024. (2024) – The Institute for Human-Centered Artificial Intelligence, Stanford University. – 502 p.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).

Guryanova A., Shestakov A., Noskov E. (2019) Digital Ethics As An Instrument For The Technological Challenges Regulation. Global Challenges and Prospects of the Modern Economic Development. – European Proceedings of Social and Behavioural Sciences. – Vol. 57. — P. 251–262.

References

Toffler È. (2002) Šok budušcego [Future Shock]: Per. s angl. — M.: OOO «Izdatel'stvo AST». — P.557. (in Russ.)

Viner N. (1966) Tvorec i robot: Obsuždenie nekotoryh problem, v kotoryh kibernetika stalkivaetsâ s religiêj [The Human Use of Human Beings: Cybernetics and Society]. — M.: Progress. — P. 104 (in Russ.)

Rassel S. (2021) Sovmestimost'. Kak kontrolorovat' iskusstvennyj intellekt [Human Compatible: Artificial Intelligence and the Problem of Control]: Per. s angl. — M.: Alpina non-fikšn. — P. 438. (in Russ.)

SIPRI Yearbook 2024: Armaments, Disarmament and International Security. (2024) — Oxford University Press. — 682 p. (in Eng.)

Argıngazin Ğ.A. (2024) K voprosu o terme «iskusstvennyj intellekt» [On the Term "Artificial Intelligence"]. Vestnik Nacional'nogo universiteta oborony Respubliki Kazakhstan. — № 2. —P. 174–177. (in Russ.)

Vardi M.Y. (2013) If machines are capable of doing almost any work humans can do, what will humans do? — P. 8 (in Eng.)

Nejman Dž. fon, Morgenshtern O. (1970) Teoriâ igr i èkonomičeskoe povedenie [Theory of Games and Economic Behavior]: Per. s angl. — M.: Izdatel'stvo «Nauka». — P. 707. (in Russ.)

Open letter on artificial intelligence. Future of Life Institute. Электронный ресурс. Режим доступа: https://futureoflife.org/open-letter/ai-open-letter/ (Дата обращения: 20.03.2025). (in Eng.)

Bostrom N. (2016) Iskusstvennyj intellekt. Ètapy. Ugrozy. Strategii [Superintelligence: Paths, Dangers, Strategies]: Per. s angl. S. Filina. — M.: Mann, Ivanov i Ferber. — P. 528. (in Russ.)

Ulam S. (1958) Tribute to John von Neumann. Bulletin of the American Mathematical Society. — P. 49. (in Eng.)

Kurzweil R. (2024) The Singularity Is Nearer: When We Merge with AI. — Random House. — P. 320. (in Eng.)

Deepfake CFO scam rocks Hong Kong. CNN – Электронный ресурс. — Режим доступа: *https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html*. (Дата обращения: 21.03.2025). (in Eng.)

Rakitov A.I. (1991) Filosofiâ komp'ûternoj revolûcii [Philosophy of the Computer Revolution]. — M.: Politizdat. — P. 287. (in Russ.)

Brokman Dž. (2017) Čto my dumaem o mašinah, kotorye dumaût: Veduşçie mirovyê učenyê ob iskusstvennom intellekte [What We Think About Machines That Think: Leading Scientists Talk About AI]: Per. s angl. — M.: Alpina non-fikšn. — P. 549. (in Russ.)

Lekun Â. (2021) Kak učitsâ mašina: Revolûciâ v oblasti nejronnyh setej i glubokogo obučeniâ [Deep Learning Revolution: How Machines Learn]: Per. s franc. — M.: Alpina PRO. — P. 335. (in Russ.)

Pikover K. (2023) Iskusstvennyj intellekt [Artificial Intelligence]: Per. s angl. A. Efimovoj. — M.: Sindbad. — P. 224. (in Russ.)

Artificial Intelligence Index Report 2024. (2024) — The Institute for Human-Centered Artificial Intelligence, Stanford University. — P. 502. (in Eng.)

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). (in Eng.)

Guryanova A., Shestakov A., Noskov E. (2019) Digital Ethics As An Instrument For The Technological Challenges Regulation. Global Challenges and Prospects of the Modern Economic Development. — European Proceedings of Social and Behavioural Sciences. —— Vol. 57. — P. 251–262. (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 66–77

https://doi.org/10.32014/2025.2518-1726.364

UDC 28.23.01

Zh.A. Abdibayev[1], S.K. Sagnayeva[1], B.B. Orazbayev[1], M. James C. Crabbe[2], K.A. Dyussekeyev*[1], 2025.
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Wolfson College, University of Oxford, Oxford, United Kingdom.
E-mail: dyussekeyev_ka@enu.kz

DEVELOPMENT OF AN EFFECTIVE WATER ACCOUNTING METHOD FOR IRRIGATION SYSTEMS FOR AUTOMATED WATER RESOURCE MANAGEMENT SYSTEMS

Abdibayev Zhanuzak — doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: abdibaev.zha@gmail.com, ORCID ID:https://orcid.org/0000-0002-3896-9701;
Sagnayeva Saule — candidate of physico-mathematical sciences, associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: sagnayeva_sk@enu.kz, ORCID ID: https://orcid.org/0000-0001-7762-8531;
Orazbayev Batyr — doctor of technical sciences, professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: orazbayev_bb@enu.kz, ORCID ID: https://orcid.org/0000-0003-2109-6999;
Michael James Cardwell Crabbe — PhD, professor, Wolfson College, University of Oxford, Oxford, United Kingdom,
E-mail: james.crabbe@wolfson.ox.ac.uk, ORCID ID: https://orcid.org/0000-0003-3609-1963;
Dyussekeyev Kanagat — candidate of technical sciences, Head of Department, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: dyussekeyev_ka@enu.kz, ORCID ID: https://orcid.org/0000-0001-7691-2506.

Abstract. Currently, due to the increasing trend of water scarcity, it is necessary to use water resources efficiently and to develop and implement effective water accounting methods and automated water resource management systems. This study aims to develop an effective and adequate method, compared to existing ones, for calculating water flow for automated water resource management systems. The presented work examines existing groups of direct and indirect methods of water accounting, identifies their main limitations and characteristic errors that significantly reduce the accuracy of measurements. *Results.* Existing groups of direct and indirect methods and tools for water accounting were analyzed, and their inaccuracies, which lead to reduced precision, along with certain limitations, were identified. For more accurate and adequate water accounting in conditions of uneven, gradually varying water movement in a non-prismatic channel with horizontal and reverse bed slopes, an effective water accounting method is

proposed. *Scientific novelty.* The equations used for a more accurate calculation of water volumes in the channels of irrigation systems in various water flow regimes based on the developed computational scheme for uneven, gradually varying water movement are provided. It is assumed that taking into account the characteristics of real water flow will improve the accuracy of water metering in automatic systems. *Practical value.* The proposed method was tested under the conditions of the K-19 channel in the Turkestan region and demonstrated its efficiency and required level of accuracy. The method can be integrated into intelligent automated water management systems, improving water accounting in irrigation systems.

Keywords: water accounting, irrigation systems, automated water resource management systems (AWMS), direct methods, indirect methods, uneven gradually varying water flow

Acknowledgements. *The research data was sponsored by the Science Committee of the Minister of Science and Higher Education of the Republic of Kazakhstan (Grant No. of the research fund AP23490181 Development of an intelligent water management system and IoT).*

© Ж.А. Әбдібаев[1], С.К. Сагнаева[1], Б.Б. Оразбаев[1], М. Джеймс К. Крэбб[2], К.А. Дюсекеев*[1], 2025.

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;

[2]Вольфсон колледжі, Оксфорд университеті, Оксфорд, Біріккен Корольдігі.

E-mail: dyussekeyev_ka@enu.kz

СУ РЕСУРСТАРЫНЫҢ АВТОМАТТАНДЫРЫЛҒАН ЖҮЙЕЛЕРІНЕ СУАРУ ЖҮЙЕЛЕРІНДЕГІ СУ ЕСЕПТЕУДІҢ ТИІМДІ ӘДІСІН ӘЗІРЛЕУ

Әбдібаев Жанұзак Архабайұлы — докторант, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: abdibaev.zha@gmail.com, ORCID ID: https://orcid.org/0000-0002-3896-9701;
Сагнаева Сауле Кайроллиевна — физика-математика ғылымдарының кандидаты, қауымдастырылған профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: sagnayeva_sk@enu.kz, ORCID ID: https://orcid.org/0000-0001-7762-8531;
Оразбаев Батыр Бидайбекович — техника ғылымдарының докторы, профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: orazbayev_bb@enu.kz, ORCID ID: https://orcid.org/0000-0003-2109-6999;
Майкл Джеймс Кардуэлл Крэбб — PhD, профессор, Вольфсон колледжі, Оксфорд университеті, Оксфорд, Біріккен Корольдігі,
E-mail: james.crabbe@wolfson.ox.ac.uk, ORCID ID: https://orcid.org/0000-0003-3609-1963;
Дюсекеев Канагат Абетович — техника ғылымдарының кандидаты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: dyussekeyev_ka@enu.kz, ORCID ID: https://orcid.org/0000-0001-7691-2506.

Аннотация. Қазіргі уақытта су тапшылығының өсу тенденциясына байланысты су ресурстарын тиімді пайдалану, суды есепке алудың тиімді тәсілдерін және су ресурстарын автоматтандырылған басқару жүйелерін әзірлеу және енгізу қажет. Бұл зерттеудің мақсаты қолданыстағы әдістермен салыстырғанда, су ресурстарын автоматтандырылған басқару жүйелері үшін, суды есептеудің тиімді және адеваттығы жоғары тәсілін әзірлеу болып табылады. Ұсынылған жұмыс суды есепке алудың тікелей және жанама әдістерінің қолданыстағы топтарын зерттейді, олардың негізгі шектеулерін және өлшемдердің дәлдігін айтарлықтай төмендететін сипаттамалық қателерді анықтайды. *Нәтижелері.* Суды есептеудің тікелей және жанама әдістері мен құралдарының қолданыстағы топтары талданады, олардың дәлдігінің төмендеуіне әкелетін қателері мен кейбір шектеулері анықталды. Көлденең және кері түбі еңістері бар призматикалық емес арнадағы біркелкі емес, біртіндеп тегіс өзгеретін су қозғалысы режимдерінде суды көлемін дәлірек және адекватты есептеудің тиімді әдісі ұсынылған. *Ғылыми жаңалығы.* Су ағынының әртүрлі режимдеріндегі суару жүйелерінің каналдарындағы су көлемін дәлірек есептеу үшін қолданылатын теңдеулер біркелкі емес біртіндеп өзгеретін су қозғалысы үшін құрылған есептеу схемасы негізінде ұсынылған және сипатталған. Нақты су ағынының сипаттамаларын есепке алу автоматты жүйелерде суды есепке алудың дәлдігін жақсартады деп болжануда. *Практикалық құндылық.* Ұсынылған су есептеу тәсілі Түркістан облысында К-19 каналы жағдайында сынақтан өтіп, оның тиімділігі мен қажетті дәлдік деңгейін қамтамасыз ететіні анықталды. Бұл әдіс суару жүйелеріндегі суды есепке алуды жетілдіре отырып, суды басқарудың интеллектуалды автоматтандырылған жүйелеріне біріктірілуі мүмкін.

Түйін сөздер: су есептеу, суару жүйелері, су ресурстарын автоматтандырылған басқару жүйелері (СР АБЖ), тікелей тәсілдер, жанама тәсілдер, су ағынының біркелкі емес біртіндеп өзгеретін қозғалысы

© Ж.А. Абдибаев[1], С.К. Сагнаева[1], Б.Б. Оразбаев[1], М. Джеймс К. Крэбб[2], К.А. Дюсекеев*[1], 2025.

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

[2]Колледж Вольфсона, Оксфордский университет, Оксфорд, Соединенное Королевство.

E-mail: dyussekeyev_ka@enu.kz

РАЗРАБОТКА ЭФФЕКТИВНОГО МЕТОДА ВОДОУЧЕТА НА ОРОСИТЕЛЬНЫХ СИСТЕМАХ ДЛЯ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ВОДНЫМИ РЕСУРСАМИ

Абдибаев Жанузак Архабайулы — докторант, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: abdibaev.zha@gmail.com, ORCID ID: https://orcid.org/0000-0002-3896-9701;
Сагнаева Сауле Кайроллиевна — кандидат физико-математических наук, ассоциированный профессор, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: sagnayeva_sk@enu.kz, ORCID ID: https://orcid.org/0000-0001-7762-8531;

Оразбаев Батыр Бидайбекович — доктор технических наук, профессор, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: orazbayev_bb@enu.kz, ORCID ID: https://orcid.org/0000-0003-2109-6999;
Майкл Джеймс Кардуэлл Крэбб — PhD, профессор, Колледж Вольфсона, Оксфордский университет, Оксфорд, Соединенное Королевство,
E-mail: james.crabbe@wolfson.ox.ac.uk, ORCID ID: https://orcid.org/0000-0003-3609-1963;
Дюсекеев Канагат Абетович — кандидат технических наук, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,
E-mail: dyussekeyev_ka@enu.kz, ORCID ID: https://orcid.org/0000-0001-7691-2506.

Аннотация. В настоящее время в связи с тенденцией роста дефицита водных ресурсов необходимо эффективно использовать водных ресурсов, разработать и внедрить эффективных методов водоучета и автоматизированных систем управления водными ресурсами. Целью настоящего исследования является разработка эффективного и адекватного, по сравнению существующими способами, метода расчета расхода воды для автоматизированных систем управления водными ресурсами. В представленной работе подвергаются анализу существующие группы прямых и косвенных методов водоучета, выявлены их основные ограничения и характерные погрешности, существенно снижающие точность измерений. Результаты. Проанализированы существующие группы прямых и косвенных методов и средств водоучета, выявлены их погрешности, приводящие к снижению их точности, и некоторые ограничения. Для проведения более точного и адекватного водоучета в режимах неравномерного плавно изменяющегося движения воды в непризматическом русле с горизонтальным и обратным уклонами дна предложен эффективный метод водоучета. Научная новизна. Приведены уравнения, используемые для более точного расчета объема воды в каналах оросительных систем в различных режимах течения воды на основе созданной расчетной схемы движения неравномерного плавно изменяющегося движения. Предполагается, что учёт особенностей реального водного потока позволит повысить точность водоучета в автоматических системах. Практическая ценность. Предложенный метод испытан в условиях канала К-19 Туркестанской области и показал свою эффективность и необходимый уровень точности. Метод может быть интегрирован в интеллектуальные автоматизированные системы управления водными ресурсами, улучшая водоучёт в оросительных системах.

Ключевые слова: водоучет, оросительные системы, автоматизированные системы управления водными ресурсам (АСУ ВР), прямые методы, косвенные методы, неравномерное плавно изменяющееся движение потока воды

Introduction. The water resources of the Republic of Kazakhstan are characterized by their scarcity compared to many other countries. The majority of water resources are supplied by the country's surface waters. Of these, 56% are formed locally, while the remaining 44% come from the runoff of transboundary rivers from Russia, China, Uzbekistan, and Kyrgyzstan. This indicates Kazakhstan's

dependence on transboundary rivers flowing in from neighboring countries (Yespolov et al., 2022). This situation significantly increases the importance of regulating transboundary flows to address the country's existing and potential water-related issues. Additionally, water resources for irrigation in Kazakhstan are distributed unevenly, and some regions experience water shortages for irrigation during summer.

In recent years, the growing trend in water consumption and the declining availability of water resources pose a threat of increasing shortages, especially in the irrigation systems of the southern regions of the country. If the efficiency of water accounting and management systems is not improved, water scarcity will intensify in the coming years, negatively affecting the water supply for the population, irrigation systems, and the state of the environment.

To improve the efficient use of water resources in Kazakhstan, it is necessary to establish reliable accounting for the collection, transportation, and distribution of water at all levels of water management facilities. This should be based on effective water accounting methods and automated water resource management systems (AWMS). However, since issues of uncertainty and fuzziness of some initial data often arise, AWMS needs to be supplemented with elements of intellectualization to ensure their efficiency. In this regard, developing an effective water accounting method for creating AWMS is a relevant scientific and practical challenge for Kazakhstan. This situation motivated this study, dedicated to developing an effective water accounting method, compared to existing methods (Bochkarev, 2012; Masumov, 2015). for use in the creation of intellectualized AWMS.

Materials and methods. The materials for this study include various sources, methods, and approaches to water accounting used in water accounting and water resource management systems within irrigation systems. Since water is one of the most scarce natural resources in the world, including in Kazakhstan, it is essential to manage the distribution and usage of water resources efficiently. In the process of water resource management, water accounting methods and water volume measurements play a critical role.

Many studies have examined water accounting and water accounting methods (Pacheco et al., 2020; Albaina et al., 2023; Arregui et al., 2010). For instance, Pacheco et al. conducted a techno-economic analysis of water accounting devices (Pacheco et al., 2020) while Albaina et al. explored the effects of various accessories installed upstream of large water meters (Albaina et al., 2023). These studies proposed approaches that combine technical and economic research, enabling more accurate assessments of measurement errors and the influence of various factors on measurement quality. The technical research includes methodologies for measuring errors that occur at different flow rates and analyzing the resulting measurements, while economic studies use various methods to determine when to replace water meters based on management needs.

Arregui et al. developed a graphical method for calculating the optimal

replacement period for water meters and proposed tools that integrate observation data with global databases to enhance data collection for water resource accounting (Arregui et al., 2010) The authors of other studies (Kamienski et al., 2019; Manimegalai et al., 2020) introduced an IoT-based platform for intelligent water resource management and a smart irrigation system with monitoring. Madias et al. demonstrated the effectiveness of smart water meters for consumers (Madias et al., 2023) However, these and other studies do not explore methods and tools for water accounting that can effectively distribute water resources across irrigation systems under various water flow conditions (e.g., with straight, horizontal, or reverse bed slopes).

The tools and methods for water accounting that can be used in irrigation systems are considered in the study (Bochkarev et al., 2013). According to modern legal and regulatory documents, direct and indirect methods are used in practice as standardized measurement methods.

Direct methods of water flow measurement are used for metrological testing, calibration, and certification of flow meters and other specialized equipment. Direct methods include volumetric, gravimetric, and volumetric-hydraulic approaches, which rely on various calculation dependencies. Because volumetric and volumetric-hydraulic methods require significant financial resources, they are primarily used for calibration of working measurement tools as per metrological standards. The volumetric method involves measuring volume and height, but its accuracy heavily depends on the devices, tools, and the operator's skill, making it less precise.

The gravimetric method offers higher accuracy compared to other water accounting methods, with errors mainly arising from determining the weight of the water, which is measured more accurately.

The **channel method** involves determining water flow and runoff by observing water levels at a control section where the relationship between the flow rate (Q) passing through the section and the corresponding water levels (H) is pre-established: $Q=f(H)$. The channel method does not require the construction of special facilities. However, it is not operationally efficient, cannot be automated, and measurement errors can occur due to channel deformation caused by erosion or silting. On fixed channels, where there are no backwater or recession effects and stable sections exist, the channel method ensures an acceptable level of accuracy for water accounting. However, there are specific conditions for applying this method: during water flow measurements, the bed and banks of the natural or artificial watercourse must be sufficiently stable. Currently, the channel method remains the primary approach for water accounting at key irrigation sources, headworks, and accounting points of main and large canals.

The hydraulic method is applied when water flow through a structure depends on the fixed dimensions of some of its elements. This method is based on the installation of hydrometric facilities or devices at specific points in irrigation systems. These facilities or devices allow the measurement of water flow passing through them.

Tracer methods involve introducing markers into the water flow, such as floats, salt solutions, or other materials. In practice, the channel and hydraulic methods are the most widely used for water accounting. In the next section, the authors of this study propose and describe a more efficient water accounting method for irrigation systems designed for automated water resource management systems.

Results and discussion. The following efficient water accounting method is proposed, which utilizes modern tools to ensure a high level of accuracy and operational efficiency in measuring water flow within automated systems. This method involves the use of water level sensors and controller equipment integrated into an automated water accounting system. For this study, the K-19 canal located in the Turkestan region, designed for irrigation systems in the area, was selected. An automated water accounting system was installed at the canal, incorporating water level sensors and controller equipment for real-time water accounting.

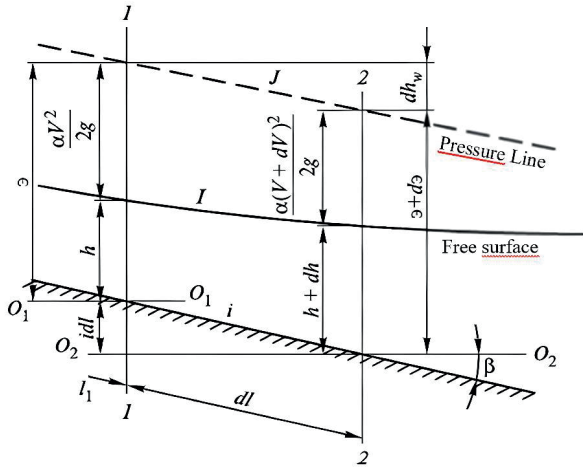
Based on the analysis and verification conducted, the proposed water accounting method demonstrated its efficiency, considering the existing backwater conditions in the K-19 canal. Graduated tables showing the relationship between flow rates and water levels were developed, and the consistency of the water accounting process with the automation requirements for this canal was evaluated. The method was validated for canals like K-19 that operate under backwater conditions.

The research utilized high-precision, calibrated instruments, including the Nikon XF laser total station, the dual-frequency GPS GS-1200 navigation device, and the GR-21M hydrometric propeller.

For calculating water flow, sensors and a tabular flow meter were used to measure water level and compute flow rates based on predefined relationships between level and flow.

The velocity regime across the K-19 canal's cross-section was studied using hydrometric techniques, with measurements of water slope and channel bed slope conducted using the laser total station and dual-frequency navigation device. The results showed that factors such as channel bed silting, overgrowth, and obstructing structures creating backwater conditions significantly influence the canal's average velocity and throughput capacity. These factors result in a dynamic regime known as "established uneven gradually varying water flow."

Established uneven gradually varying water flow is characterized by changes in flow properties along its length. If these changes (e.g., velocity and flow rate) occur gradually over the flow's length, the flow is described as gradually varying. The primary task in studying such uneven flow in an open channel is to determine the variation in depth along the flow, i.e., to construct the curve of the free surface of the water flow. The computational scheme for uneven gradually varying water flow is presented in Figure 1.



J – Pressure line; *I* – Free surface; *α* – Kinetic energy coefficient; *V* – Flow velocity; *g* – Acceleration due to gravity; *h* – Height from the channel bed to the free surface of the water; *i* – Variable bed slope; *l* – Length from the horizontal line $O_2 - O_2$ to the horizontal line $O_1 - O_1$ from the beginning of the slope to the intersection with vertical line 1 – 1; *β* – Angle between the end of the slope and the horizontal line $O_2 - O_2$; э – Distance between the free surface and the pressure line; h_w – Distance from the pressure line to the intersection with the horizontal line from the start of the pressure line.

Figure 1. Computational scheme for uneven gradually varying water flow

The equation for uneven gradually varying flow in a non-prismatic channel with a straight sloped bed can be written as:

$$\frac{dh}{dl} = \frac{i_0 - \frac{Q^2}{\omega^2 C^2 R} + \frac{\alpha Q^2}{g\omega^3}\frac{d\omega}{dl}}{1 - \frac{\alpha Q^2 B}{g\omega^3}} . \quad (1)$$

where *dh/dl* – change in flow depth along the channel length; i_0 – bed slope of the channel; *Q* – water discharge in the channel; *α* – kinetic energy coefficient; *ω* – flow cross-sectional area; *C* – chezy coefficient (hydraulic coefficient); *R* – hydraulic radius, defined as the ratio of the flow cross-sectional area to the wetted perimeter; *B* – width of the channel; *g* – acceleration due to gravity.

Similar expressions, considering the slope's sign, can be derived for prismatic channels with horizontal and reverse bottom slopes. In prismatic channels, the live cross-sectional area of the flow can only vary due to changes in depth. By substituting *dω/dl*=0 into Equation (1), we obtain the differential equation for uneven, gradually varying flow in prismatic channels with a positive bottom slope:

$$\frac{dh}{dl} = \frac{i_0 - \dfrac{Q^2}{\sim^2 C^2 R}}{1 - \dfrac{aQ^2 B}{g\sim^3}}. \qquad (2)$$

Introducing the (2) kinetic parameter and using the concept of the flow characteristic $K = \sim C\sqrt{R}$ for an arbitrary depth h of uneven flow, we derive:

$$\frac{dh}{dl} = \frac{i_0 - Q^2/K^2}{1 - \Pi_K}, \qquad (3)$$

where K – flow characteristic; $\Pi_к$ – kinetic parameter. Other parameters are as described above.

Expressing the flow rate Q using Chezy's formula via the flow characteristic K_0, which corresponds to the normal depth h_0 in the channel for a given slope i_0, we can write:

$$\frac{dh}{dl} = i_0 \frac{1 - (K_0/K)^2}{1 - \Pi_K},$$

where K_0 – is the flow characteristic corresponding to the normal depth.

Using the hydraulic index of the channel:

$$\left(\frac{K_1}{K_2}\right)^x = \left(\frac{h_1}{h_2}\right)^x. \qquad (4)$$

We obtain the equation for uneven flow in prismatic channels of regular shape:

$$\frac{dh}{dl} = i_0 \frac{1 - (h_0/h)^x}{1 - \Pi_K}. \qquad (5)$$

For prismatic channels with a horizontal bottom i_0=0:

$$\frac{dh}{dl} = -\frac{\dfrac{Q^2}{K^2}}{1 - \Pi_K}.$$

For channels with a reverse slope (i_0<0)

$$\frac{dh}{dl} = -\frac{i_0 + Q^2/K^2}{1 - \Pi_K}.$$

When considering differential equations (3) and (4), the flow rate Q should be taken as constant. The variables are the flow characteristic K and the kinetic parameter $\Pi_к$, as they depend on the cross-sectional characteristics of the flow ω,

x, B, R, C, which vary along the length of the prismatic channel due to changes in depth *h* during uneven flow. It is evident that at certain depth values *h* the flow characteristic *K* and the kinetic parameter $\Pi_к$ may take such values that the term in parentheses in the denominator of the right-hand side of these equations approaches zero.

For channels with a slope i_0>0 when the numerator of equation (3) equals zero, we obtain:

$$i_0 - Q^2 = 0. \quad (6)$$

From this, *dh/dl*=0, which corresponds to the constancy of flow depth along the channel, i.e., uniform flow ($h=h_0$). This is also directly derived from expression (6), which represents Chezy's formula $Q=K_0\sqrt{i_0}$ for uniform flow. Thus, it is confirmed that uniform flow is possible in a prismatic channel with a positive bottom slope i_0>0. The derivative *dh/dl=tg*0, where 0 – is the angle between the tangent to the curve of the free water surface and the line *N–N* of normal depth or the line *K–K* of critical depth. Consequently, if the depth of uneven flow in a channel with a slope i_0>0 approaches the normal depth $h \to h_0$, then *dh/dl=tg*0→0, meaning the free surface asymptotically approaches the *N–N* line.

Based on the theoretical description of established uneven gradually varying water flow in the K-19 channel, it was determined that to obtain accurate discharge data while simultaneously considering various factors affecting the hydraulic flow regime (artificial backwater, backwater in branch channels, siltation, and vegetation growth), the water flow chart must be frequently adjusted. Thus, the results of the velocity study along the length of the K-19 channel indicate that the average velocity and discharge are influenced by the backwater regime created by obstructive structures along the channel, siltation of the channel bed, and vegetation growth. Collectively, these factors create a dynamic regime referred to as "established uneven gradually varying water flow." This type of flow impacts the velocity regime of water in the channel and leads to changes in discharge at the same water level. Therefore, it is necessary to continuously adjust the water flow chart for a backwater-type channel like K-19.

During the research, the installed equipment (a table-based flow meter) in the pilot project on the K-19 channel functioned properly. Water flow rate calculations were performed using the provided graduated level/discharge tables. However, discrepancies in water discharge measurements between the hydrometric propeller and the installed devices at the head section of K-19 were due to the absence of corrections in the water flow chart. The graduated dependencies (charts) provided by the RGP "Su-Metrology" branch did not account for the backwater regime, indicating that the methodology for constructing these charts should consider established uneven gradually varying water flow.

Conclusions: The proposed and described method for measuring water flow in open channels transitions traditional accounting methods into a digital format.

However, it is essential to calculate and implement adjustments in the water accounting sensors' programs based on the operational regimes of the channel (e.g., backwater, siltation, vegetation growth). The program must automatically adjust water accounting when backwater conditions arise. This is the primary reason for inaccuracies in sensor readings. This observation applies to all sensors using ultrasonic methods to measure water flow levels, regardless of the manufacturer or brand. These sensors operate based on the following principle: the sensor emits ultrasonic pulses, the product surface reflects these pulses, and the sensor registers them again. The time taken for the reflected ultrasonic signal to return is directly proportional to the distance traveled.

Based on the conducted research, it can be concluded that despite operational difficulties when using modern water accounting devices in channels with slight slopes operating in automatic mode, automating the water accounting process is considered feasible. However, it requires continuous adjustment of the water flow chart, depending on the flow type and the operational state of the channel. Regarding the accuracy and reliability of the data provided by the water accounting sensors studied in the K-19 channel case, it can be affirmed that sensors of this class and type operate correctly within the specified technical accuracy limits.

References

Albaina I., Bidaguren I., Izquierdo U. et al. (2023) Influence of Various Accessories Upstream Large Water Meters. Water Resources Management. 37. — P. 4693–4708. https://doi.org/10.1007/s11269-023-03573-2. (in English)

Arregui F.J., Espert V. (2010) A graphical method to calculate the optimum replacement period of water meters. Journal of Water Resources Planning and Management. Volume 137, Issue 1. — P. 143-146. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000100. (in English)

Bochkarev V.Ya. (2012) Novyye tekhnologii i sredstva izmereniy, metody organizatsii vodoucheta na orositel'nykh sistemakh [New technologies and measuring instruments, methods of organizing water metering in irrigation systems]. V.Ya. Bochkarev; FGBNU "RosNIIPM". — Novocherkassk. — P.227 http://cawater-info.net/bk/improvement-irrigated-agriculture/files/bochkarev.pdf. (in Russian)

Kamienski C., Soininen J-P., Taumberger M., Dantas R., Toscano A., Cinotti T.S., Maia R.F., André Neto A.T. (2019) Smart Water Management Platform: IoT-Based Precision Irrigation for Agriculture. Sensors.https://doi.org/10.3390/s19020276. (in English)

Madias K., Szymkowiak A., Borusiak B. (2023) What builds consumer intention to use smart water meters – Extended TAM-based explanation. Water Resources and Economics. Vol. 44, 100233. https://doi.org/10.1016/j.wre.2023.100233. (in English)

Manimegalai V., Little Judy A., Gayathri A., Ashadevi S., Mohanapriya V. (2020) Smart Irrigation System with Monitoring and Controlling using IoT. International Journal of Engineering and Advanced Technology. Vol. 9, No 4. — P. 1373-1376. https://doi.org/10.35940/ijeat.C6586.049420. (in English)

Masumov R.R. (2015) Metody izmereniya raskhoda vody na rekakh i kanalakh, v napornykh truboprovodakh nasosnykh stantsiy i orositel'nykh system [Methods of measuring water flow in rivers and canals, in pressure pipelines of pumping stations and irrigation systems]. R.R. Masumov. — Tashkent: Scientific and information center MKVK. — P. 84 http://cawater-info.net/library/rus/watlib/watlib-11-2015.pdf. (in Russian)

Pacheco V.F, Valdés R.E, Gil E.B. et al (2020) Techno-economic analysis of residential water meters: A practical example. Water Resources Management. 34: 2471–2484. https://doi.org/10.1007/s11269-020-02564-x. (in English)

V.Ya. Bochkarev, Ya.V. Bochkarev (2013) Avtomatizatsiya vodoraspredeleniya na kanalakh

orositel'nykh sistem ravninnoy zony metodom neposredstvennogo otbora raskhodov [Water-distribution Automation at the Canals of Plain Irrigation Systems by the Method of Direct Discharge Withdrawal]. Scientific journal of the Russian Research Institute of Land Reclamation Problems, No. 1(09). — P. 32-41. https://rosniipm-sm.ru/dl_files/udb_files/udb4-rec622-field12.pdf. (in Russian)

Yespolov T., Tireuov K., Kerimova U. (2022) Vodnyye resursy v sel'skom khozyaystve Respubliki Kazakhstan: vzglyad uchenykh na ratsional'noye ispol'zovaniye, perspektivy i upravleniye [Water resources in agriculture of the Republic of Kazakhstan: a view of scientists on rational use, prospects and management]. Problems of AgriMarket. (3):155-163. https://doi.org/10.46666/2022-3.2708-9991.17. (in Russian)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 78–92

https://doi.org/10.32014/2025.2518-1726.365

ҒТМР 20.19.27:
ОӘЖ 004.8

Zh. Bazarbek, N. Toyganbaeva*, M. Mansurova, T. Sarsembayeva, M. Sakypbekova, 2025.
Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: bodinaz@mail.ru

DEVELOPING A DATASET FOR CREATING A LARGE LANGUAGE MODEL (LLM) FOR THE KAZAKH LANGUAGE

Madina Mansurova — Candidate of Physico-mathematical Sciences, Professor, Head of the Department of Artificial Intelligence and Big Data at Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Madina.Mansurova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-9680-2758;
Toyganbaeva Nazgul — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: bodinaz@mail.ru, ORCID ID: https://orcid.org/0000-0003-2661-8661;
Bazarbek Zhaniya — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Zhaniya.Bazarbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-7838-2104;
Sarsembayeva Talshyn — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: talshyn.sagdatbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0001-7668-2640;
Sakypbekova Meruyert — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Meruert.Sakypbekova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-6652-1357.

Abstract. This study focuses on the development of a Large Language Model (LLM) for the Kazakh language and provides a comprehensive overview of its theoretical foundations, methodological aspects, and practical applications. A model designed with consideration of the morphological, syntactic, dialectal, and orthographic features of the Kazakh language represents a significant step in advancing artificial intelligence. The relevance of this research lies in the need to create intelligent systems capable of professionally processing Kazakh texts and providing users with accurate and well-grounded responses. Within the framework of the project, a specialized dataset was collected and structured. The purpose of this article is to create a premium dataset suitable for LLM training by collecting data in the Kazakh language and presenting it in a high-quality and accessible form. The dataset development addressed key challenges such as filling gaps in linguistic materials, covering dialectal diversity, incorporating orthographic variations, and

including diverse usage scenarios. The application of OCR technology enabled the digitization of materials from multiple sources and their transformation into formats convenient for processing. Furthermore, methods for annotation, structuring, and systematization of data were proposed, contributing to improved model reliability and accuracy. The study also analyzes advanced methodologies such as MBERT and GPT, emphasizing their limitations in processing the Kazakh language. The importance of building unique datasets for low-resource languages is particularly highlighted. The findings demonstrate the potential for applying AI in government, education, healthcare, business, and digital services. Thus, this work contributes to reducing informational inequality, integrating the Kazakh language into the global AI ecosystem, and fostering Kazakhstan's technological progress.

Keywords: Kazakh language, large language model (LLM), mBERT, dataset, natural language processing (NLP)

Acknowledgment: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001-OT-24).

Ж.П. Базарбек, Н.А. Тойганбаева*, М.Е. Мансурова, Т.С. Сарсембаева, М.Ж. Сакыпбекова, 2025.
Әл-Фараби атындағы Қазақ ұлттық университеті.
E-mail: bodinaz@mail.ru

ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН ҮЛКЕН ТІЛ МОДЕЛІН (LLM) ЖАСАУ ҮШІН DATASET ӘЗІРЛЕУ

Мансурова Мадина Есимхановна — ф.-м. ғ.к., профессор, әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының меңгерушісі, Алматы, Қазақстан,
E-mail: Madina.Mansurova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-9680-2758;
Тойганбаева Назгуль Абеновна — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: bodinaz@mail.ru ORCID ID: https://orcid.org/0000-0003-2661-8661;
Базарбек Жания Пархатқызы — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: Zhaniya.Bazarbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-7838-2104;
Сарсембаева Талшын Сағдатбекқызы — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: talshyn.sagdatbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0001-7668-2640;
Сакыпбекова Меруерт Жумабековна — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: Meruert.Sakypbekova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-6652-*1357,*

Аннотация. Бұл зерттеу қазақ тілі үшін үлкен тіл моделін (LLM) құруға арналған және оның ғылыми-теориялық негіздерін, әдістемелік қырларын, сондай-ақ практикалық қолдану мүмкіндіктерін кешенді қарастырады. Қазақ тілінің морфологиялық, синтаксистік, диалектілік және

әртүрлі графикалық жүйелердегі ерекшеліктерін ескеріп жасалған модель жасанды интеллектті дамытудағы маңызды қадам болып саналады. Мұндай бағыттың маңыздылығы қазақ тілінде кәсіби деңгейде мәтіндермен өзара әрекеттесетін, пайдаланушы сұрақтарына толық әрі орынды жауап беретін интеллектуалды жүйелерді қалыптастыру қажеттілігімен байланысты. Жоба аясында арнайы dataset жиналып, құрылымдалды. Осы мақаланың мақсаты – қазақ тіліндегі деректерді жинақтап, сапалы әрі қолжетімді формада ұсыну арқылы LLM оқытуға жарамды премиум-жиынтық құру. Dataset әзірлеу барысында жетіспейтін мәтіндік ресурстарды толықтыру, диалектілерді қамту, түрлі орфографиялық нұсқаларды енгізу және әртүрлі сценарийлерді ескеру мәселелері шешілді. OCR технологиясын қолдану деректерді әртүрлі көздерден цифрландыруға және өңдеуге ыңғайлы пішімдерге түрлендіруге мүмкіндік берді. Сонымен бірге деректерді аннотациялау, құрылымдау және жүйелеу әдістері ұсынылып, олардың модельдің сенімділігі мен дәлдігін арттыруға ықпалы айқындалды. Зерттеу MBERT және GPT сияқты алдыңғы қатарлы әдіснамаларды талдауды қамтиды. Бұл модельдердің қазақ тілімен жұмыс істеудегі шектеулері көрсетіліп, ресурсы шектеулі тілдер үшін арнайы dataset қалыптастырудың маңыздылығы ерекше атап өтілді. Жобаның нәтижелері мемлекеттік басқару, білім беру, медицина, бизнес және цифрлық қызмет көрсету салаларында жасанды интеллектті кеңінен қолдануға мүмкіндік береді. Осылайша, атқарылған жұмыс қазақ тілінің ақпараттық теңсіздігін азайтуға, оны жаһандық жасанды интеллект экожүйесіне енгізуге және Қазақстанның технологиялық дамуына жаңа серпін беруге бағытталған.

Түйін сөздер: жасанды интеллект, қазақ тілі, табиғи тілді өңдеу, машиналық оқыту, үлкен тіл моделі (LLM), mBERT, GPT синтетикалық деректер

Алғыс: Бұл зерттеу Қазақстан Республикасының Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (Грант № BR24993001-OT-24).

Ж.П. Базарбек, Н.А. Тойганбаева*, М.Е. Мансурова, Т.С. Сарсембаева, М.Ж. Сакыпбекова, 2025.

Казахский Национальный университет им. аль-Фараби, Алматы, Казахстан.
E-mail: bodinaz@mail.ru

РАЗРАБОТКА ДАТАСЕТА ДЛЯ СОЗДАНИЯ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ (LLM) ДЛЯ КАЗАХСКОГО ЯЗЫКА

Мансурова Мадина Есимхановна — к.ф.-м.-н., профессор, Заведующая кафедрой Искусственного интеллекта и Big Data КазНУ имени Аль-Фараби, Алматы, Казахстан,
E-mail: Madina.Mansurova@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0002-9680-2758;
Тойганбаева Назгуль Абеновна — старший преподаватель кафедры Искусственного интеллекта и Big Data КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: bodinaz@mail.ru, ORCID ID: https://orcid.org/0000-0003-2661-8661;

Базарбек Жания Пархатқызы — старший преподаватель кафедры Искусственного интеллекта и Big Data КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: Zhaniya.Bazarbek@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0002-7838-2104;
Сарсембаева Талшын Сағдатбекқызы — старший преподаватель кафедры Искусственного интеллекта и Big Data КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: talshyn.sagdatbek@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0001-7668-2640;
Сакыпбекова Меруерт Жумабековна — Әл-Фараби атындағы ҚазҰУ жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Казахстан,
E-mail: Meruert.Sakypbekova@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0002-6652-1357.

Аннотация. Данное исследование посвящено созданию большой языковой модели (LLM) для казахского языка и комплексно рассматривает её научно-теоретические основы, методологические аспекты, а также практические возможности применения. Модель, разработанная с учётом морфологических, синтаксических, диалектологических и графических особенностей казахского языка, является важным шагом в развитии искусственного интеллекта. Актуальность данного направления обусловлена необходимостью формирования интеллектуальных систем, способных профессионально работать с текстами на казахском языке и давать пользователям полные и обоснованные ответы. В рамках проекта был собран и структурирован специализированный dataset. Цель данной статьи – создание премиального набора данных, пригодного для обучения LLM, за счёт систематизации и представления казахскоязычных текстов в удобной форме. В процессе разработки были решены задачи восполнения недостатка языковых материалов, охвата диалектов, включения различных орфографических вариантов и сценариев использования. Применение технологии OCR позволило оцифровать материалы из различных источников и преобразовать их в удобные для обработки форматы. Кроме того, были предложены методы аннотирования, структурирования и систематизации данных, способствующие повышению точности и надёжности моделей. Исследование включает анализ передовых методологий, таких как MBERT и GPT, с указанием их ограничений при работе с казахским языком. Особо подчёркнута значимость формирования уникальных наборов данных для языков с ограниченными ресурсами. Полученные результаты открывают возможности широкого применения ИИ в сфере государственного управления, образования, медицины, бизнеса и цифровых услуг. Таким образом, проделанная работа направлена на сокращение информационного неравенства, интеграцию казахского языка в глобальную экосистему искусственного интеллекта и придание нового импульса технологическому развитию Казахстана.

Ключевые слова: искусственный интеллект, казахский язык, обработка естественного языка, машинное обучение, большая языковая модель (LLM), mBERT, GPT синтетические данные

Благодарность: Данное исследование было профинансировано Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант № BR24993001-OT-24).

Кіріспе. Үлкен Тілдік Модельдер (LLM) — бұл табиғи тілді адамға жақын деңгейде құруға және өңдеуге қабілетті жасанды интеллект. Олар автоматты түрде аудармада, мәтінді талдауда, дауыстық көмекшілерде және басқа салаларда қолданылады. Қазақ тілінің бірегей морфологиялық және синтаксистік ерекшеліктері бар. Осыған сүйене отырып, ағылшын және басқа да әлемдік тілдерге бағытталған қолданыстағы LLM-ді өңдеу қиын. Мамандандырылған модельді құру қазақ тіліндегі мәтіндерді неғұрлым нақты және табиғи түсінуді және қалыптастыруды қамтамасыз етеді. Тілдік модельдің сапасы оқыту деректерінің көлемі мен оның түрлілігіне тікелей байланысты. Деректер қоры неғұрлым бай және таза болса, модельдің нәтижелері соғұрлым жақсы болады. Дереккөздердің сан түрлілігін ескеру және әртүрлі стильдер мен диалектілер арасындағы тепе-теңдікті сақтау маңызды.

Үлкен тілдік модельдер саласындағы зерттеулер табиғи тілдерді өңдеу технологияларының қарқынды дамуын көрсетеді. mBERT сияқты трансформаторлық модельдердің пайда болуы маңызды кезеңдердің бірі болды (Devlin et al., 2018). GPT-3 және GPT-4 модельдері жоғары сапалы мәтіндерді жасауды, автоматты аударманы, чатбот жасауды және басқа да көптеген тапсырмаларды орындай алды (OpenAI 2023). Тиісті зерттеу бағыттарының ішінде мыналар бар: Дәл баптау: мамандандырылған деректер корпусында үлкен үлгілерді оқыту медицина, құқық және техникалық аударма сияқты жоғары мамандандырылған салаларда дәлдікті арттыруға мүмкіндік береді (Howard & Ruder, 2018). Көптілді модельдер үшін зерттеулерде көрсеткендей, mBERT және BLOOM сияқты модельдер мәтіндерді бірнеше тілде өңдей алады, бірақ ресурстары аз тілдерді өңдеу сапасы әлі де үлкен мәселе болып қала береді. LLM-нің біржақтылығы мен этикасында ғалымдар деректердің модельдік бейімділікке әсерін және деректерді сүзу және теңгерімдеу арқылы оны азайту жолдарын зерттейді (Bender et al., 2021). Никбахт Р. Және басқа ғалымдардың ұсынған ғылыми жұмысында 3gpp (3-Generation Partnership Project) техникалық сипаттамаларын түсіну үшін үлкен тілдік үлгілерді (LLM) оқытуға арналған ашық деректер жинағын ұсынады. Зерттеушілер деректер жиынтығының сапасын бағалау үшін сұрақтарға жауап бермес бұрын тиісті ақпаратты алу үшін RAG жүйесін қолданған (Nikbakht et al. 2024). Есептеу шығындарын оңтайландыру және азайтуда жаңа зерттеулер модельдердің көлемін азайтуға және олардың сапасын айтарлықтай жоғалтпай тездетуге бағытталған. Зерттеудің бұл бағыттары қазақ тіліне арналған мамандандырылған модельдерді қоса алғанда, LLM-нің дамуына негіз болады.

Материалдар мен әдістер.

Қазақ тілінің морфологиялық және синтаксистік күрделілігі. Қазақ тілі агглютинативті болып табылады. Яғни, бұл сөздердің мағынасын өзгертетін көптеген аффикстердің болуын білдіреді. Бұл мәтінді лемматизациялау мен токенизациялауды қиындатады. Латын, кириллица, қысқартылған жазу мәселелері бойынша қазақ тілі бірнеше алфавиттерді қолданады: тарихи контексте кириллица, латын және араб жазуы. Осылайша аралас пайдалану мәтіндерді өңдеуде қиындықтар туғызады. Диалектілер және стилистикалық ерекшеліктер, диалектілердегі айырмашылықтар мен сөйлеу мәнері (ресми, ауызекі, жастар жаргоны) әмбебап модель құру үшін деректерді жинауды және қалыпқа келтіруді талап етеді. Бұл морфологиялық күрделілік LLM әзірлеуде ерекше назар аударуды талап етеді:

- Көптік аффикстер жүйесі: кітап → кітаптар → кітаптарым
- Көптеген септік формалары: кітап → кітапқа, кітаптан, кітаппен
- Диалектілік және стилистикалық айырмашылықтар: ресми (Үкімет шешім қабылдады), ауызекі (Бүгін кешке кездесеміз бе?), жастар жаргоны (Түсінбедім, қазір тексеремін)
- Әліпбилердің әртүрлілігі: кириллица, латын графикасы, араб жазуы

Бұл ерекшеліктерді ескермей құрылған модель қазақ тіліндегі мәліметтерді дәл өңдей алмайды. №1 кестеде LLM мен басқа да машиналық оқыту жобалары үшін қажетті дерекқорларды құруда қолданылатын негізгі дерек көздері әртүрлі болады.

№1 кесте - Деректерді жинау және өңдеу

Ашық дереккөздер	
Дереккөз түрі	Мысалдар
Кітаптар, Мақалалар, энциклопедиялар	Әдеби шығармалар, ғылыми басылымдар, газеттер
Жаңалықтар сайттары, блогтар, форумдар	Онлайн басылымдар, талқылау платформалары
Әлеуметтік желі	Платформаларда пайдаланушы жасаған мазмұнды мәтіндер Twitter, Facebook
Жабық дереккөздер	
Диссертациялық жұмыстар	Философия докторы (PhD) дәрежесін алу үшін дайындалған диссертациялар
Кітаптар, энциклопедиялар	Әл-Фараби атындағы ҚазҰУ кітапханасынан әдеби шығармалар, ғылыми басылымдар, кітаптар

Мәтіндерді алдын ала өңдеу қадамдары. Бірінші, дубликаттарды жою – бірдей мәтіндердің бірнеше рет қолданылуын болдырмау. Екінші, орфографиялық және грамматикалық қателерді түзету. Дұрыс белгіленген деректер сапаны арттырады. Үшінші, қалыпқа келтіру (Normalization) – жазу стилін біріздендіру. Төртінші форматтау және құрылымдау – JSON, XML немесе CSV форматтарында сақтау.

Датасетті "docx" форматында құжаттарды автоматты түрде өңдеуге арналған бағдарлама әзірленді және python ортасында өңделді. Негізгі мақсат – мәтінді шығару, оны жеке сөйлемдерге бөлу және одан әрі талдау немесе өңдеу үшін `txt` форматында сақтау. Және ол сол құжаттағы басқа тілдерді, формула, суреттерді жойып, жаңа форматқа сақтайды.

Негізгі жұмысы келесідей:

1. Бағдарлама 'docx' файлын ашады және барлық мәтінді шығарады.

2. Содан кейін мәтін артық бос орындардан, жолдарды тасымалдаудан және басқа қажетсіз таңбалардан тазартылады.

3. Алынған мәтін тыныс белгілері ережелері арқылы сөйлемдерге бөлінеді.

4. Барлық сөйлемдер ' txt ' файлында сақталады, мұнда әр сөйлем бөлек жолда орналасады.

Бұл тәсіл мәтіндік ақпаратты құрылымдалған түрде ыңғайлы сақтауға, сондай-ақ оны әрі қарай талдау, табиғи тілді өңдеу және басқа да тапсырмалар үшін пайдалануға мүмкіндік береді.

Қазақ тіліне арналған LLM құрудағы негізгі қиындықтардың бірі — деректер көлемінің шектеулі болуы. Осы мәселені шешу үшін синтетикалық деректерді генерациялау әдістері қолданылады, оның ішінде мәтіндерді парафразалау, синонимдерді қолдану, статистикалық және нейрондық генерациялау тәсілдері қолданылады. Сондай-ақ, деректерді қолмен аннотациялау және белгілеу процесі маңызды рөл атқарады, бұл деректердің сапасын арттырып, модельдің түсіну қабілетін жақсартады. Сонымен қатар, интернеттен және түрлі сандық архивтерден автоматты түрде деректер жинау әдістері де қарастырылады.

Dataset сапасын бағалау үшін толықтық, әртүрлілік, синтаксистік дұрыстық және белгілеу стандарттарына сәйкестік көрсеткіштері қолданылады. Деректердің теңгерімділігі, стильдік вариативтілігі және тақырыпты қамту деңгейі де маңызды. Сонымен қатар, корпус сапасын жақсарту үшін автоматтандырылған валидация және адамдық тексеру әдістері қатар қолданылады. Машиналық оқыту жүйелері үшін деректерді бөлудің (train/test/validation split) дұрыстығын сақтау модельдің генерализациялау қабілетін арттырады.

LLM үшін dataset таңдау және дайындау – оның өнімділігі мен тиімділігіне тікелей әсер ететін маңызды қадам. №2 кестеде көптеген зерттеу мен өндірістік жобаларда қолданылатын дайын Dataset-тер.

№2 - Датасеттерге шолу

Датасеттер	**Сипаттамасы**	**Жылы**	**Қолжазба тілі**
Common Crawl (Common Crawl, 2025)	шамамен 1 триллион сөздер бар.	2007	ағылшын
The Pile (EleutherAI, EleutherAI. The Pile: GitHub repository, 2025)	800 ГБ дереккөзі, (кітаптар, код, ғылыми мақалалар).	2020	ағылшын

PubMed (PubMed, 2025)	биомедициналық әдебиеттеріне, жаратылыстану журналдарына және онлайн кітаптарға 37 миллионнан астам сілтемелерді қамтиды.	1997	ағылшын
Open Corpora (OpenCorpora, 2025)	Орыс тіліндегі мәтіндердің үлкен коллекциясы. Әдеби шығармалар, ресми құжаттар, энциклопедиялық мәтіндер.	2009	орыс
Yandex Toloka Datasets (Toloka.ai, 2025)	Toloka платформасы арқылы жиналған және өңделген орыс тіліндегі деректер. Қамтитын мазмұн: Жаңалықтар, техникалық мәтіндер, пікірлер.	2014	орыс
Қазақ тілінің ұлттық корпусы (Qazcorpus, 2025)	қазақ тілінің лексика-грамматикалық жүйесін толық қамтыған (терең аннотацияланған) миллиондаған сөзқолданыстан тұратын электронды пішіндегі көлемді мәтіндер жинағы, Жалпы сөзқолданыс саны – 65 000 000. 16 ішкорпустан тұрады.	2012	қазақ
KOHTD	3000 қолжазба емтихан жұмысы мен 140335-тен астам сегменттелген суреттері бар және шамамен 922010 таңбадан тұратын қазақ тіліндегі офлайн қолжазба мәтіндік деректер жинағы (Kazakh offline Handwritten Text dataset - KOHTD) (Toiganbayeva et al, 2022)	2022	қазақ
Kaz_txt_Dataset	Әр түрлі сала бойынша қазақ тіліндегі кең датасет	2024	қазақ

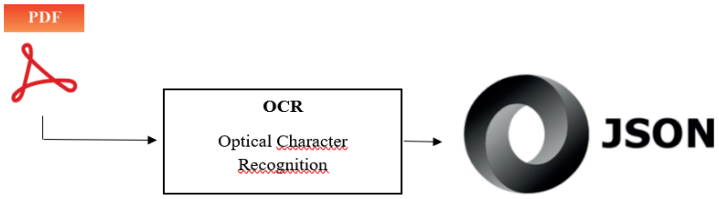
Қазақ тілі үшін сапалы дереккөздердің жетіспеуі LLM өнімділігін шектейді. Қазақ тілі үшін сапалы dataset құру – үлкен тілдік модельдерді (LLM) оқыту мен дамытудың маңызды аспектілерінің бірі. Бұл өзектілікті бірнеше негізгі факторлармен түсіндіруге болады:

- Тілдік ресурстар тапшылығы – қазақ тілінде үлкен әрі әртүрлі деректер жиынтығы жоқтың қасы.
- LLM қазақ тілінде нашар жұмыс істейді – Қолданыстағы GPT, LLaMA, Mistral модельдері қазақ тіліндегі мәтіндерді өңдеуде жиі қателеседі.
- Ақпараттық теңсіздік – Қазақ тілді қолданушылар үшін жасанды интеллект мүмкіндіктері шектеулі болып отыр.
- Сапалы dataset қазақ тілінде дамыған GPT сияқты модельдерді жасауға мүмкіндік береді.
- Мемлекеттік қызмет, білім беру, бизнес, медицина және тағы басқа салаларда жасанды интеллект құралдарын тиімді қолдануға жол ашады.
- Егер қазақ тілі үшін сапалы деректер жинақталмаса, ол үлкен тілдік модельдердің көлеңкесінде қалып қоюы мүмкін.
- Өзіміздің dataset-іміз болмаса, AI қазақ тілін қате немесе бұрмаланған түрде түсіндіруі мүмкін.

Міне, осы өзектілі мәселелер «Қазақ тілі мен технологиялық прогресті қолдау үшін үлкен тіл моделін (LLM) құру» жобасы аясында шешімін тауып,

Kaz_txt_Dataset құрылып жатыр. Kaz_txt_Dataset құру үшін pdf форматтағы материалдар OCR технологиясына негізделіп, txt, JSON форматтарына аударылып (1-сурет.), жинақталған мәтіндер әр сала бойынша іріктелген.

1-сурет. pdf форматтағы материалдар OCR технологиясына негізделіп JSON форматтарына аудару



OCR технологиялары dataset-ті жаңа деректер көздерімен толықтыру арқылы баспа және қолжазба мәтіндерін цифрландыруға көмектеседі. OCR жүйелері латын, кириллица және араб әліпбиіндегі мәтіндерді тану және конвертациялау мүмкіндігін ескеруі керек. Бұған қоса, тарихи және сирек кездесетін құжаттардан алынған мәтіндерді тану, түзету және валидациялау әдістерін жетілдіру маңызды. OCR-дің қазақ тіліне бейімделуі үшін арнайы нейрондық желілерді оқыту қажет, бұл кириллица мен латын графикасын қатар тануға мүмкіндік береді.

Dataset болашақта өзін-өзі оқытатын модельдерге арналған нейрондық желілермен интеграциялануы мүмкін, бұл деректердің сапасын автоматты түрде жақсартуға мүмкіндік береді. Сонымен қатар, әртүрлі салаларға арналған мамандандырылған модельдерді дамыту (құқықтану, медицина, білім беру) маңызды бағыттардың бірі болып табылады. Бұдан бөлек, мультимодальды модельдер құру — яғни мәтін, аудио және бейнені біріктіру арқылы кеңейтілген LLM жасау да өзекті. Мысалы, сөйлеу тілін тану және генерациялау мүмкіндіктерін жетілдіру қазақ тіліндегі қолданушыларға ыңғайлырақ жүйелер жасауға мүмкіндік береді. Қазақ тілінде LLM дамытудағы маңызды қадамдардың бірі — диалектілік ерекшеліктерді есепке алатын және морфологиялық құрылымға сезімтал модельдер жасау, бұл өз кезегінде қазақ тілінің цифрлық трансформациясына айтарлықтай үлес қосады.

Қазақ тілінде LLM құруға байланысты бірнеше зерттеу жұмыстары бар. MBERT және GPT-4 сияқты көп тілді модельдер қазақ тілімен жұмыс істеу қабілетін көрсетсе де, олар негізінен ағылшын және басқа да жоғары ресурстық тілдерге бейімделген. Қазақ тілінің лингвистикалық ерекшеліктерін неғұрлым терең және дәл өңдеу үшін мынадай тәсілдер әзірленді:

- KazNERD (Yeshpanov et al., 2022) - атаулар, ұйымдар және географиялық атаулар сияқты субъектілердің қолмен аннотациясын қамтитын қазақ тілінде аталған субъектілерді тану міндеті үшін құрылған датасет.
- KazGPT-жергілікті бастамалар шеңберінде әзірленген қазақ мәтінін генерациялауға бағытталған прототиптік тілдік модель.

- Қолмен белгіленген корпустарға негізделген модельдер — морфологиялық және синтаксистік талдау үшін, соның ішінде semeval және Universal Dependencies халықаралық жобалары аясында қолданылды.

Біз ұсынатын әдіс мамандандырылған қазақ тілді деректер жиынтығын және бейімделген трансформаторлық архитектураны қолдану есебінен мәтінді генерациялау және түсіну сапасын арттыру мақсатында аталған модельдердің артықшылықтарын біріктіруді көздейді. Қазақ тіліне арналған LLM әзірлеу үшін бірнеше эксперимент жүргізілді.

Эксперимент барысында әртүрлі дереккөздерден алынған мәтіндер өңделді. Негізгі қадамдар:

- Мәтіндерді лемматизациялау және токенизациялау – қазақ тілінің морфологиялық ерекшеліктеріне сәйкес арнайы өңдеу әдістері қолданылды.
- Диалектілерді сәйкестендіру – әртүрлі стильдер мен аймақтық ерекшеліктерді ескеру үшін корпусқа әртүрлі көздерден алынған мәтіндер қосылды.
- Деректерді аннотациялау – POS-тегтеу, синтаксистік құрылымдарды белгілеу және семантикалық аннотациялар енгізілді.

Модельді баптау (Fine-tuning)

Зерттеуде mBERT және GPT-3 модельдері қазақ тіліндегі деректер жиынтығында нақтылап бапталды. Баптау кезеңдері:

1. Алдын ала оқытылған модельді пайдалану – mBERT және BLOOM көптілді модельдері қазақ тіліндегі мәтіндерге бейімделді.
2. Қазақ тіліндегі корпус негізінде қайта оқыту – арнайы жинақталған dataset қолданылды.
3. Тілдік ерекшеліктерге бейімдеу – аффиксация, септеу және морфологиялық өзгешеліктер ескерілді.

Зерттеу нәтижелері мен талқылау.

Модельдердің тиімділігі бірнеше метрика бойынша бағаланды:

- Perplexity (PPL) – тілдік модельдің қаншалықты сәйкес келетінін өлшеу.
- BLEU score – аударма және генерацияланған мәтін сапасын бағалау.
- F1-score – атау есімдерді тану (NER) және морфологиялық талдау дәлдігін өлшеу.

Зерттеу барысында қазақ тіліне арналған LLM әртүрлі үлгілері талданды. Пысықтау нәтижелері төмендегі №3 кестеде келтірілген.

№3 кесте - Нәтижелер

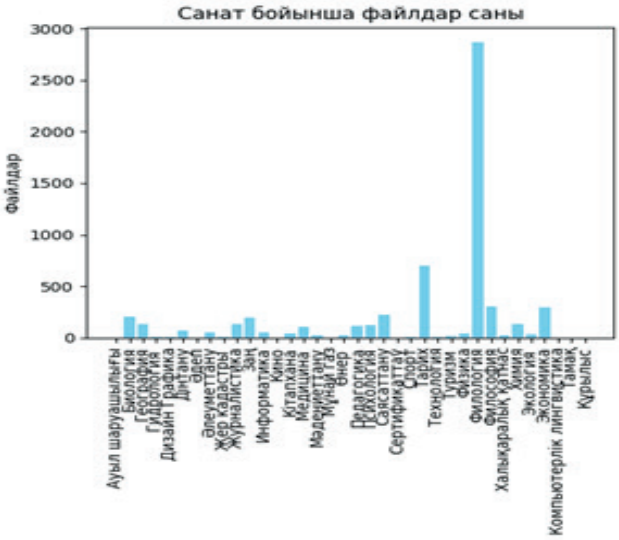
Модель	Perplexity ↓	BLEU ↑	F1-score ↑
mBERT	25.3	32.7	78.4
GPT-3	19.8	45.2	85.1
BLOOM	22.1	38.5	81.3
LLaMA	18.5	47.1	86.4
Mistral	17.3	50.5	88.2

Қазақ тіліне арналған GPT-3 және LLaMA модельдерін дәл баптау процесін қолдана отырып жетілдіру кезінде бірнеше тест мәтіндері қолданылды. Мысалы, bert моделі бастапқыда "Бүгінгі ауа-райы қандай? деген сұраққа "Мен бүгін ауа райы деректерін көрсете алмаймын" деп жауап берді. Ал, Mistral моделі "Бүгін Алматыда күн ашық, ауа температурасы +10°c" дәлірек жауап берді. Бұл қазақ тіліндегі мәтіндердің мәнмәтінін түсінудегі соңғы үлгілердің артықшылықтарын көрсетеді.

Корпус деректерін визуалды талдау. Қазақ тіліне арналған үлкен тілдік корпусты әзірлеу кезінде деректердің жалпы көлемі бойынша және оларды тақырыптық санаттар бойынша бөлу қажет. Дереккөздердің әртүрлілігі және тақырыптар бойынша деректердің тепе-теңдігі қорытынды тіл моделінің сапасына тікелей әсер етеді. Бұл жұмыста жиналған корпусқа талдау жасалды және оны алынған мәліметтер негізінде датасеттің құрылымын бейнелейтін үш диаграмма арқылы көрсетеміз.

2- суртеттегі гистограммада әртүрлі тақырыптық санаттар бойынша файлдар санын көрсетміз. Бірқатар санаттарда мәтіндердің едәуір саны бар екендігі байқалады, ал басқаларында деректер айтарлықтай аз. Мысалы, филология, тарих, экономика санаттары ең жоғары көрсеткіштерге ие, бұл осы салалардағы мәтіндердің жоғары қолжетімділігін көрсетеді. Сонымен қатар, компьютерлік лингвистика, құрылыс, кино санаттарында әлдеқайда аз деректер бар, бұл осы салалардағы мәтіндерді өңдеу кезінде модель сапасының жеткіліксіздігіне әкелуі мүмкін.

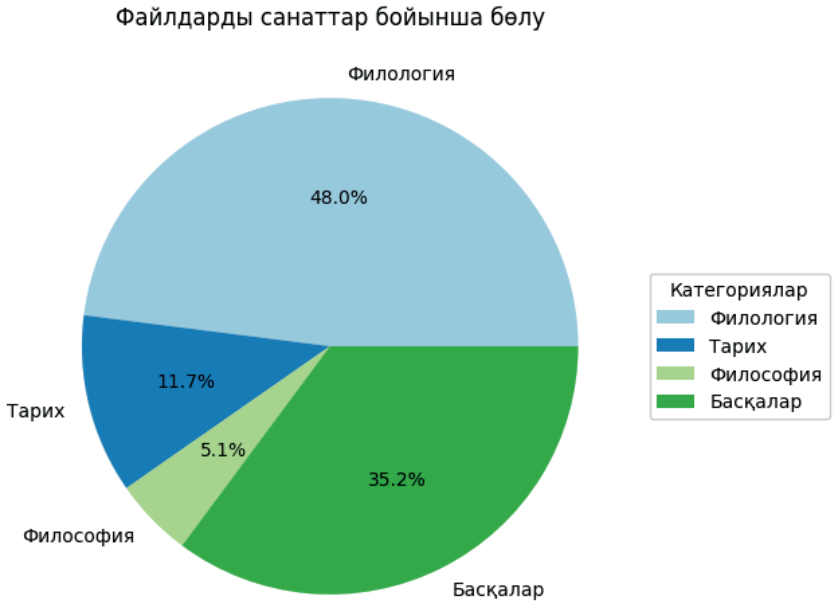
2-сурет. Деректерді тарату гистограммасы



Бұл талдау корпусты теңдестіру үшін бірқатар бағыттар бойынша қосымша мәліметтер жинау қажеттілігі туралы қорытынды жасауға мүмкіндік береді.

3-суреттегі дөңгелек диаграмма тақырыптық санаттар бойынша файлдардың пайыздық таралуын көрсетеді. Визуализацияның бұл түрі корпуста қандай тақырыптар басым екенін және деректердің қаншалықты біркелкі бөлінгенін бағалауға мүмкіндік береді.

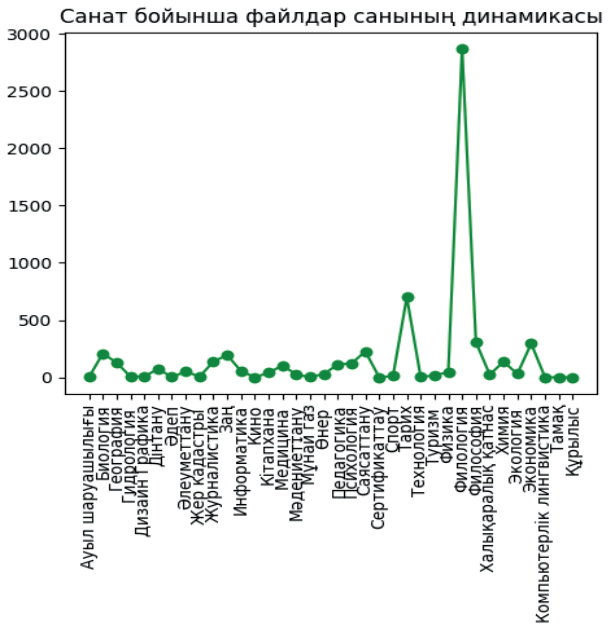
3-сурет. Корпус құрылымының дөңгелек диаграммасы



Диаграммада көрсетілгендей, филология, тарих, экономика корпустың көп бөлігін алады, ал қалған санаттар аз ғана пайызды құрайды. Мұндай теңгерімсіздік тілдік модельдің сапасына әсер етуі мүмкін, өйткені тек тақырыптар басым болатын корпуста оқытылған Машиналық оқыту алгоритмдері аз ұсынылған салалардағы мәтіндерді өңдеу кезінде нашар өнімділікті көрсетуі мүмкін. Оңтайлы шешім қосымша деректерді мақсатты түрде жинау, сондай-ақ санаттар бойынша файлдар санын қалыпқа келтіру арқылы корпусты теңестіру болуы мүмкін.

4-суретте сызықтық график санаттар бойынша файлдар санының өзгеруін көрсетеді. Бұл корпус құрылымындағы трендтерді анықтауға және әртүрлі бағыттар бойынша деректерді жинау қарқынын нақты салыстыруға мүмкіндік береді. Деректер көлемі біркелкі бөлінбегенін көруге болады, бұл корпусты теңестіру бойынша қосымша жұмыс істеу қажеттілігін тағы бір рет растайды.

4-сурет. Деректер динамикасының сызықтық графигі



Корпустың визуалды деректерін талдау бірнеше негізгі қорытынды жасауға мүмкіндік береді:

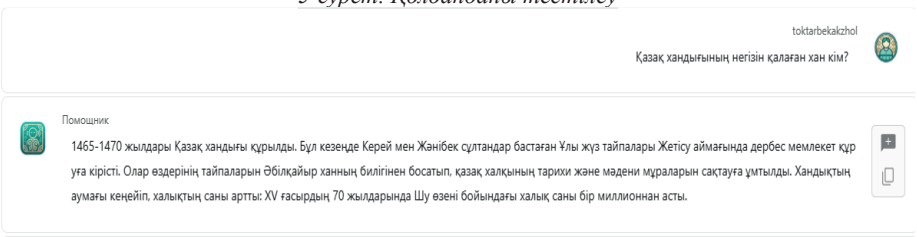
1. Деректер теңгерімсіздігі- белгілі бір санаттарда басқаларға қарағанда едәуір көп файлдар бар, бұл модельдің басым тақырыптарға ауысуына әкелуі мүмкін.

2. Корпусты кеңейту қажеттілігі — ең аз мәтіндері бар Санаттар датасеттің жалпы өкілдігін жақсарту үшін қосымша толтыруды қажет етеді.

3. Деректер құрылымын оңтайландыру- қайта ұсынылған санаттардағы қайталанатын деректерді сүзу немесе жетіспейтін тақырыптардағы деректерді жасанды түрде көбейту сияқты теңдестіру әдістері мүмкін.

5-суретте модель құрылған датасет негізінде сұраққа жауап берді:

5-сурет. Қолданбаны тестілеу



«Қазақ хандығының негізін қалаған хан кім?» сұрағына AI Assistant берген жауабы өте мазмұнды әрі құрылымды болды.

Бұл нәтижелер үлкен тілдік корпусты құруда маңызды рөл атқарады және оны одан әрі жақсартудың негізгі бағыттарын анықтауға көмектеседі. Санаттар арасында деректерді біркелкі бөлуге қол жеткізу әртүрлі тақырыптық салалардағы мәтіндерді тиімді өңдеуге қабілетті жоғары сапалы және әмбебап тілдік модельді қамтамасыз етеді.

Қорытынды. Қазақ тілі үшін үлкен тіл моделін (LLM) құру осы тілдегі мәтіндермен тиімді жұмыс істеуге қабілетті жасанды интеллектті дамытудағы маңызды қадам болып табылады. Қазақ тілінің морфологиялық, синтаксистік және диалектілік ерекшеліктерін зерделеу, сондай-ақ әртүрлі алфавиттерді есепке алу табысты модельді әзірлеу кезінде шешуші болып табылады. Бұл жолдағы басты проблемалардың бірі Қазақ тілі үшін тілдік деректердің шектелуі болып табылады, бұл LLM-ді оқытуды қиындатады және олардың өнімділігін төмендетеді.

Осы жоба аясында dataset жиналды, оның мақсаты — қазақ тілді модельді оқыту үшін деректерді жинау және құрылымдау. OCR сияқты заманауи технологияларды пайдалану әртүрлі көздерден материалдарды цифрландыруға және өңдеуге оңай пішімдерге түрлендіруге мүмкіндік береді, бұл қол жетімді деректер көлемін айтарлықтай кеңейтеді. AI Assistant сұрақтарға орынды және толық жауап беру үшін KazLLM-ге арналған датасетті әлі де толықтыру жұмыстарын жүргізу керек.

Сонымен қатар, қазақ тілі үшін сапалы және алуан түрлі dataset құру жасанды интеллектті дамыту, оны мемлекеттік басқару, білім беру, медицина және бизнес сияқты салаларда қолдану үшін мүмкіндіктер ашады. Толыққанды және аннотацияланған деректердің болуы қазақ тілімен адам түсінігіне жақын деңгейде жұмыс істей алатын дәл және тиімді модельдерді әзірлеуге мүмкіндік береді.

Осылайша, осы жоба шеңберінде атқарылған жұмыс ақпараттық теңсіздікті жоюға ықпал етеді және оның жасанды интеллекттің жаһандық әзірлемелеріне толыққанды қатысуын қамтамасыз ете отырып, қазақ тілін технологиялық прогресте қолдауды қамтамасыз етеді.

References

Bender E.M. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. P. 610–623. Available at: https://dl.acm.org/doi/10.1145/3442188.3445922 (in Eng.)

Brown T. et al. Language Models are Few-Shot Learners [Electronic resource]. arXiv:2005.14165. Available at: https://arxiv.org/abs/2005.14165 (in Eng.)

Common Crawl [Electronic resource]. — Access mode: https://commoncrawl.org (date accessed: 08.04.2025) (in Eng.)

Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resource]. — arXiv:1810.04805. – Access mode: https://arxiv.org/abs/1810.04805 (in Eng.)

EleutherAI [Electronic resource]. — Access mode: https://www.eleuther.ai (date accessed: 08.04.2025) (in Eng.)

Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification. Proceedings

of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. — P. 328–339. — Access mode: https://aclanthology.org/P18-1031 (in Eng.)

Nikbakht S.R., Benzaghta M., Geraci G. TSpec-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications [Electronic resource]. – arXiv:2406.01768. — Access mode: https://arxiv.org/abs/2406.01768 (in Eng.)

OpenAI. GPT-4 Technical Report [Electronic resource]. – Access mode: https://openai.com/research/gpt-4 (in Eng.)

OpenCorpora [Electronic resource]. — Access mode: https://opencorpora.org (date of access: 08.04.2025) (in Eng.)

PubMed [Electronic resource]. – Access mode: https://pubmed.ncbi.nlm.nih.gov (date accessed: 08.04.2025) (in Eng.)

QazCorpus – Kazakh Language Corpus, version 2 [Electronic resource]. – Access mode: https://v2.qazcorpus.kz (date accessed: 08.04.2025) (in Eng.)

The Pile – GitHub repository [Electronic resource]. — Access mode: https://github.com/EleutherAI/the-pile (date accessed: 08.04.2025) (in Eng.)

Toloka.ai [Electronic resource]. — Access mode: https://toloka.ai (access date: 04/08/2025) (in Eng.)

Toiganbayeva N, Kasem M., Abdimanap G, Bostanbekov K., Abdelrahman A., Alimova A., Nurseitov D. Toiganbayeva, N. A. et al. (2022) “KOHTD: Kazakh Offline Handwritten Text Dataset.”. Signal Processing: Image Communication. — Volume 108. https://www.sciencedirect.com/science/article/pii/S0923596522001217 (in Eng.)

Yeshpanov R., Khassanov Y., Varol H. A. KazNERD: Kazakh Named Entity Recognition Dataset. Proceedings of the Language Resources and Evaluation Conference (LREC 2022). — P. 406–414. — Access mode: https://aclanthology.org/2022.lrec-1.44/ (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 93–109

https://doi.org/10.32014/2025.2518-1726.366

IRSTI 20.51.23
UDC 004.891

© A. Bekarystankyzy*, M. Baizakova, A. Kassenkhan, M. Iglikova*, 2025.
Narxoz University, Almaty, Kazakhstan.
E-mail: akbayan.b@gmail.com

RECOMMENDATION ALGORITHMS FOR EDUCATIONAL PREFERENCES: A REVIEW

Bekarystankyzy Akbayan — PHD, Associate Professor, School of Digital Technologies, Narxoz University, Almaty, Kazakhstan, E-mail: akbayan.b@gmail.com; ORCID ID: http://orcid.org/0000-0003-3984-2718;
Baizakova Madina — Bachelor's student, School of Digital Technologies, Narxoz University, Almaty, Kazakhstan,
E-mail: nevloba.mmm@gmail.com; ORCID ID: http://orcid.org/0009-0000-6027-1011;
Kassenkhan Aray — Doctor PHD, Associate Professor, Department of Software Engineering, Satbayev University, Almaty, Kazakhstan,
E-mail: a.kassenkhan@satbayev.university; ORCID ID: http://orcid.org/0000-0002-6355-9544;
Iglikova Mereilim — Lecturer, Department of Software Engineering, Satbayev University, Almaty, Kazakhstan,
E-mail: m.iglikova@satbayev.university; ORCID ID: http://orcid.org/0000-0001-8389-5820.

Abstract. Educational Recommender Systems (ERS) have become a critical component of modern digital learning environments, offering personalized learning pathways tailored to individual student needs, preferences, and behaviors. These systems utilize a wide range of data sources—including demographic information, academic performance, cognitive characteristics, and behavioral patterns—to recommend relevant courses, learning materials, and strategies that support student engagement and success. This paper presents a comprehensive review of current approaches used in the development of ERS, including collaborative filtering, content-based filtering, hybrid recommendation models, machine learning, deep learning algorithms, knowledge graphs, and explainable artificial intelligence (XAI). Particular attention is given to the advantages and limitations of each approach, as well as key challenges that persist in the implementation of ERS. These include the cold-start problem, data sparsity, scalability issues, lack of transparency in decision-making, and concerns regarding user privacy and algorithmic fairness. The paper also explores practical applications of ERS in higher education and large-scale online platforms such as MOOCs, where such systems have demonstrated positive impacts on learner motivation, retention,

and academic performance. Furthermore, the ethical dimensions of educational recommendations—such as inclusivity, bias mitigation, and transparent design—are discussed as essential components of trustworthy and student-centered systems. The review emphasizes the growing importance of adaptable, interpretable, and ethically aligned recommendation models that can evolve with the dynamic nature of modern education. Finally, the study identifies key directions for future research, including the development of reinforcement learning-based systems, improved handling of large and heterogeneous data sets, and the integration of personalized recommendation mechanisms that enhance the overall quality, transparency, and fairness of digital education worldwide.

Key words: Personalized learning, Educational Recommendation Systems (ERS), Collaborative Filtering, Hybrid Models, Large-Scale Datasets, Explainable Artificial Intelligence

Information about funding

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993072)

© А. Бекарыстанқызы [*], М. Байзакова, А. Қасенхан, М. Игликова[*], 2025.

«Нархоз университеті», Алматы, Қазақстан.
E-mail: akbayan.b@gmail.com

БІЛІМ АЛУДЫ ЖАҚСАРТУ ҮШІН ҰСЫНЫС БЕРЕТІН АЛГОРИТМДЕРГЕ ШОЛУ

Бекарыстанқызы Ақбаян — PhD, қауымдастырылған профессор, Цифрлық технологиялар мектебі, Нархоз университеті, Алматы, Қазақстан,
E-mail: akbayan.b@gmail.com; ORCID ID: http://orcid.org/0000-0003-3984-2718;
Байзақова Мадина — бакалавр студенті, Цифрлық технологиялар мектебі, Нархоз университеті, Алматы, Қазақстан,
E-mail: nevloba.mmm@gmail.com; ORCID ID: http://orcid.org/0009-0000-6027-1011;
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; ORCID ID: http://orcid.org/0000-0002-6355-9544;
Игликова Мерейлім — оқытушы, Программалық инженерия кафедрасы, Қ.И. Сәтбаев атындағы Қазақ Ұлттық Техникалық Зерттеу университеті, Алматы, Қазақстан,
E-mail:m.iglikova@satbayev.university; ORCID ID: http://orcid.org/0000-0001-8389-5820.

Аннотация: Білім беру бойынша ұсыныс жүйелері (Educational Recommender Systems, ERS) соңғы жылдары цифрлық оқытудың ажырамас құрамдас бөлігіне айналып, жекелендірілген оқыту әдістерін іске асыруға және білім беру сапасын арттыруға айтарлықтай үлес қосып келеді. Мұндай жүйелер студенттердің академиялық үлгерімі, оқу барысындағы мінез-құлық ерекшеліктері, жеке қалаулары, когнитивтік стильдері мен

демографиялық мәліметтері сияқты әртүрлі деректерді сараптай отырып, оқыту траекториясын жеке бейімдеуге мүмкіндік береді. Бұл мақалада ERS құруда қолданылатын заманауи әдістерге жан-жақты шолу жасалады. Атап айтқанда, коллаборативті сұрыптау, мазмұнға негізделген сүзгілеу, гибридті модельдер, машиналық және терең оқыту алгоритмдері, білім графтары мен түсіндірмелі жасанды интеллект (XAI) технологиялары қарастырылады. Сонымен қатар, деректердің сиректігі, суық бастау мәселесі, жүйені кеңейту кезіндегі қиындықтар, алгоритмдердің ашықтығы мен пайдаланушылардың дербес деректерін қорғау қажеттілігі сияқты өзекті мәселелер талданады. ERS-тің ашық онлайн-оқыту платформаларында (MOOC) және жоғары оқу орындарында қолданылу мысалдары келтіріліп, олардың студенттердің мотивациясын арттыруға, оқу процесіне тартылуын күшейтуге және оқудан шығу деңгейін төмендетуге ықпал ететіні көрсетіледі. Сонымен қатар, ұсыныстар әділдігі, инклюзивтілік және этикалық тұрақтылық мәселелері де қарастырылады. Авторлар білім беру саласындағы өзгермелі талаптарға бейімделе алатын, икемді, түсінікті және әділ ұсыныс жүйелерін әзірлеудің маңыздылығын атап өтеді. Бұл шолу зерттеу білім беру жүйелерін одан әрі дамытуға негіз бола отырып, ұсыныс жүйелерінің тиімділігін, этикалық жауапкершілігін және жекелендіру деңгейін арттыру жолдарын ұсынады.

Түйін сөздер: Білім беру бойынша ұсыныс жүйелері (БҰЖ), жеке оқыту, коллаборативті сұрыптау, гибридті модельдер, түсіндірмелі жасанды интеллект (XAI), кең көлемді деректер жиынтығы, ұсыныс дәлдігі

А. Бекарыстанқызы*, М. Байзакова, А. Қасенхан, М. Игликова*, 2025.
«Университет Нархоз», Алматы, Казахстан.
E-mail: akbayan.b@gmail.com

ОБЗОР РЕКОМЕНДАТЕЛЬНЫХ АЛГОРИТМОВ ДЛЯ ОБРАЗОВАТЕЛЬНЫХ ПРЕДПОЧТЕНИЙ

Бекарыстанкызы Акбаян — PhD, доцент, Школа цифровых технологий, Университет Нархоз, Алматы, Казахстан,
E-mail: akbayan.b@gmail.com, ORCID ID: http://orcid.org/0000-0003-3984-2718;
Байзакова Мадина — студентка бакалавриата, Школа цифровых технологий, Университет Нархоз, Алматы, Казахстан,
E-mail: nevloba.mmm@gmail.com, ORCID ID: http://orcid.org/0009-0000-6027-1011;
Касенхан Арай — PhD, доцент, кафедра Программной инженерии, Казахский Национальный Исследовательский Технический университет К.И.Сатпаева, Алматы, Казахстан,
E-mail: a.kassenkhan@satbayev.university, ORCID ID: http://orcid.org/0000-0002-6355-9544;
Игликова Мерейлім — преподаватель, кафедра Программной инженерии, Казахский Национальный Исследовательский Технический университет К.И.Сатпаева, Алматы, Казахстан,
E-mail: m.iglikova@satbayev.university, ORCID ID: http://orcid.org/0000-0001-8389-5820.

Аннотация: Образовательные рекомендательные системы (Educational Recommender Systems, ERS) в последние годы становятся неотъемлемой

частью цифровых образовательных платформ, так как позволяют реализовать персонализированный подход к обучению и повысить его эффективность. Эти системы используют современные алгоритмы анализа данных и искусственного интеллекта для подбора учебных материалов, курсов и индивидуальных траекторий обучения, опираясь на широкий спектр информации: академическую успеваемость, поведенческие особенности, предпочтения, когнитивные стили, а также демографические данные обучающихся. Настоящая статья представляет собой всесторонний обзор современных подходов к построению ERS, включая коллаборативную и контентную фильтрацию, гибридные модели, алгоритмы машинного и глубокого обучения, применение графов знаний и объяснимого искусственного интеллекта (XAI). Рассматриваются ключевые проблемы, с которыми сталкиваются такие системы: эффект «холодного старта», разреженность данных, сложности масштабирования, необходимость обеспечения прозрачности алгоритмов и защиты персональных данных. Также анализируются практические примеры применения ERS в контексте онлайн-обучения и высшего образования (в том числе на платформах MOOCs), где системы рекомендаций демонстрируют высокую результативность, способствуя снижению отсева студентов, повышению их мотивации и вовлечённости в учебный процесс. Отдельное внимание уделено вопросам этики, инклюзии, а также справедливости в рекомендациях. В статье подчёркивается необходимость развития устойчивых, адаптивных и интерпретируемых моделей, способных учитывать изменения в образовательных потребностях и контексте. Представленное исследование не только систематизирует существующие подходы, но и обозначает направления дальнейших разработок, направленных на повышение качества, прозрачности и персонализации цифрового образования на глобальном уровне.

Ключевые слова: образовательные рекомендательные системы (ERS), персонализированные обучение, коллаборативная фильтрация, гибридные модели, объяснимый ИИ, крупномасштабные наборы данных, точность рекомендации

Introduction. Educational systems with recommending functionalities are very popular and their application and usage are getting wider according to their capability to proffer materials, which meets individual preferences of learners. By performing analyze on data like materials of past courses, contents of searches by learners ERS generate suggestions using filtering types, like collaborative, content-based and hybrid. However, challenges such as the cold-start problem and data sparsity persist, particularly for new users or courses (Masthoff, 2023; Thammarak, Kesornsit, & Sirisathitkul, 2024; Zhou & Zhang, 2024).

Hybrid systems that combine CF and CBF have demonstrated improvements in addressing these limitations, offering more accurate and robust recommendations (Shuang & Hang, 2024; Alahmadi & Alruwaili, 2021). In addition, deep learning and matrix factorization methods like Singular Value Decomposition (SVD)

have significantly enhanced the performance of ERS, especially in large-scale environments like MOOCs (Wu, 2023; Alfaifi, 2024).

The inclusion of knowledge graphs and explainable AI (XAI) offers further advancements by providing more personalized, context-aware recommendations and ensuring transparency in the recommendation process (Shuang et al., 2024; Zheng, 2022; Yazdi et al., 2024). While progress has been made, challenges related to privacy, data management, and recommendation quality remain, requiring ongoing research to further enhance the effectiveness of ERS (Dhananjaya, Goudar, Kulkarni, & Rathod, 2024; Xiong, Li, Liu, Chen, Zhou, Rong, & Ouyang, 2024).

Materials and Methods. Implication of recommender systems to educational process requires significant efforts due to the fact that it has enormous aspects. The next research questions were studied in the current paper in order to obtain majority of these aspects (Table 1):

Table 1. Research questions

Research questions
Background and Key Concepts
Main Approaches to Educational Recommendations
Challenges in Educational Recommender Systems
Case Studies and Applications
Evaluation of Educational Recommender Systems
Future Directions and Research Gaps

Approaches to Material Search. Relevant material was sourced using a methodical strategy to guarantee a thorough review. The following techniques were applied:

1. Database Selection: Google Scholar, Elsevier's ScienceDirect, IEEE Xplore, ACM Digital Library, and Springer were used as major academic databases. For the most recent scientific findings, preprint sources such as arXiv were also investigated.

2. Method of Search: To find pertinent material, a mix of Boolean operators and keywords was employed. Among the sample search terms were: "Educational recommender systems" AND "personalized learning"

● "Collaborative filtering in education" OR "content-based recommendation for students"

● "Challenges in adaptive learning platforms"

● "Hybrid recommendation models in MOOCs"

Inclusion and Exclusion Criteria:

● **Inclusion Criteria:** Peer-reviewed journal articles, conference papers, and authoritative surveys published in the last decade (2020-2025) focusing on recommender systems for educational settings.

● **Exclusion Criteria:** Papers that do not explicitly discuss recommendation algorithms in education, duplicate studies, and works that lack empirical validation.

Selection Process: Titles and abstracts were screened to determine relevance. Full-text articles were reviewed if they contained significant discussions on educational recommender systems.

Classification of Sources: The collected materials were categorized into different themes aligned with the research questions in Table 1. This ensured structured analysis and synthesis of information.

Methodology. To answer the research questions, an integrative literature review methodology was employed. The process involved:

- **Thematic Analysis**: Identifying and grouping themes such as user modeling, algorithmic approaches, and evaluation techniques.
- **Analyzing by comparing**: compare different techniques of recommendations in order to detect their strength, weakness and possible contexts of application.
- **Case Study Review**: Examining real-world implementations of educational recommender systems to assess their impact.

By employing these methodologies, this study provides a structured and comprehensive analysis of recommendation algorithms in educational settings, ensuring coverage of both theoretical and applied aspects of the field.

2.1 Background and Key Concepts

2.1.1 Overview of Recommender Systems. Recommender systems (RS) are software tools and techniques designed to assist users in making decisions by providing personalized suggestions. The types of systems analyze behaviors of users according to their previous activities in a system in order to give suggestions and recommendations about products, services, information resources and courses. Typically, RS employ methods like collaborative filtering, content-based filtering, and hybrid approaches, each with its own strengths and limitations (Aucancela, 2023; Li, Zhang, & Zhang, 2021; Kamal et al., 2024; Urdaneta-Ponte et al., 2021).

The focus Collaborative Filtering (CF) is directed to the interaction of users with items. This approach generalized the preferences of users with similar behaviors for proposing recommendations. The main problem in the usage of this method can the sparsity of data and cold-start issue. On the other hand, Contend-Based Filtering (CBF) pays attention to the attributes of items and makes suggestions on the basis of user's interaction with cognate items in the past. While CBF can handle new items effectively, it may struggle with offering diverse suggestions (Zhou et al., 2024; Alahmadi et al., 2021; Xiong et.al., 2024). Hybrid systems combine CF and CBF techniques to address their respective shortcomings, offering improved accuracy and robustness (Thammarak et al., 2024; Alahmadi et al., 2021; Yazdi et al., 2024).

The wide use of recommendation systems can be found in domains like tourism, healthcare, e-commerce and education. Their ability to enhance user experience, decision-making, and engagement has positioned them as crucial tools across industries (Urdaneta-Ponte et al., 2021; Xiong et al., 2024).

2.1.2 Distinction Between General Recommender Systems and Educational Recommender Systems. Educational recommender systems (ERS) are a

specialized subset of RS designed to support educational decision-making by providing personalized recommendations for courses, learning materials, career paths, and other academic resources. Unlike general RS, which often prioritize maximizing user satisfaction or sales, ERS focus on improving educational outcomes, engagement, and retention rates (Aucancela, 2023; Butmeh & Abu-Issa, 2024; Dhananjaya et al., 2024).

In order to create suitable recommendations to learners, ERS usually consume unique types of data like demographic information, learning behaviors and academic rates. These systems pursue the aim to solve specific issues in education like helping students guiding suitable course, improve the development of their various skills and the rates of possible dropouts. Techniques such as collaborative filtering, content-based filtering, and hybrid models are commonly employed in ERS, with an emphasis on integrating domain-specific knowledge and learner preferences (Butmeh et al., 2024; Shuang et al., 2024).

One key distinction is the importance of pedagogical considerations in ERS. These types of systems basically use educational theories in their core logic to meet learners expectations and enhance their capabilities, the examples example these theories can be cognitive diagnosis and knowledge mapping (Shuang et al., 2024; Butmeh et al., 2024; Silva et al., 2022). Furthermore, Educational Recommender Systems (ERS) are progressively incorporating sophisticated methodologies, such as deep learning and knowledge graphs, to improve recommendation precision and better accommodate the dynamic needs of learners (Wu, 2023; Zheng, 2022; Guo et al., 2024).

2.1.3 Common Data Types Used in Educational Recommender Systems. ERS rely on diverse data types to deliver personalized and effective recommendations. Commonly used data types include:

1. Demographic Data: Information such as age, gender, and location helps contextualize recommendations, ensuring relevance to the learner's background and circumstances (Butmeh, et al., 2024; Dhananjaya, et al., 2024; Yazdi, et al., 2024).
2. Academic Data: This includes grades, test scores, and course enrollments, providing insights into the learner's academic performance and progress (Phalle & Bhushan, 2024; Dhananjaya, et al., 2024; Xiong, et al., 2024).
3. Behavioral Data: Data on learners' interactions with educational platforms, such as time spent on tasks, resource usage patterns, and engagement levels, helps model preferences and predict future needs (Ma et al., 2023; Phalle, et al., 2024; Xiong, et al., 2024).
4. Cognitive and Learning Preferences insights into learners' knowledge levels, cognitive styles, and favored instructional strategies facilitate the generation of personalized recommendations tailored to their unique abilities and learning objectives (Thammarak, et al., 2024; Rahman et al., 2022; Guo, et al., 2024).

By synthesizing these data dimensions, Educational Recommender Systems (ERS) can construct detailed learner profiles, thereby supporting the creation of

recommendations that not only improve learning outcomes but also effectively mitigate issues like the cold-start problem and information overload (Thammarak, et al., 2024; Alahmadi et al., 2021; Raza et al., 2024).

2.2 Main Approaches to Educational Recommendations

2.2.1 Collaborative Filtering (CF). Collaborative Filtering (CF) is a prominent methodology in recommendation systems, particularly in educational recommender systems (ERS). It leverages the preferences and behaviors of users to predict items or services that may interest others with similar patterns. CF operates primarily through two approaches: **user-based filtering** and **item-based filtering**.

- **User-based filtering** identifies users with similar preferences or behaviors. Recommendations are made based on what similar users have liked or interacted with.
- **Item-based filtering** analyzes relationships between items, recommending items that are closely related to those a user has previously interacted with.

CF can effectively propose personalized recommendations especially if there is available a big amount of user-interaction data. For instance, it is frequently employed in course recommendation systems, where courses preferred by users with similar profiles are highlighted as beneficial options (Aucancela, 2023; Butmeh et al., 2024; Zhou & Zhang, 2024; Dhananjaya, et al., 2024) However, CF faces significant challenges:

- **Data sparsity**: When user interaction data is limited, the system struggles to make accurate predictions.
- **Cold-start problem**: The absence of sufficient data for new users or items hinders recommendation accuracy.
- **Scalability**: As datasets grow, the computational cost increases, requiring optimization to maintain efficiency.

To overcome these challenges, researchers propose hybrid approaches that combine CF with other methods, such as content-based filtering, and the integration of trust metrics to enhance accuracy and adaptability (Alahmadi & Alruwaili, 2021; Dhananjaya, et al., 2024; Xiong, et al., 2024). To tackle these issues, some research recommends using hybrid approaches and integrating trust metrics to improve the accuracy and adaptability of recommendations (Raza et al., 2024).

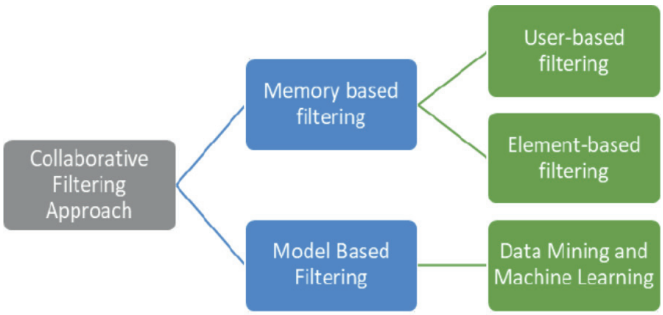


Figure 1 – Categorization of the collaborative filtering approach (Aucancela, 2023)

2.2.2 Content-Based Filtering (CBF). Content-Based Filtering (CBF) relies on analyzing item attributes to recommend items similar to those a user has interacted with in the past. It utilizes algorithms such as decision trees and Bayesian classifiers to match user preferences(Figure 2) with item features, including course topics, difficulty levels, or skill requirements (Zhou & Zhang, 2024; Alahmadi & Alruwaili, 2021). CBF's strengths lie in its ability to handle new users by leveraging their explicitly provided preferences and its independence from other users' data.

However, CBF is prone to overspecialization, where the system repeatedly recommends items similar to those already consumed, leading to a lack of diversity (Zhou & Zhang, 2024; Xiong, et al., 2024). For example, in educational settings, CBF might recommend courses closely aligned with a user's past selections without introducing diverse learning opportunities. Additionally, CBF struggles with new item recommendations since it relies on pre-existing item descriptions. Addressing these challenges often involves hybrid methods that combine the strengths of multiple approaches (Zhou & Zhang, 2024; Dhananjaya, et al., 2024).

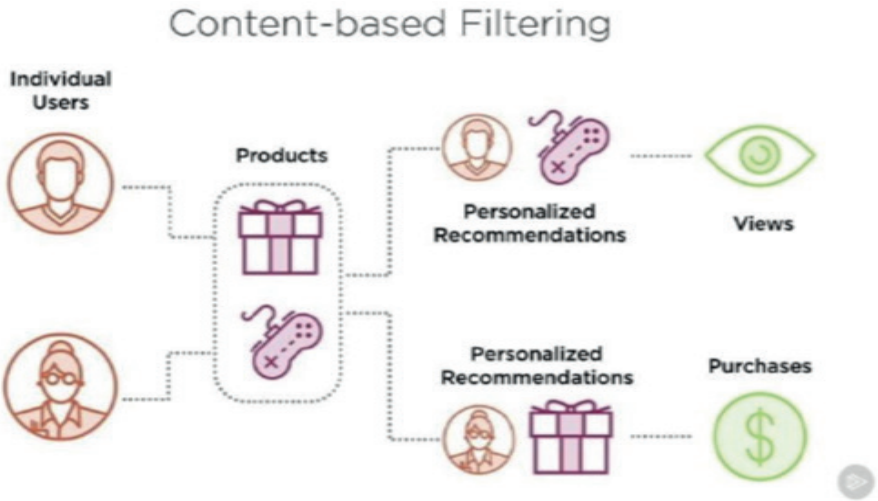


Figure 2 – Content-based Filtering (Zhou & Zhang, 2024)

2.2.3 Hybrid Methods

Hybrid recommendation systems integrate multiple approaches, such as CF and CBF, to overcome their individual limitations and enhance recommendation quality. For instance, hybrid systems can mitigate CF's cold-start problem by incorporating content-based techniques to generate recommendations for new users or items. They can also reduce CBF's (Figure 3) overspecialization by introducing diversity through collaborative methods (Thammarak, et al., 2024; Zhou & Zhang, 2024; Alahmadi & Alruwaili, 2021; Yazdi, et al., 2024).

Several hybridization techniques are employed in ERS, including weighted, switching, mixed, and feature augmentation methods. These approaches combine

CF and CBF dynamically, optimizing performance based on user data and context (Alahmadi & Alruwaili, 2021; Yazdi, et al., 2024). For example, a hybrid attribute-based system tested in an e-learning environment successfully integrated user preferences, learning styles, and knowledge levels to recommend personalized courses while addressing data sparsity and cold-start challenges (Thammarak, et al., 2024). Such systems have demonstrated superior accuracy and adaptability, particularly in addressing diverse learner needs and preferences (Butmeh, et al., 2024; Thammarak, et al., 2024).

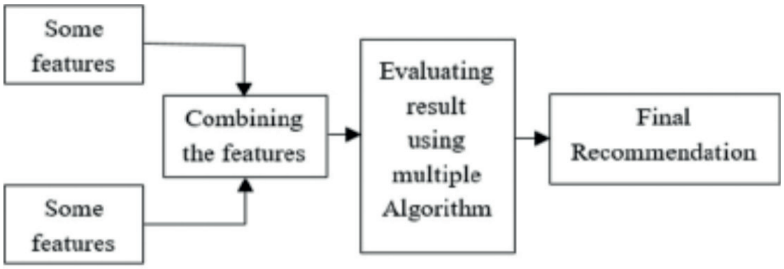


Figure 3 – Hybrid Recommendation System (Alahmadi & Alruwaili, 2021)

2.2.4 AI-Based Techniques

AI-based techniques have transformed educational recommendations by introducing advanced methods like machine learning (ML), deep learning (DL), and knowledge graphs. ML algorithms, including decision trees, random forests, and gradient boosting, have been effectively applied to match students with suitable courses based on their skills and preferences. These methods enhance recommendation accuracy by addressing imbalanced data and incorporating dynamic user behavior (Phalle & Bhushan, 2024).

Deep learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks, have further advanced ERS by capturing complex patterns in user interactions and preferences. For instance, deep learning models like DeepFM and Bidirectional LSTM have demonstrated high performance in personalized course recommendations, addressing cold-start issues and improving temporal dynamics (Masthoff, 2023; et al., 2024).

Knowledge graphs (KGs) offer another innovative approach by structuring data into semantic networks, enabling more precise and context-aware recommendations. In educational contexts, KGs have been used to address data sparsity and enhance adaptability by combining semantic similarity with item features. For example, the TransAR-CF model improved recommendation accuracy and user satisfaction by leveraging KGs for course recommendations in higher education (Shuang et al., 2024; Rahman et al., 2022).

The integration of these AI-based techniques has significantly enhanced the

scalability, accuracy, and personalization of ERS, addressing challenges like the cold-start problem, data sparsity, and scalability while ensuring a more effective and engaging learning experience (Shuang, et al., 2024; Rahman, et al., 2022; Guo, et al., 2024; Raza et al., 2024).

2.3 Challenges in Educational Recommender Systems

Educational Recommender Systems (ERS) face several challenges that limit their efficiency, scalability, and user satisfaction. Below, we discuss the most pressing issues and explore proposed solutions based on the latest research.

2.3.1 Cold-Start Problem

The cold-start problem arises when ERS encounter new users or items with insufficient data for accurate recommendations (Thammarak, et al., 2024; Yazdi, et al., 2024). Traditional collaborative filtering (CF) methods rely heavily on user interactions, making it difficult to generate recommendations for new learners or courses. Hybrid models, which combine CF with content-based filtering (CBF), have been widely adopted to address this challenge (Alahmadi & Alruwaili, 2021). For example, the hybrid attribute-based system proposed integrates user preferences, learning styles, and knowledge levels to create more accurate recommendations for new users (Thammarak, et al., 2024). Knowledge graphs, as highlighted, also help mitigate this issue by providing a semantic understanding of relationships between users and items, enriching the recommendation process (Shuang & Hang, 2024).

2.3.2 Data Sparsity and Scalability

Data sparsity refers to the lack of sufficient user-item interactions, which hinders the performance of CF algorithms (Butmeh, et al., 2024; Xiong, et.al. 2024). Matrix factorization techniques, such as singular value decomposition (SVD) and BiVAE, have been effective in addressing sparsity by capturing latent user and item features (Aldayel et al., 2023; Kamal, et al., 2024). Scalability is another critical challenge, particularly in large-scale educational systems. Distributed computing approaches, including the use of parallel processing and cloud-based infrastructures, have been proposed to improve scalability (Phalle & Bhushan, 2024). These methods enable ERS to process vast datasets efficiently, ensuring timely and relevant recommendations for a growing user base.

2.3.3 Personalization

Personalization is central to ERS, aiming to tailor recommendations to individual learning needs and preferences. However, achieving true personalization is challenging due to the dynamic nature of learners' goals and behaviors (Alfaifi, 2024; Guo et al., 2024). Advances in machine learning (ML) and deep learning (DL) have significantly enhanced the ability of ERS to adapt to individual learning paths.

This study proposes a dynamic recommendation system that leverages bidirectional LSTM networks in combination with mindfulness mechanisms to effectively model and adapt to both the short-term behaviors and long-term preferences of users (Guo et al., 2024). Another approach introduces a hybrid method that clusters students based on their multidimensional skill profiles, aiming

to generate more personalized and precise recommendations tailored to each learner (Alfaifi, 2024).

2.3.4 Privacy and Ethical Issues

Privacy and ethical considerations play a crucial role in educational recommendation systems (ERS), particularly when dealing with sensitive student data. Ensuring transparency, fairness, and robust data protection is essential for maintaining user trust and adhering to ethical standards. Research indicates that privacy-preserving technologies, such as differential privacy and encryption techniques, can effectively safeguard user information. At the same time, ongoing monitoring is required to address ethical concerns like algorithmic bias and disparities in access to personalized learning opportunities. User-centered design methodologies are seen as a promising direction for developing fair, trustworthy algorithms and promoting equity in educational settings (Xiong, et.al., 2024).

2.4 Case Studies and Applications

2.4.1 Case Study 1: Course Recommendation Systems in Higher Education

In recent years, course recommendation systems in universities have become a subject of considerable interest among researchers. A systematic review of publications over the past decade identified collaborative filtering as the predominant approach in the development of such systems, with hybrid methods and content-based filtering also being widely applied. These systems make recommendations based on student demographics, academic performance, and personal learning preferences. They have proven to be highly effective in improving retention rates, increasing student engagement, and supporting the creation of personalized learning pathways. For example, several systems use demographic data to tailor course recommendations, fostering better learning experiences for students. Future research directions include exploring multi-view data approaches and incorporating fairness in recommendation processes (Butmeh et al., 2024; Dhananjaya et al., 2024; Thammarak et al., 2024).

2.4.2 Case Study 2: MOOC Recommendation Systems

Massive Open Online Course (MOOC) recommendation systems have gained considerable attention due to their potential to personalize learning at scale. DeepFM and DORIS are two prominent deep learning-based systems used to recommend courses in MOOCs. These systems are designed to overcome common challenges such as cold-start issues, scalability, and data sparsity. DORIS, in particular, leverages deep learning techniques, including DeepFM and TextRank, to enhance recommendation accuracy by processing textual data. The system has shown significant improvements in recommending courses that match learners' preferences and needs. These systems utilize personalized algorithms to suggest courses, taking into account dynamic user preferences and interactions with educational content. Further improvements could focus on integrating advanced deep learning techniques and addressing real-time user feedback to refine recommendations (Zheng, 2022; Yazdi, et al., 2024;Wu, 2023).

2.4.3 Case Study 3: Personalized Learning in E-Learning Platforms

Personalized learning systems in e-learning platforms aim to optimize the educational experience by tailoring content and resources to individual learners. An attribute-based hybrid recommendation system, which combines collaborative filtering and content-based approaches, has been proposed to address the commonly encountered cold start problem in e-learning environments. This approach builds models of both users and courses by taking into account learners' preferences, cognitive styles, and prior knowledge. Empirical studies have demonstrated that when integrated with machine learning techniques—particularly decision trees and random forests—this method proves effective in delivering targeted course recommendations. The use of machine learning algorithms enhances the accuracy of predicting students' course preferences, which in turn contributes to improving the quality of the learning process. Personalized recommendations support students by suggesting courses aligned with their current needs and academic goals (Butmeh et al., 2024; Thammarak, et al., 2024; Phalle & Bhushan, 2024).

2.5 Evaluation of Educational Recommender Systems

2.5.1 Metrics for Evaluation

The evaluation of educational delivery systems (ERS) requires the use of various performance indicators to determine their accuracy and effectiveness. The most commonly used indicators include precision, recall, accuracy, novelty, and diversity. Accuracy and recall — are important for assessing the compliance of the courses offered with user demand, and accuracy — indicates how correctly the system can predict user preferences. Innovation and Variety also play an important role, as they characterize the ability of the system to provide different recommendations, and not monotonous ones, which, in turn, will improve the overall user experience. These metrics help deliver content that is relevant, diverse, and meaningful, thereby improving user satisfaction and engagement (Li et al., 2024; Butmeh et al., 2024; Shuang et al., 2024).

2.5.2 Evaluation Approaches

Various assessment methods are used to evaluate the performance of educational offering systems (ERS). Prediction accuracy is usually determined by metrics such as mean absolute error (MAE) and mean square root error (RMSE), which measure the difference between predicted ratings and actual user feedback. In addition, user-oriented assessment approaches based on surveys and user opinions play an important role in assessing the subjective quality of recommendations. These methods allow you to assess how well the recommendations meet the needs and expectations of users.

In addition, modern assessment methods based on the theoretical foundations of decision-making are being studied, which make it possible to comprehensively assess the performance of the system, taking into account factors such as user satisfaction, adaptive capabilities of the system and long-term efficiency (Li et al., 2024; Butmeh et al., 2024; Urdaneta-Ponte et al., 2021).

2.5.3 Challenges in Evaluation

Assessment of the effectiveness of educational systems (ERS) presents a number of difficulties, primarily due to the complex and multidimensional nature of educational data. Furthermore, the diversity of students' needs and preferences necessitates the creation of a comprehensive evaluation system that can capture the varying impacts of recommendations in different educational contexts. Additionally, ethical considerations and ensuring fairness remain crucial issues, particularly when processing students' confidential data. Researchers are exploring multidimensional evaluation systems that better capture the nuances of educational recommendations, including factors such as diversity, fairness, and privacy (Butmeh et al., 2024; Dhananjaya et al., 2024; Xiong et.al., 2024).

Results and discussion. One of the most promising areas of future research in educational recommender systems (ERS) is the integration of advanced artificial intelligence (AI) models such as learning reinforcement, deep learning and neural networks. According to recent research, MOOC and e-learning environments are deep learning approaches such as convolutional neural networks (CNN), repetitive neural networks (RNNs), and long-term short-term memory (LSTM) architectures that demonstrate important potential in improving the accuracy of representations in contexts such as (Wu, 2023; Alfaifi, 2024). It still requires the study of dynamic and adaptive recommendation processes that can continuously learn from the models of users of artificial intelligence systems and provide personalized course recommendations in real time. In addition, the ability to optimize recommendation systems by creating a responsive educational environment has an impact on changes in the preferences and learning behavior of users of reinforcement learning (Wu, 2023; Guo et al., 2024). Future research aims to provide personalized learning pathways, improving long-term learning outcomes, while developing the use of reinforcement learning based on the student's progress.

One of the main problems in environments where new users or elements (e.g. courses) are often introduced remains the problem of cold start education recommendation systems (ERS). To solve this problem, hybrid models and approaches are offered on the following eLearning and MOOC platforms. In solving cold startup problems, attribute-based hybrid recommendation systems that combine collaborative filtering into content with methods using user data and element attributes to create recommendations for new users or courses have shown promising results (Thammarak, et al., 2024; Phalle et al., 2021). In addition, machine learning techniques such as decision trees and random forests that group elements based on common features are used to improve scalability and offer quality by working with large data sets (Phalle & Bhushan, 2024). Continued development in this area should focus on integrating new data sources and improving the adaptability of these models to handle cold-start problems across diverse educational contexts.

As ERS become more prevalent in personalized education, addressing ethical concerns and ensuring data privacy are crucial. This is due to the large accumulation of personal and academic data needed to personalize learning, especially confidential

data of minors and students, which raises serious concerns about confidentiality. The researchers noted the importance of designing recommendation systems that provide clear and meaningful recommendations while ensuring the protection of users ' privacy. At the same time, in the development of artificial intelligence models, special attention should be paid to ensuring that the recommendations offered by educational systems (ERS) are fair and impartial and that they are transparent and interpretable in order to build user trust. From the point of view of ethical issues, the risk of increasing bias or limiting educational opportunities should also be taken into account (Xiong et.al., 2024). Future research should focus on developing the foundations of ethical artificial intelligence in education, considering recommender systems (ERS) as a tool that contributes to the formation of an inclusive and fair learning environment.

Conclusion. The survey performed on recommendation algorithms used to detect the preferences in education has exposed the immense capacity of recommender systems in empowering the process of personalized learning. The most used approaches like filtering types: collaborative and content-based, and their hybrid types have been used in studies which pursued to offer students accurate recommendations and concurrently address issues and challenges related with development which include the problems of data sparsity, scalability, and cold-start issue. The application of artificial intelligence approaches in the implementation of educational recommender systems have increased accuracy of recommendations and have shown promising results in the improvement of adaptability of ERS. As modern conditions of learning require from students to evolve in their knowledge as fast as possible, the recommendations of e-learning systems are becoming more and more dynamic and this requires sophisticated algorithms, which can easily adapt to behaviors, preferences and growing learning contexts of students.

Future research should focus on the improvement of hybrid approaches, knowledge graphs and building new model architectures of deep learning, because they have shown the possibility of overcoming actual limitations in ERS. Moreover, widening the usage of ERS to various contexts of education can lead to the further improvement of this field. The future pathway of ERS evolution will focus not only on the personalization of recommendations, but it can also play significant role in detecting the gaps and involvement of learners in the diversity of learning environments.

References

Bing Wu, (2023) “Personalized Hybrid Recommendation Algorithm for MOOCs Based on Learners’ Dynamic Preferences and Multidimensional Capabilities”. https://doi.org/10.3390/app13095548 (In English)

Dimah Alahmadi and Fatimah Alruwaili (2021) “Deep Learning for MOOCs Course Recommendation Systems: State of the Art Survey”, vol. 12, issue 11. (In English)

Dhananjaya Gm, R.H. Goudar, Anjanabhargavi A. Kulkarni and Vijayalaxmi Rathod (2024) “A Digital Recommendation System for Personalized Learning to Enhance Online Education: A Review”. https://doi.org/10.1109/access.2024.3369901 (In English)

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

Fangxia Zheng, (2022) "Personalized Education Based on Hybrid Intelligent Recommendation System". https://doi.org/10.1155/2022/1313711 (In English)

Hala Butmeh and Abdallatif Abu-Issa (2024) "Hybrid attribute-based recommender system for personalized e-learning with emphasis on cold start problem". https://doi.org/10.3389/fcomp.2024.1404391 (In English)

Hadis Ahmadian Yazdi, Seyyed Javad Seyyed Mahdavi and Hooman Ahmadian Yazdi (2024) "Dynamic educational recommender system based on Improved LSTM neural network". https://doi.org/10.1038/s41598-024-54729-y (In English)

Judith Masthoff, (2023) "Hybrid session-aware recommendation with feature-based models", vol. 34. — P. 691–728 (In English)

Madhusree Kuanr and Puspanjali Mohapatra (2021) "Assessment Methods for Evaluation of Recommender Systems: A Survey", vol. 46. (In English)

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

Lianhuan Li, Zheng Zhang and Shaoda Zhang (2021) "Hybrid Algorithm Based on Content and Collaborative Filtering in Recommendation System Optimization and Simulation". https://doi.org/10.1155/2021/7427409 (In English)

Luoshuang Shuang and Chu Hang (2024) "A Study on the Role of Knowledge Mapping Technology in Constructing a Knowledge System for Student Management and Promoting Innovation in and Political Education". https://doi.org/10.2478/amns-2024-2364 (In English)

Nabila Kamal, Farhana Sarkar, Arifur Rahman and Sazzad Hossain (2024) "Recommender System in Academic Choices of Higher Education: A Systematic Review". https://doi.org/10.1109/access.2024.3368058 (In English)

Margarita Alejandra Aucancela, (2023) "Educational Recommender Systems: A Systematic Literature Review". https://doi.org/10.22492/issn.2435-9467.2023.74 (In English)

Mashael Aldayel, Abeer Al-Nafjan, Waleed M. Al-Nuwaiser, Ghadeer Alrehaili and Ghadi Alyahya (2023) "Collaborative Filtering-Based Recommendation Systems for Touristic Businesses, Attractions, and Destinations". https://doi.org/10.3390/electronics12194047 (In English)

Ms. Tejashri Sharad Phalle and Prof. Shivendu Bhushan (2024) "Content Based Filtering and Collaborative Filtering: A Comparative Study". https://doi.org/10.53555/jaz.v45is4.4158 (In English)

Md. Mijanur Rahman, Ismat Ara Shama, Siamur Rahman and Rahmatullah Nabil (2022) "Hybrid Recommendation System to Solve Cold Start Problem". https://doi.org/10.5281/zenodo.7026121 (In English)

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

Pengyu Guo, Mohd Khalid Mohamad Nasir and Yishuai Xu (2024) "Collaborative Filtering Recommender System for Online Learning Resources with Integrated Dynamic Time Weighting and Trust Value Calculation". https://doi.org/10.18421/tem132-49 (In English)

Rui Zhou and Nan Zhang (2024) "Research on Recommendation Methods for Scientific and Technological Information and Their Application in College Education — Based on Knowledge Graphs". https://doi.org/10.2478/amns-2024-2832 (In English)

Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah and Amirmohammad Kazemeini (2024) "A Comprehensive Review of Recommender Systems: Transitioning from Theory to Practice". https://doi.org/10.48550/arXiv.2407.13699 (In English)

Yousef H. Alfaifi, (2024) "Recommender Systems Applications: Data Sources, Features, and Challenges". https://doi.org/10.3390/info15100660 (In English)

Yinping Ma, Rongbin Ouyang, Xinzheng Long, Zhitong Gao, Tianping Lai and Chun Fan (2023) “DORIS: Personalized Course Recommendation System Based on Deep Learning”. https://doi.org/10.1371/journal.pone.0284687 (In English)

Zhang Xiong, Haoxuan Li, Zhuang Liu, Zhuofan Chen, Hao Zhou, Wenge Rong and Yuanxin Ouyang (2024) “A Review of Data Mining in Personalized Education: Current Trends and Future Prospects”. https://doi.org/10.1007/s44366-024-0019-6 (In English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 110–131

https://doi.org/10.32014/2025.2518-1726.367

UDC 004.912; 004.93; 004.8; 004.5; 004.6

A. Yerimbetova[1,2], U. Berzhanova[1,2], E. Daiyrbayeva[1,3], B. Sakenov[1], M. Sambetbayeva[1,4], 2025.

[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

DEVELOPMENT OF A PARALLEL CORPUS FOR KAZAKH SIGN LANGUAGE TRANSLATION AND TRAINING OF THE TRANSFORMER MODEL

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: nurbekkyzy_e@mail.ru, 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: Kazakh Sign Language (KSL) serves as a primary means of communication for people with hearing and speech impairments. This study focuses on analyzing the syntactic structure of KSL and identifying its differences from the features of spoken Kazakh language. Additionally, a specialized linguistic analyzer was developed to transform Kazakh texts into KSL glosses, enabling the creation of a parallel corpus. The study of KSL is significant not only from

a scientific perspective but also for the development of tools that contribute to building an inclusive society. As part of the research, a machine translation system based on the Transformer model was trained using the parallel corpus. As a result, high translation accuracy was achieved, demonstrating the potential for enhancing communication accessibility. This approach represents an important step in the automated processing of KSL. Such technologies aim to improve the educational process and social integration of people with special needs. To analyze KSL syntax, Python 3.10, Stanza, and libraries such as PyTorch, NumPy, and Pandas were utilized. Over 500 sentences were analyzed, revealing the flexibility of word order and the features of visual-spatial structure. The analysis of sentences, including reversible, nonreversible, locative, animate, inanimate, heavy, and nonheavy, enabled the description of the grammatical structure of KSL. The research results serve as a foundation not only for identifying the syntactic patterns of the language but also for developing new translation models. The research highlighted key challenges, such as the limited availability of annotated data. Future work will focus on integrating video data and expanding evaluation metrics. The proposed methods form the basis for developing inclusive information technologies and improving communication with people with special needs. This work plays an important role in expanding the inclusive potential of technologies and opens new avenues for scientific research.

Key words: Kazakh Sign Language, parallel corpus, Transformer architecture, NLP, machine learning

А.С. Еримбетова [1,2], У.Г. Бержанова [1,2], Э.Н. Дайырбаева [1,3], Б.Е. Сәкенов [1], М.А. Сәмбетбаева [1,4], 2025.

[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

ҚАЗАҚ ЫМ ТІЛІНЕ АУДАРУ ҮШІН ПАРАЛЛЕЛЬ КОРПУС ҚҰРУ ЖӘНЕ TRANSFORMER МОДЕЛІН ОҚЫТУ

Еримбетова Айгерим Сембековна — PhD, техн. ғылым. ғандидаты, 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 Габитқызы — 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 Нурбекқызы — магистр, Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университетінің бағдарламалық инженерия кафедрасының аға оқытушысы, Алматы, Қазақстан; ҚР ҒЖБМ ҒК Ақпараттық және есептеуіш технологиялар институтының ғылыми қызметкері, Алматы, Қазақстан,
E-mail: nurbekkyzy_e@mail.ru; ORCID ID: https://orcid.org/0000-0002-4255-5456;
Сакенов Бакжан Ерланұлы — 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;
Сәмбетбаева Мадина Аралбаевна — PhD, Л.Н.Гумилев атындағы Еуразия ұлттық университетінің доценті, Алматы Қазақстан; 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. Қазақ ым тілі (ҚЫТ) есту және сөйлеу қабілеті бұзылған адамдар үшін негізгі қарым-қатынас құралы болып табылады. Бұл зерттеу ҚЫТ-нің синтаксистік құрылымын талдауға және қазақ тілінің ауызекі тілі ерекшеліктерінен айырмашылықтарын анықтауға бағытталған. Сонымен қатар, қазақ тіліндегі мәтіндерді ҚЫТ глоссқа айналдыру үшін арнайы лингвистикалық талдаушы жасалды, бұл параллель корпус құруға мүмкіндік берді. ҚЫТ-нің зерттелуі тек ғылыми маңыздылыққа ғана емес, сонымен қатар инклюзивті қоғам құруға қажетті құралдар жасауға ықпал етеді. Зерттеуде параллель корпусты қолдана отырып, Transformer моделі негізінде машиналық аударма жүйесі оқытылды. Нәтижесінде аударманың жоғары дәлдігіне қол жеткізілді, бұл коммуникация қолжетімділігін арттырудың әлеуетін көрсетті. Бұл тәсіл қазақ ым тілін автоматтандырылған түрде өңдеуде маңызды қадам болып табылады. Мұндай технологиялар ерекше қажеттіліктері бар адамдардың білім алу және әлеуметтік интеграция үдерістерін жақсартуға бағытталған. ҚЫТ синтаксисін талдау үшін Python 3.10, Stanza, PyTorch, NumPy және Pandas кітапханалары қолданылды. 500-ден астам сөйлемдер зерттеліп, сөз тәртібінің икемділігі мен визуалды-кеңістіктік құрылымының ерекшеліктері анықталды. 500-ден астам сөйлемдер бойынша талдау жүргізіліп, негізгі назар қайтымды, қайтымсыз, мекенді, жанды, жансыз, күрделі және күрделі емес сөйлемдер бойынша жүргізілген талдау қазақ ым тілінің грамматикалық құрылымын сипаттауға мүмкіндік берді. Зерттеудің нәтижелері тілдің синтаксистік заңдылықтарын анықтауға ғана емес, сонымен қатар жаңа аударма модельдерін әзірлеуге де негіз болады. Зерттеу барысында аннотацияланған деректердің шектеулі болуы сияқты негізгі қиындықтар анықталды. Болашақта бейне деректерді біріктіру және бағалау көрсеткіштерін кеңейту бағытында жұмыс істеу жоспарлануда. Ұсынылған әдістер инклюзивті ақпараттық технологияларды дамытуға және ерекше қажеттіліктері бар адамдармен қарым-қатынасты жақсартуға негіз болады. Бұл жұмыс технологиялардың инклюзивті әлеуетін кеңейтуде және жаңа ғылыми зерттеулерге жол ашуда маңызды рөл атқарады.

Түйін сөздер: Қазақ ым тілі, параллель корпус, Transformer архитектурасы, NLP, машиналық оқыту

Алғыс: *Бұл зерттеу Қазақстан Республикасының Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (Грант № BR24992875).*

А.С. Еримбетова[1,2], У.Г. Бержанова [1,2], Э.Н. Дайырбаева [1, 3], Б.Е. Сакенов [1], М.А. Сәмбетбаева[1,4], 2025.

[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, Астана, Қазақстан.

E-mail: berzhanovaulmekenn@gmail.com

РАЗРАБОТКА ПАРАЛЛЕЛЬНОГО КОРПУСА ДЛЯ ПЕРЕВОДА КАЗАХСКОГО ЖЕСТОВОГО ЯЗЫКА И ОБУЧЕНИЕ МОДЕЛИ TRANSFORMER

Еримбетова Айгерим Сембековна — PhD, к.т.н., ассистент 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 Габиткызы — докторант 2-го курса по специальности 8D06101 - Информационные системы Казахского национального университета имени Аль-Фараби; младший научный сотрудник Института информационных и вычислительных технологий Комитета науки Министерства образования и науки РК, Алматы, Казахстан,
E-mail: berzhanovaulmekenn@gmail.com; ORCID ID: https://orcid.org/0009-0000-2467-5721;
Дайырбаева Эльмира Нурбеккызы — магистр, старший преподаватель кафедры Программной инженерия Казахского национального исследовательского технического университета имени К.И. Сатпаева, Алматы, Казахстан; научный сотрудник Института информационных и вычислительных технологий Комитета науки Министерства образования и науки РК, Алматы, Казахстан,
E-mail: nurbekkyzy_e@mail.ru; ORCID ID: https://orcid.org/0000-0002-4255-5456;
Сакенов Бакжан Ерланулы — магистр, инженер-программист Института информационных и вычислительных технологий Комитета науки Министерства образования и науки РК, Алматы, Казахстан,
E-mail: sbakzhan22@gmail.com; ORCID ID: https://orcid.org/0000-0002-9849-6176;
Самбетбаева Мадина Аралбаевна — PhD, доцент Евразийского национального университета им. Л.Н. Гумилева, Казахстан, Алматы, Сатпаев 22; ведущий научный сотрудник Института информационных и вычислительных технологий Комитета науки Министерства образования и науки РК, Астана, Казахстан,
E-mail: madina_jgtu@mail.ru; ORCID ID: https://orcid.org/0000-0001-9358-1614.

Аннотация: Казахский жестовый язык (КЖЯ) является основным средством общения для людей с нарушениями слуха и речи. Это исследование

направлено на анализ синтаксической структуры КЖЯ и выявление отличий от особенностей разговорного казахского языка. Кроме того, для преобразования текстов на казахском языке в глоссы КЖЯ был создан специализированный лингвистический анализатор, что позволило разработать параллельный корпус. Изучение КЖЯ важно не только с научной точки зрения, но и для разработки инструментов, способствующих созданию инклюзивного общества. В рамках исследования была обучена система машинного перевода на основе модели Transformer, используя параллельный корпус. В результате была достигнута высокая точность перевода, что продемонстрировало потенциал для повышения доступности коммуникации. Этот подход представляет собой важный шаг в автоматизированной обработке казахского жестового языка. Подобные технологии направлены на улучшение процесса обучения и социальной интеграции людей с особыми потребностями. Для анализа синтаксиса КЖЯ использовались Python 3.10, Stanza, библиотеки PyTorch, NumPy и Pandas. Было исследовано более 500 предложений, что позволило выявить особенности гибкости порядка слов и визуально-пространственной структуры. Анализ предложений, включая обратимые, необратимые, локативные, одушевлённые, неодушевлённые, сложные и простые конструкции, позволил описать грамматическую структуру казахского жестового языка. Результаты исследования послужили основой не только для выявления синтаксических закономерностей языка, но и для разработки новых моделей перевода. В ходе исследования были выявлены основные трудности, такие как ограниченное количество аннотированных данных. В будущем планируется работа над интеграцией видеоданных и расширением оценочных показателей. Предложенные методы лежат в основе развития инклюзивных информационных технологий и улучшения коммуникации с людьми с особыми потребностями. Эта работа играет важную роль в расширении инклюзивного потенциала технологий и открывает новые направления для научных исследований.

Ключевые слова: Казахский жестовый язык, параллельный корпус, архитектура Transformer, NLP, машинное обучение

Кіріспе. Дүниежүзілік денсаулық сақтау ұйымының (ДДҰ) мәліметтері бойынша, әлем халқының шамамен 5%-ы (460 миллион адам) есту қабілетінен айрылған. Есту қабілеті нашар адамдар үшін ымдау тілдері негізгі қарым-қатынас құралы болып табылады, әлемде шамамен 300 түрлі ым тілі бар. Алайда, бұл тілдерді халықтың тек 1%-ы ғана меңгерген, олардың көпшілігі есту қабілеті нашар адамдар мен олардың отбасылары (Perea-Trigo et al., 2024). Есту және сөйлеу қабілеті бұзылған адамдар үшін ым тілі негізгі қарым-қатынас құралы болып табылады. Алайда, олар қоғаммен өзара әрекеттесуде жиі қиындықтарға ұшырайды. Қазақстанда шамамен жарты миллион адам ым тілін қолданады (Amangeldi et al., 2020a), бірақ олардың толыққанды қарым-қатынас жасау мүмкіндігі шектеулі болып қала береді

Ым тілі (Amangeldi et al., 2020b) – ақпарат алмасу үшін қол қимылдарын, бет-әлпет мимикасын және дене қозғалыстарын қолданатын визуалды байланыс тәсілі. Есту және сөйлеу қабілеті шектеулі адамдар үшін коммуникациялық кедергілерді жоюға бағытталған маңызды шешімдердің бірі болып - машиналық аударма жүйелері табылады. Алайда, ауызекі тілге арналған технологиялармен салыстырғанда, ым тілдерін аударуға арналған әдістер мен жүйелер айтарлықтай дамымаған. Бұл, ең алдымен, ым тілдерінің визуалды-кеңістіктік құрылымына байланысты, өйткені олар қол қимылдарының, саусақтардың орналасуының, мимиканың және дене қалпының үш өлшемді кеңістікте әрекеттесуіне негізделеді. Сонымен қатар, ым тілдерінде стандартты грамматикалық құрылымдардың болмауы жазбаша мәтінді ым тіліне аударуды айтарлықтай қиындатады.

Көпмасштабты кеңістіктік-уақыттық графтық жиынтықталу (MS-G3D) архитектурасы дене және саусақ буындарынан тұратын қаңқалық графты пайдаланып, зерттеуінде ұсынылған. Авторлар жұмыстары (Vázquez-Enríquez et al., 2021) көрсеткендей, бұл тәсіл семантикалық тұрғыда маңызды өзара байланыстарды дәл түсіруге мүмкіндік беріп, AUTSL деректер жиынтығында жоғары нәтижелерге қол жеткізген. Сонымен қатар, 3D-CNN (S3D) архитектурасымен салыстыру және трансферлік оқыту тәжірибелері MS-G3D әдісінің ISLR саласында бәсекеге қабілетті әрі бейімделгіш екенін дәлелдеген.

ҚЫТ параллель корпусын әзірлеу және зерттеу есту және сөйлеу қабілеті шектеулі адамдардың қазақстандық қоғамға ықпалдасуын жеңілдететін және коммуникацияны едәуір жақсартатын маңызды бағыт болып табылады. Ым тілі есту және сөйлеу қабілеті бұзылған жандар үшін негізгі қарым-қатынас құралының рөлін атқарады. Қазақстанда, көптеген басқа мемлекеттердегідей, бұл санаттағы адамдар білім алуда, ақпаратқа қол жеткізуде және әлеуметтік ортаға бейімделуде елеулі қиындықтарға тап болады. Мұндай кедергілердің басты себебі – ым тілі туралы қоғамның хабардарлығының төмендігі, білім беру бағдарламаларының жеткіліксіздігі, сондай-ақ осы саланы дамытуға арналған зерттеулер мен ресурстардың тапшылығы. ҚЫТ зерттеуге бағытталған ғылыми бастамалар есту және сөйлеу қабілеті бұзылған адамдардың білім мен ақпаратқа қол жеткізуін қамтамасыз етуге, сондай-ақ олардың мәдени және лингвистикалық ерекшеліктеріне қатысты түсінікті тереңдетуге бағытталған. ҚЫТ үшін параллель корпустар құру және ҚЫТ нұсқауларын жетілдіру зерттеушілерге, тәрбиешілерге және мамандарға таптырмас құрал бола алады. Бұл жұмыстар ҚЫТ оқытуға арналған заманауи әдістемелерді әзірлеуге, аударма ресурстарын жетілдіруге және тілдік белгілердің құрылымдық ерекшеліктерін тереңірек түсінуге мүмкіндік береді. Осы бөлімде ҚЫТ және параллель корпусты дамыту саласындағы зерттеулердің негізгі мақсаттары мен кездесетін қиындықтары талқыланады.

Параллель корпус сурдоаударма жүйелерін дамытуда маңызды рөл атқарады. Сондай-ақ, ым тілдерінің өзіндік грамматикалық құрылымы мен

лексикасы оларды мәтінге аудару жүйелерін әзірлеуде елеулі қиындықтар тудырады. Осындай жүйелерді құруға (Bertin-Lemée et al., 2022; Jiao et al., 2024) зерттеулері түрлі тәсілдерді қарастырып, мәліметтердің шектеулілігі, қолданыстағы модельдердің төмен дәлдігі және ым тілдерінің лингвистикалық ерекшеліктерін ескерудің маңыздылығы сияқты негізгі мәселелерді анықтайды.

Bertin-Lemée және басқалары (2022) зерттеуінде мәтінді ым тілдерінің иерархиялық формалды сипаттамаларына AZee аудару әдісі ұсынылған. Бұл әдіс параллель корпустарды қолданып, жаңа мәлімдемелерді аудару үшін мәтін сегменттерін қайталап ауыстыру арқылы бірнеше аударма нұсқаларын құрастырады. Әдіс ым тілі құрылымдарын мүмкіндігінше сақтауға мүмкіндік береді және нәтижелерді аватарларды синтездеу үшін қолдануға негізделген.

Ым тілінен мәтінге аудару (SLT) бейнелерді мәтіндік сөйлемдерге түрлендіруге бағытталған маңызды міндет (Jiao et al., 2024). Дәстүрлі тәсілдерде визуалды ұсынуды үйрету үшін глосс аннотациялары қолданылса да, олардың жоғары құны масштабталуды қиындатады. Осы мәселені шешу үшін зерттеулерде визуалды туралауды алдын ала дайындау (VAP) әдісі ұсынылып, ол визуалды және мәтіндік токендерді сәйкестендіру арқылы семантикалық тұрғыда мәнді ақпаратты тиімді пайдалануға мүмкіндік берді. Тәжірибелік нәтижелер VAP тәсілінің SLT өнімділігін арттырып, глосс негізіндегі әдістермен салыстырғанда айырмашылықты едәуір қысқартатынын көрсетті. Зерттеулерде қолданылатын әртүрлі әдістер мәтіндер мен бейнелерді ым тілдеріне аудару мәселелерінің көпқырлылығын көрсетеді. Заманауи нейрондық желі модельдері айтарлықтай жетістіктерге қол жеткізгенімен, деректердің шектеулілігі мен ым тілдерінің ерекшеліктерін ескерудегі қиындықтар әлі де бар.

ҚЫТ және параллель корпусты дамыту зерттеушілердің алдында тұрған негізгі міндеттері мәдени мұраны сақтау, білімге қолжетімділікті кеңейту, қарым-қатынасты нығайту, лингвистикалық зерттеулерді дамыту, әлеуметтік интеграцияны қолдау және кәсіби аудармашыларға көмек көрсету болып табылады. Бұл зерттеу қазақ тіліндегі мәтіндерді ҚЫТ глоссына аударуды автоматтандыру және параллель корпусты дамыту арқылы есту және сөйлеу қабілеті бұзылған адамдар үшін тілдік және мәдени кедергілерді жоюға бағытталған. Жоба ҚЫТ-нің грамматикалық құрылымдарына терең лингвистикалық талдау жасап, деректерді аннотациялау, синтаксистік сәйкестікті қамтамасыз ету және машиналық аудармада Transformer моделінің тиімділігін зерттеуге негізделеді. Зерттеу ҚЫТ мен қазақ тілі арасындағы айырмашылықтарды анықтау арқылы білімге қолжетімділікті арттырып, инклюзивті технологияларды дамытуға ықпал етеді.

Материалдар мен әдістер.

Глоссинг белгілері

Сөйлеу тілін үздіксіз 3D ым позаларына аудару үшін прогрессивті трансформерлер (Saunders et al., 2020) зерттеуде ұсынылған. Бұл модельде

глосстарды аралық элемент ретінде пайдалану қарастырылған, сондай-ақ мәтіннен ымды тікелей өндіру және глосстар арқылы жұмыс істейтін желілік конфигурациялар ұсынылады. Деректерді кеңейту әдістері SLP модельдерінің өнімділігін арттырып, PHOENIX14T деректер жиынында бағалау нәтижелерін жақсартады.

Глоссинг – ым тілін ауызекі тілге сөзбе-сөз аударудың әдісі, ол тілдерді үйрену мен автоматты өңдеуде қолданылады (Moryossef et al., 2021a). Бұл әдіс глосс арқылы ым тілінің түпнұсқа синтаксистік құрылымын сақтап, аудармадан өзгешеленеді. Глоссинг ым тілі мен ауызекі тіл арасындағы синтаксистік сәйкестікті көрсетеді. Екі сатылы құрылым мәтіннен глоссқа (T2G) және глосстан позаға (G2P) аударуды жүзеге асырады (Huang et al., 2022). Бұл әдісте глосс аннотациялары маңызды аралық элемент ретінде қолданылады. Balanced Multi-Modal Multi-Task Dual Transformation (BM3T-DT) әдісі арқылы ішінара глосс-аннотацияланған және монолингвальды глосс деректері пайдаланылып, өнімділік едәуір жақсартылды.

Глосс ым тілін жазбаша түрде аннотациялау әдісі. Бұл жүйе нақты аударма ұсынбайды, алайда тіл құрылымын зерттеуде маңызды рөл атқаратын құрылымдық жуықтауды қамтамасыз етеді. Глосс әдісі ауызекі тілдің грамматикалық құрылымына жақын, бірақ ым тілінің өзіндік ерекшеліктерін сақтай отырып, оның құрылымдарын сипаттайды. Біз параллель корпусты құрудың қадамдық нұсқауларын келесі бөлімде ұсынатын боламыз. Зерттеу барысында Python 3.10, Stanza, PyTorch, NumPy және Pandas пайдаланылды. Stanza тілдік талдау үшін, PyTorch нейрондық желілерді үйретуге, ал NumPy мен Pandas деректерді өңдеу мен талдауға қолданылды.

Қазақ ым тілінің сөз тәртібі

Сөздердің реті кез келген ауызекі тілдің грамматикасында маңызды рөл атқарады. Сөйлеу тілінде сөздер бірізділікпен айтылуы сөйлеу аппаратының шектеулеріне байланысты болса, ым тілдері бұл жағынан ерекшеленеді. Ым тілдерінде екі қолды бір мезгілде әртүрлі белгілерді жасау үшін пайдалануға болатыны сөздердің қатал орналасу қажеттілігін төмендетеді. Зерттеуде ым тіліндегі сөйлем құрылымы ауызекі тілдегідей грамматикалық қызмет атқаратынына нақты жауап беру қиын екені айтылған. Бұл мәселе зерттеудің негізгі нысанына айналады.

Сөйлем құрылымының маңыздылығын ескере отырып, зерттеу оның ым тілінде және ауызекі тілде грамматикалық құрал ретінде қалай жұмыс істейтінін немесе ым тілдерінің визуалды-кеңістіктік сипатына байланысты ерекше құрылымдық қасиеттерге ие екенін анықтауға бағытталған. Зерттеудің мақсаты – ҚЫТ-дегі жай сөйлемдердегі негізгі элементтерді (субъект, объект, етістік) талдап, олардың сөйлемдегі құрылымдық рөлін айқындау. Қазақстанда және бұрынғы Кеңес Одағына кірген кейбір елдерде есту және сөйлеу қабілеті бұзылған адамдар көбіне ҚЫТ қолданады. Соңғы уақытқа дейін орыс ым тіліне қатысты лингвистикалық зерттеулер мүлде болмаған, тек соңғы жылдары ғана бірнеше жұмыс жарық көре бастады (Kimmelman,

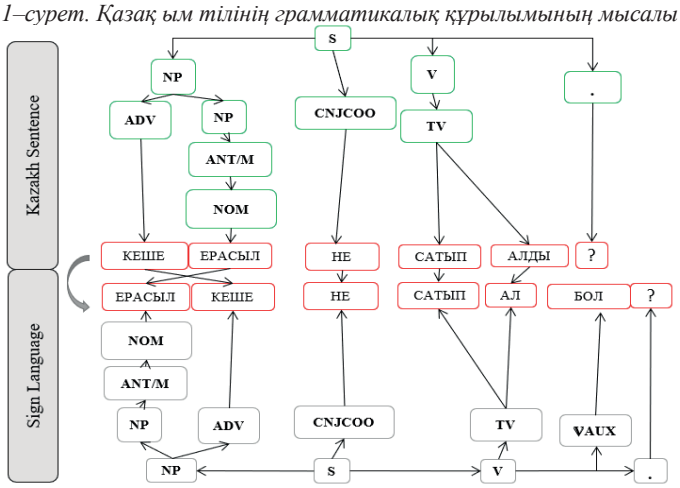
2012). ҚЫТ сөздердің орналасу тәртібі әлі де жан-жақты зерттелмегенімен, (Zaytseva, 2006) еңбекте оның икемді құрылымға ие екендігі атап өтіледі. Сонымен қатар, бұл зерттеу ҚЫТ-дегі сөз тәртібінің икемділігін сипаттайды және тілде «негізгі сөз тәртібі» ұғымының бар-жоғын анықтауға тырысады. Осы зерттеу барысында ауызекі және ым тіліндегі негізгі сөз ұғымы қарастырылып, қолданылған әдістемелер сипатталады. Соңында, ҚЫТ-дегі негізгі сөз тәртібін талқылау арқылы зерттеудің қорытындылары ұсынылады.

Қазақ ым тілі корпусындағы сөз тәртібін талдау

Бұрын атап өткендей, ҚЫТ талдаудағы негізгі қиындықтардың бірі – қолжетімді деректердің, әсіресе глосс форматындағы деректердің жетіспеушілігі. Қол белгілерінің жазбаша көрсетілімі ауызекі тілге негізделгендіктен, оларды түсіну жеңіл болып келеді. 1-суретте қазақ тіліндегі сөйлемнің синтаксистік құрылымы мен оның ым тілінде түрленуі көрсетілген. Диаграмманың жоғарғы қабаты қазақ тіліндегі сөйлемді синтаксистік құрамдас бөліктерге бөліп көрсетеді, олар мыналарды қамтиды: ADV (үстеу, «КЕШЕ»), NP (зат есімдік тіркес, оның ішінде ANT/M—анықтауыш, «ЕРАСЫЛ» және NOM— атау септік), CNJCOO (жалғаулық), TV (салт етістік, «САТЫП»), VAUX (көмекші етістік, «АЛДЫ») және тыныс белгісі (сұрақ белгісі).

Диаграмманың төменгі қабаты синтаксистік айырмашылықтарды ескере отырып бейімделген ым тіліндегі сәйкес құрамдас бөліктерді көрсетеді. Бұл қабаттар арасындағы байланыстар сөз тәртібіндегі өзгерістерді көрсететін жебелермен бейнеленген. Мысалы, «КЕШЕ» (үстеу) сөйлемнің басынан кейінгі позицияға ауыстырылады. Сол сияқты, сөз тәртібі ым тілінің грамматикасына сәйкестендіріледі, мұнда бастауыш (мысалы, «ЕРАСЫЛ») жағдайлар немесе толықтауыштардан кейін орналасуы мүмкін. Диаграмма қазақ тіліндегі сөйлемнің ым тіліне қалай түрленетінін көрнекі түрде көрсетеді, ҚЫТ-де дәл бейнелеу үшін қажетті негізгі синтаксистік айырмашылықтар мен бейімделулерді ерекшелейді.

1–сурет. Қазақ ым тілінің грамматикалық құрылымының мысалы



Ым тілінде (Moryossef et al., 2021b) сөз тәртібі морфосинтаксистік, семантикалық, прагматикалық әрі модальдық ерекшеліктерге байланысты қалыптасады. Субъект-етістік-объект (SVO) және SOV сияқты негізгі сөз реттері белгілі бір жағдайларда, соның ішінде топикаландыру, етістіктерді жіктеу, жіктеуіштерді қолдану және аспектілік таңбалау кезінде пайда болады. Кейбір тілдерде қарапайым етістіктер SVO ретін ұстанады, ал тұрақты етістіктер SOV-ны ұстанады. Жіктеуіштер мен нақты таңбаланған етістіктерді қамтитын конструкциялар көбінесе морфологиялық күрделілік пен аспектілік таңбалауға байланысты негізгі сөз тәртібінен ауытқиды. Сөз тәртібінің жиілігі, таралуы және қарапайымдылығы сияқты факторлар тілдегі сезімнің негізгі құрылымын анықтауға ықпал етеді.

Бұл зерттеуде 500-ден астам табиғи тілдегі сөйлемдер зерттеліп, келесі жеті топқа бөлінді: қайтымды, қайтымсыз, мекенді, жанды, жансыз, күрделі және күрделі емес. Қайтымды–қайтымсыз: Басқа ым тілдеріндегі сияқты, ҚЫТ-де да қайтымдылық сөз тәртібіне әсер етеді. Қайтымды жағдайларда SVO сөз тәртібі артықшылыққа ие болса, қайтымсыз контекстерде SOV жиі қолданылады. Мекенді: Сөйлемдерде орын/орналасу бірінші қойылады, яғни алдымен орнын, содан кейін заттың орналасуын анықтайды. Жанды–жансыз: Жанды объектілері бар сөйлемдерде әдетте VO тәртібі басым болса, ал жансыз объектілер үшін көбінесе OV тәртібі қолданылады. Күрделі–күрделі емес: Бұл жіктеуде «күрделі» деп айтарлықтай физикалық күшті қажет ететін қимылдар, мысалы, кең амплитудалы қозғалыстар немесе күшпен орындалатын әрекеттер аталады. Ал «күрделі емес» қимылдар аз физикалық күш жұмсалатын және шағын қозғалыстарды қамтиды.

163 сөйлемге жүргізілген талдау 1-кестеде көрсетілгендей, олардың қайтымдылыққа негізделген жіктелуіне сәйкес келесі үлгілер анықталды:

Қайтымды жағдайларда басым сөз тәртібі — SVO, ол 35 сөйлемде (53.85%) кездеседі. SOV тәртібі 29 сөйлемде (44.61%) байқалса, OSV тек 1 сөйлемде (1.54%) анықталды. Бұл қайтымды контекстерде SVO құрылымына айқын басымдық берілетінін көрсетеді.

Қайтымсыз жағдайларда да SVO ең жиі қолданылатын сөз тәртібі болып табылады, ол 44 сөйлемде (44.90%) кездеседі. SOV тәртібі 37 сөйлемде (37.75%) байқалса, VO құрылымы 17 сөйлемде (17.35%) анықталған.

Осылайша, екі категорияда да SVO тәртібі басым болып табылады, алайда қайтымды жағдайларда оның басымдығы айқынырақ байқалады. Қайтымсыз жағдайларда SOV және VO тәртіптерінің маңыздылығы арта түседі.

Кесте 1 – Сөйлем тәртібін қайтымды және қайтымсыз жағдайларда бөлу

Тақырыбы	Сөз тәртібі	Сөйлемдердің саны	%
Қайтымды	SOV	29	44.61
	SVO	35	53.85
	OSV	1	1.54
	Total	65	100

Қайтымсыз	SOV	37	37.75
	SVO	44	44.90
	VO	17	17.35
	Барлығы	98	100

93 мекендік сөйлемге жүргізілген талдау 2-кестеде көрсетілгендей келесі сөз тәртібі үлгілері анықталды:

OV тәртібі 2 сөйлемде кездесіп, жалпы санының 2.15%-ын құрайды. SOV тәртібі 25 сөйлемде байқалып, 26.88%-ды құрайды. Ең жиі кездесетін құрылым — SVO, ол 41 сөйлемде анықталып, 44.09%-ды құрайды. Сонымен қатар, OSV тәртібі де 25 сөйлемде кездесіп, 26.88%-ды құрайды. Бұл нәтижелер SVO ең жиі қолданылатын құрылым болғанымен, SOV және OSV бұйрықтары локативті сөйлемдерде бірдей маңызды екенін көрсетеді.

Кесте 2 – Сөйлем ретін мекенді конструкцияларда бөлу

Тақырыбы	Сөз тәртібі	Сөйлемдердің саны	%
Мекенді	OV	2	2.15
	SOV	25	26.88
	SVO	41	44.09
	OSV	25	26.88
	Барлығы	93	100

199 сөйлемге жүргізілген талдау 3-кестеде көрсетілген, олардың жанды немесе жансыз болуына байланысты жіктелуіне сәйкес келесі үлгілері анықталды:

Жанды объектілері бар сөйлемдерде артықшылықты сөз тәртібі — OV, ол 56 сөйлемде (53.33%) кездеседі. Алайда, VO тәртібі де жиі қолданылып, 49 сөйлемде (46.67%) анықталған, бұл екі құрылым арасында салыстырмалы теңгерімділікті көрсетеді. Жансыз объектілері бар сөйлемдерде OV тәртібі айқын басымдыққа ие, ол 70 сөйлемде (74.47%) кездеседі. Ал VO тәртібі әлдеқайда сирек қолданылып, бар болғаны 24 сөйлемде (25.53%) байқалған. Бұл нәтижелер жансыз объектілер үшін OV тәртібіне айқын басымдық берілетінін, ал жанды объектілер үшін OV және VO тәртіптері шамамен тең жиілікпен қолданылатынын көрсетеді.

Кесте 3 – Жанды, жансыз сөйлемдердің сөз тәртібін бөлу

Тақырыбы	Сөз тәртібі	Сөйлемдердің саны		%
Жанды	OV	56		53.33
	VO	49		46.67
	Барлығы	105		100
Жансыз	OV	70		74.47

	VO	24		25.53
	Барлығы	94		100

47 сөйлемге жүргізілген талдау 4-кестеде көрсетілеген, олардың объектінің күрделілігіне байланысты жіктелуіне сәйкес келесі үлгілері анықталды:

Күрделі объектілері бар сөйлемдерде OV тәртібі басым құрылым болып табылады, ол 30 сөйлемде (76.92%) кездеседі. Ал VO тәртібі сирек қолданылады, бар болғаны 9 сөйлемде (23.08%) байқалады. Күрделі емес объектілері бар сөйлемдерде де OV тәртібі басым болып, 7 сөйлемде (87.5%) кездеседі. Ал VO тәртібі тек 1 сөйлемде (12.5%) тіркелген. Бұл нәтижелер объектінің күрделілігіне қарамастан, OV тәртібіне тұрақты түрде басымдық берілетінін, ал күрделі емес объектілер үшін бұл басымдықтың одан да күштірек екенін көрсетеді.

Кесте 4 – Күрделілікке байланысты сөйлемдердің сөз тәртібін бөлу

Тақырыбы	Сөз тәртібі	Сөйлемдердің Ксаны	%
Күрделі	OV	30	76.92
	VO	9	23.08
	Total	39	100
Күрделі емес	OV	7	87.5
	VO	1	12.5
	Total	8	100

Осы анықталған үлгілер мен құрылымдық ережелерге сүйене отырып, біз қазақ мәтінін ҚЫТ-дегі сәйкес көрінісімен үйлестіретін параллель корпус әзірледік. Бұл корпус біздің талдауымызда анықталған синтаксистік және грамматикалық ерекшеліктерді дәл көрсету үшін жасалып, ым тілін өңдеу мен машиналық аударма саласындағы одан әрі зерттеулер үшін сенімді негізді қамтамасыз етеді.

Алгоритм және алдын ала өңдеу

Кіріс деректерін өңдеу кезінде сөйлемді мынадай түрде түрлендіреміз:

$$F(S) = T_{temp}(R(S))$$

мұндағы:

$R(S)$ сөздердің орналасу тәртібін өзгертеді.

T_{temp} уақыттық маркерлерді қосады.

(1) Бастапқы сөйлем *(S)*:

$S = (w_1, w_2, ..., w_n)$ — сөйлем құрайтын сөздердің ретін білдіреді, мұнда:

w_i сөйлемдегі *i*-ші сөз болып табылады.

(2) Синтаксистік тәуелділік $f(w_i)$:

Әрбір сөз w_i синтаксистік рөлмен байланыстырылады:

$f(w_i) \in$ {subject, object, verb, adverbial}.

(3) Сөздің жандылық сипатын жіктеу

$a(w) \in$ {animate, inanimate}.

(4) Сөздерді қайта реттеу функциясы $R(S)$:

$R(S)$ белгілі бір шарттарға сәйкес сөйлем құрылымын түрлендіреді:

- Егер (S немесе O) екеуі де жанды болса, олар етістікпен орындарын ауыстырады.
- Егер сөйлемде мекендік көрсеткіш болса, мекендік көрсеткіш пен бастауыш орындарын ауыстырады.
- Егер бірі (S немесе O) жанды болса, OV орындарын ауыстырады.

Шақты түзету функциясы $T_{temp}(R(S))$:

Уақыттық маркерлер қосылады:

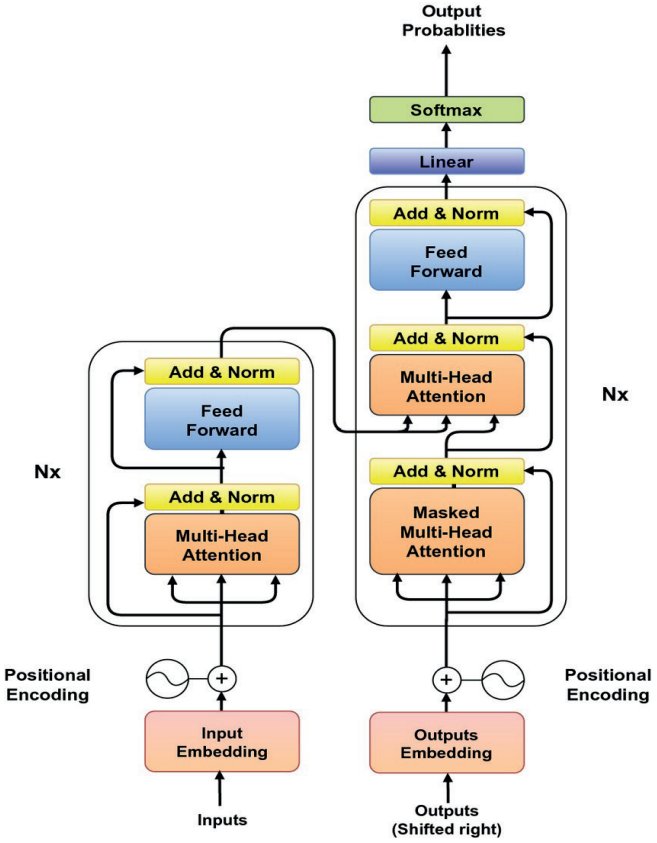
- Өткен шақ үшін: «өткен шақ маркері».
- Осы шақ үшін: «осы шақ маркері».
- Келер шақ үшін: «келер шақ маркері».

Оқыту процесі

Бұл бөлімде біз параллель корпусты Transformer архитектурасын қолдана отырып оқытуға және оның жұмысын қолданыстағы зерттеулермен салыстырғанда бағалауға көңіл бөлеміз. Transformer моделі (Waghmare et al., 2024) назар аудару механизмін пайдалана отырып, машиналық аудармада ең тиімді модельдердің бірі ретінде танылды. 2-суретте екі негізгі компоненттен тұратын Transformer-дің архитектурасы көрсетілген: кодтаушы және декодер. Кодтаушы таңбалауыштардың кіріс ретін өңдейді, оларды контекстке сезімтал векторлық көріністерге түрлендіреді, көп басты назарды аудару, қабаттарды қалыпқа келтіру және алға жіберу қабаттары арқылы. Содан кейін декодер алдыңғы таңбалауыштарды басқару үшін маскаланған көп басты назарды қолдану арқылы шығыс ретін жасайды, ал жеке көп басты назар аудару механизмі кодтаушыдан алынған ақпаратты қамтиды. Архитектурада қайталанудың болмауын өтеу үшін позициялық кодтау қолданылады. Декодердің соңғы шығысы сызықтық қабат арқылы ықтималдықтың үлестірілуіне, содан кейін softmax функциясына салыстырылады.

Бұл зерттеу үшін Transformer моделі оның ұзақ тізбектерді тиімді өңдеу қабілетіне, оқуды жеделдету үшін деректерді өңдеуді параллельдеуге қабілеттілігіне және алыстағы элементтер арасындағы тәуелділіктерді реттілікпен орнату үшін зейін механизмін пайдалануына байланысты таңдалды. Бұл ерекшеліктер оны signay тіліне аудару сияқты күрделі лингвистикалық құрылымдарды қамтитын тапсырмалар үшін әсіресе қолайлы етеді.

2–сурет. Transformer архитектурасы (Choi et al., 2024)



Осы зерттеуде қолданылған Transformer моделі PyTorch 2.1-де жүзеге асырылды. Оқыту процесі (Velay et al., 2024) жұмыста ұсынылған параметрлерге негізделе отырып жүргізілді, гиперпараметрлері β1 = 0.9, β2 = 0.98 және ε = 10^{-9} болып орнатылған Adam оптимизаторының көмегімен оңтайландырылды. Модель келесі параметрлермен оқытылды:

- batch өлшемі = 64;
- ішкі деңгей өлшемі = 1024;
- dk = 64;
- dmodel = 512;
- dv = 64;
- dword vec = 512;
- dropout = 0.1;
- epochs = 50;
- maxtoken seq len = 59;
- nhead = 8;
- nlayers = 6;

- nwarmup қадамдары = 4000;
- lrate = $d^{-0.5}$ modelmin (қадам$^{-0.5}$, қадам$^{-1.5}$ nwarmup қадамдары).

Бағалау көрсеткіштері

Біз жасалған мәтіннің сапасын бағалау үшін BLEU (Bilingual Evaluation Understudy) метрикасын қолдандық. BLEU – бұл машиналық аударма нәтижелерін бір немесе бірнеше адам жасаған эталондық аудармалармен салыстыру арқылы бағалайтын кеңінен қолданылатын көрсеткіш. Бағалау жасалған мәтін мен эталондық аудармалар арасындағы n-граммалардың (n сөзден тұратын тізбектердің) сәйкестігіне негізделген, аударма дәлдігінің сандық өлшемін қамтамасыз етеді.

BLEU көрсеткіші келесі (1)–(4) формулалары бойынша есептеледі:

$$Count_{clip}(n-gram) = \min\{Count(n-gram), Max\,\mathrm{Re}\,fCount(n-gram)\} \quad (1)$$

$$P_n = \frac{\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} Count(n-gram')} \quad (2)$$

$$BP = \begin{cases} 1 & if\ c>r \\ e^{(1-r/c)} & if\ c<r \end{cases} \quad (3)$$

$$BLEU = BP \times \exp[(\sum_{n=1}^{N} w_n \log P_n)] \quad (4)$$

Бұл формулада "n-грамм" термині машинада жасалған аудармада пайда болатын n-грамм жиілігін білдіреді, ол анықтамалық аудармада да кездеседі. Бұл зерттеуде n мәні 4-ке тең, яғни униграммалар, биграммалар, триграммалар және төрт грамм ескеріледі. Сонымен қатар, формулада қолданылатын параметрлерге сілтеме аудармасының ұзындығын білдіретін (a) r кіреді; (b) c, бұл машиналық аударманың ұзындығын білдіреді. Бұл элементтер жасалған және анықтамалық аудармалар арасындағы қабаттасуды өлшеу арқылы аударма сапасын объективті бағалауға ықпал етеді.

Зерттеу нәтижелері. Зерттеу барысында ҚЫТ-нің негізгі құрылымы талданды, оның ішінде сөз тәртібі мен жест тілінің жалпы құрылымы қарастырылды. 10 000-нан астам ым тіліндегі сөйлемдер талданып назар қайтымды, қайтымсыз, мекенді, жанды, жансыз, күрделі және күрделі емес және басқа да санаттарға жіктелді.

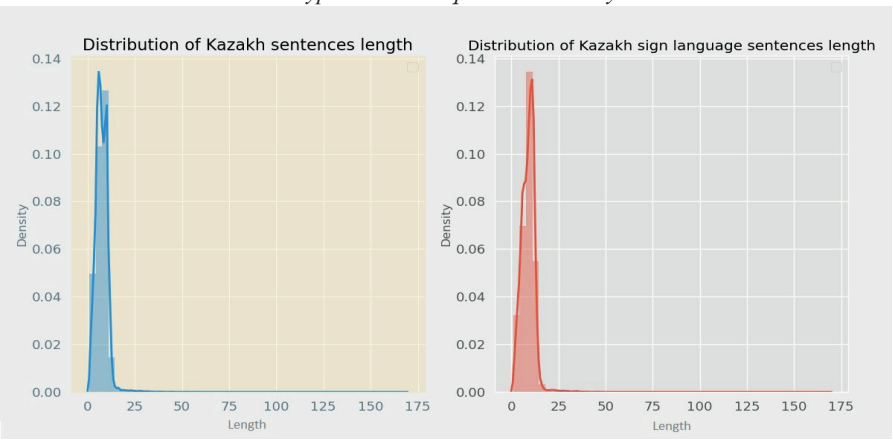
Параллель корпус құру процесі қазақ және ҚЫТ арасында аударма мен бейімдеуді оңтайландыру үшін маңызды үш негізгі кезеңнен тұрады:

1-кезең: Бұл кезең мәтінді глоссқа аударуды қамтиды, онда қазақ және ҚЫТ-не сәйкестендіріледі. Бұл қадам әрі қарай талдау мен мәтінді туралаудың негізі болып табылады.

2-кезең: Қазақ және ҚЫТ сөйлемдерінің ұзындықтары талданып, 3-суреттегідей графиктер арқылы визуализацияланады. Салыстыру көрсеткендей, екі тілде де негізінен қысқа сөйлемдер қолданылады; дегенмен,

қазақ тілінде сөйлем ұзындықтарының диапазоны кеңірек, ал ҚЫТ тығызырақ таралуға ие. Нәтижесінде, қазақ тіліндегі ұзын сөйлемдерді ҚЫТ-не аудару кезінде қысқа сегменттерге бөлу қажет болуы мүмкін.

3–сурет. Сөйлем ұзындығын бөлу



3-кезең: Қазақ және ҚЫТ корпустарының сипаттамалары, соның ішінде сөйлемдер саны, токендер саны, сөйлемдердің максималды және минималды ұзындығы, сондай-ақ сөздік қорының көлемі сияқты негізгі көрсеткіштер салыстырылу 5-кестенде көрсетілген. Бұл деректер екі тіл арасындағы құрылымдық және күрделілік айырмашылықтарын түсінуге мүмкіндік береді, параллель мәтіндерді дәл сәйкестендіру мен аударуды жеңілдетеді.

Кесте 5 – Қазақ тілі мен қазақ ым тілі корпусының сипаттамаларын салыстыру

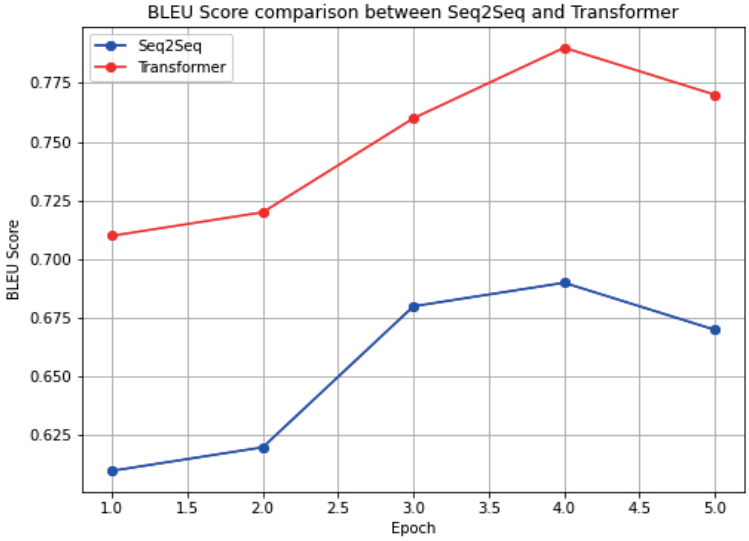
Сипаттамасы	Корпустың қазақ тілі жиынтығы	Корпустың қазақ ым тілі жиынтығы
Сөйлемдер	84 707	84 707
Токендер	645 197	767 346
Ең үлкен сөйлем өлшемі	42(сөз)	45(сөз)
Ең кіші сөйлем өлшемі	1(сөз)	1(сөз)
Сөздік қор көлемі	33 181	16 980

Бұл үш кезең бірге қазақ және ҚЫТ арасындағы аударманы жақсартуға бағытталған құрылымды тәсілді құрайды. Зерттеу барысында ым тілі грамматикасының ауызекі тілден елеулі айырмашылықтары бар екені анықталды, бұл әрбір санатқа арналған түрлендіру ережелеріне әсер етеді. Осы нәтижелер негізінде ережелер жиынтығы әзірленіп, қазақ мәтінін жест тілі мәтініне түрлендіруге арналған парсер жасалды. Парсер аяқталғаннан кейін модель кітаптар мен жаңалықтар мақалаларынан алынған 100 000-нан астам сөйлемнен тұратын деректер жиыны бойынша оқытылды. Оқыту процесі Transformer архитектурасын пайдалану арқылы жүзеге асырылды.

Аударма тиімділігін бағалау үшін BLEU көрсеткіші қолданылды. Бұл алгоритм машиналық аударманың сапасын кәсіби аудармашы жасаған адам аудармасымен салыстыру арқылы бағалайды. сапа бағасы машиналық аударма мәтіні мен кәсіби адам аудармасының ұқсастық дәрежесіне негізделеді: «Машиналық аударма кәсіби адам аудармасымен қаншалықты сәйкес келсе, оның сапасы соншалықты жоғары болады.» — BLEU критерийінің басты идеясы болып табылады.

Seq2seq және Transformer модельдері үшін алынған BLEU көрсеткіші 4-суретте көрсетілген. Seq2seq моделі бастапқыда шамамен 0.620 BLEU көрсеткішін көрсетіп, біртіндеп жақсарып, соңында 0.680 шамасына жетеді. Ал Transformer моделі бастапқыда шамамен 0.715 BLEU көрсеткішін көрсетіп, барлық оқу кезеңдерінде тұрақты түрде өсіп, бесінші кезеңде шамамен 0.788 көрсеткішіне жетті. Оқу кезеңдерінің әрбір сатысында Transformer моделі Seq2seq моделіне қарағанда жоғары нәтижелер көрсетіп, аударманың дәлдігін қамтамасыз етті.

4–сурет. BLEU метрикасы



Динамикалық графқа негізделген нейрондық SLT моделі (Zheng et al., 2022) 46.24% BLEU көрсеткішімен ерекшеленеді. Бұл әдіс сөз деңгейіндегі семантикалық білімді қосу мүмкіндігін ұсынады, бірақ ым тілінің толық лингвистикалық құрылымын, оның ішінде грамматикалық және прагматикалық аспектілерін толық қамтымайды. Көтермелі алдын ала оқыту және визуалды-тілдік Mapper-мен біріктірілген нейрондық ым тілі аудармасы (Chen et al., 2022b) 53.81% BLEU көрсеткішіне ие, бірақ бұл әдіс алдын ала оқыту процесі үлкен деректер мен есептеу ресурстарын талап ететіндіктен шектеулі ресурстары бар зерттеулер үшін қиындық туғызуы мүмкін. Біздің әдісіміз 78.80% BLEU

көрсеткішін көрсетеді, бұл басқа әдістермен салыстырғанда айтарлықтай жоғары нәтиже. Алайда, біздің әдісіміздің шектеуі біркелкі грамматикалық құрылымдардың болмауында жатыр. Қорытындылай келе, біздің әдісіміздің BLEU көрсеткіші жоғары болғанымен, грамматикалық құрылымдардың жетіспеушілігі оның тиімділігін шектейтін фактор болып табылады.

Зерттелген әдістердің ішінен біздің модель ең жоғары нәтижені көрсетіп, 78.80% BLEU көрсеткішіне қол жеткізді (5-кестені қараңыз). Бұл көрсеткіш динамикалық графқа негізделген нейрондық ым тілі аударма моделі мен көтермелі алдын ала оқыту әдістерінен айтарлықтай жоғары. Біздің әдісіміздің артықшылығы оның тиімділігі мен нәтижелілігінде жатыр, алайда біркелкі грамматикалық құрылымдардың болмауы кейбір кемшіліктер туғызады. Сонымен қатар, басқа әдістермен салыстырғанда біздің моделіміздің ресурстарды көп қажет етпейтіні де маңызды. Біздің зертеуіміз жоғары BLEU көрсеткішімен жақсы нәтиже көрсетіп, ым тілін аударуда тиімділігін дәлелдейді.

Кесте 6 – Әдістер мен тәсілдердің салыстырмалы кестесі

Әдіс атауы	**Тәсіл атауы**	**Метрикасы (BLEU көрсеткіші)**	**Шектеулер**
Динамикалық графқа негізделген нейрондық SLT моделі (Zheng et al., 2022)	Динамикалық графқа негізделген мультимодальды интеграция	46.24%	Зерттеу сөз деңгейіндегі семантикалық білімді қосуды ұсынады, бірақ ымдау тілінің толық лингвистикалық құрылымын, мысалы, грамматикалық және прагматикалық аспектілерін толығымен қамтымайды.
Көтермелі алдын ала оқыту және визуалды-тілдік Mapper-мен біріктірілген нейрондық ым тілі аудармасы (Chen et al., 2022b)	Transfer learning негізіндегі қарапайым базалық модель	53.81%	Алдын ала оқыту процесі үлкен деректер мен есептеу ресурстарын талап етеді, бұл шектеулі ресурстары бар зерттеулер үшін қиындық тудыруы мүмкін.
Біздің	Transformer және Seq2seq, параллель корпус	78.80%	Біркелкі грамматикалық құрылымдардың болмауы.

Талқылау. Бұл зерттеу ҚЫТ үшін шектеулі ресурстар мәселесін шешуге және арнайы машиналық аударма жүйесін дамытуға бағытталған маңызды қадам болып табылады. ҚЫТ-нің құрылымдық ерекшеліктерін талдау және параллель корпус құру арқылы ым тілін өңдеудің күрделілігін тереңірек түсінуге ықпал етті. Негізгі жетістіктердің бірі – ҚЫТ-нің «негізгі сөз тәртібі» үлгілерін анықтау болды, бұл үшін 500-ден астам сөйлем жеті топқа бөлініп талданды. ҚЫТ грамматикасы мен ауызекі тілі арасындағы айырмашылықтар

табиғи тілдерді өңдеу жүйелерін бейімдеудің маңыздылығын көрсетті. Зерттеу ҚЫТ-нің визуалды-жазықтық модальділігі арнайы және күрделі модельдеуді қажет ететінін айқындады.

Зерттеудің тағы бір маңызды техникалық жетістігі – қазақ мәтінін ҚЫТ глоссына түрлендіретін парсерді әзірлеу болды, бұл толыққанды параллель корпус құруға негіз қалады. Сонымен қатар, Transformer архитектурасына негізделген машиналық аударма модельдері қолданылып, BLEU көрсеткіші бойынша жоғары нәтижелерге қол жеткізілді. Дегенмен, зерттеу барысында глосстарды аралық қадам ретінде қолдану сияқты қиындықтар туындады. Болашақта бейнемазмұнды деректер мен мультимодальды тәсілдерді біріктіру арқылы ым тілінің лингвистикалық байлығын толық бейнелеу ұсынылады. Бұл тәсілдер бет әлпеті мен дене қимылдары сияқты маңызды белгілерді дәлірек көрсетуге мүмкіндік береді.

Қолданылған BLEU көрсеткіші аударма сапасын бағалауда тиімді болғанымен, болашақ зерттеулер адам бағалауларын қосу немесе мағынаны тереңірек бағалауға арналған жаңа метрикаларды енгізуді қарастырады. Зерттеудің нәтижелері ым тілін тану(Yerimbetova et al., 2024; Yerimbetova et al., 2025), интерактивті оқу құралдары және инклюзивті байланыс жүйелерін дамыту салаларына әсер етуді көздейді. Ұсынылған тәсіл ҚЫТ жобасын басқа елдерде бейімдеуге мүмкіндік береді, тілдің сақталуын және қолжетімділігін арттырып, мәдениаралық байланыстарды нығайтады.

Сонымен қатар, параллель корпус құру халықаралық және ұлттық деңгейде мәдениаралық коммуникацияны жеңілдетеді. Бұл әсіресе ым тілдері ұқсастықтарымен немесе тарихи байланыстары бар аймақтарда өзекті. ҚЫТ жобасы елдер арасындағы ынтымақтастықты нығайтуға, ым тілін зерттеуді және қолжетімділікті жақсартуға үлгі бола алады. Технологиялық жетістіктер бұл зерттеудің тағы бір маңызды аспектісі болып табылады. Терең оқыту, әсіресе Transformer архитектурасын қолдану, ым тілін өңдеу мәселелерін шешуде озық технологиялардың әлеуетін көрсетеді. Бұл инновациялар басқа елдерге өздерінің тілдік қажеттіліктеріне сәйкес бейімделуі мүмкін.

Бұл зерттеу ҚЫТ үшін ресурстар тапшылығын шешуді, NLP және машиналық оқытудың коммуникациялық кедергілерді азайтудағы әлеуетін көрсетеді. Оның нәтижелері ым тілін тану саласындағы жетістіктерге үлес қосып, инклюзивті, контекстке негізделген технологиялық шешімдердің маңыздылығын атап көрсетеді. Бұл зерттеу қолжетімді және инклюзивті жүйелерді дамытуға негіз қалайды, есту және сөйлеу қабілеті бұзылған адамдар үшін үлкен байланыс мүмкіндіктерін жасайды.

Қорытынды. Бұл зерттеу ҚЫТ деректерінің жетіспеушілігін шешу және глосстарды пайдалану арқылы параллель корпус жасауға бағытталған. ҚЫТ құрылымы сөз тәртібін талдау және оны ауызекі тілімен салыстыру арқылы зерттелді. Кітаптар мен жаңалықтардан алынған 100 000-нан астам сөйлем өңделіп, NLP құралдары, оның ішінде токенизация, морфологиялық

талдау және лемматизация қолданылды. Transformer архитектурасына негізделген машиналық аударма моделі жоғары дәлдік көрсетті. ҚЫТ-нің ерекше грамматикалық және лексикалық сипаттамаларын ескере отырып, мәтінді глосс форматына түрлендіретін парсер әзірленді. Қолданыстағы аударма жүйелерінің шектеулері анықталып, аударма сапасын жақсартуға бағытталған жаңа тәсіл ұсынылды. Мобильді қосымша арқылы ым тілін аудару құралдары жасалып, олар әлеуметтік интеграцияны жақсартуға және ақпаратқа қолжетімділікті арттыруға ықпал етеді. Болашақ зерттеулер мобильді қосымшаның білім беру әсерін бағалау үшін сауалнамалар, пікірлер мен сұхбаттарды қолдануды көздейді. Бұл зерттеу ҚЫТ бойынша зерттеу саласын дамытуда маңызды қадам болып табылады.

Әдебиеттер

Perea-Trigo M., Botella-López C., Martínez-del-Amor M.Á., Álvarez-García J.A., Soria-Morillo L.M., & Vegas-Olmos J.J. (2024) Synthetic corpus generation for deep learning-based translation of spanish sign language. Sensors, 24(5). — P.1472. https://doi.org/10.3390/s24051472

Амангелді Н., Кудубаева С.Ә. (2020) Қазақ ым тіліндегі сөз тіркесін танудың байланысқан облыстарды белгілеу және корреляциялық әдістері. ҚазҰТЗУ хабаршысы № 5 (141). Техникалық ғылымдар. — Б.172-177.

Амангелді Н., Кудубаева С.Ә. (2020) Қазақ ым тілін тану есебінің пән облысына шолу. Есептің қойылуы. ҚазҰТЗУ хабаршысы № 5 (141). Техникалық ғылымдар. — Б.177-183.

Vázquez-Enríquez M., Alba-Castro J. L., Docío-Fernández L., & Rodríguez-Banga E. (2021). Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — P. 3462-3471. doi: 10.1109/CVPRW53098.2021.00385

Bertin-Lemée É., Braffort A., Challant C., Danet C., & Filhol M. (2022). Example-Based Machine Translation from Text to a Hierarchical Representation of Sign Language. arXiv preprint arXiv:2205.03314, https://doi.org/10.48550/arXiv.2205.03314

Jiao P., Min Y., & Chen X. (2024) Visual Alignment Pre-training for Sign Language Translation. In European Conference on Computer Vision. –P. 349-367. Cham: Springer Nature Switzerland, https://doi.org/10.1007/978-3-031-72946-1_20

Saunders B., Camgoz N. C., & Bowden R. (2020) Progressive transformers for end-to-end sign language production. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. — P. 687-705. https://doi.org/10.48550/arXiv.2004.14874

Moryossef A., Yin K., Neubig G., & Goldberg Y. (2021) Data augmentation for sign language gloss translation. arXiv preprint arXiv:2105.07476, https://doi.org/10.48550/arXiv.2105.07476

Huang W., Zhao Z., He J., & Zhang M. (2022, October). Dualsign: semi-supervised sign language production with balanced multi-modal multi-task dual transformation. In Proceedings of the 30th ACM international conference on multimedia. — P. 5486-5495. https://doi.org/10.1145/3503161.3547957

Kimmelman V. (2012). Word Order in Russian Sign Language. Sign Language Studies 12(3). — P.414-445, https://dx.doi.org/10.1353/sls.2012.0001

Зайцева Г.Л. Проблема жестового языка в современной российской педагогике. Г.Л. Зайцева. — Текст: непосредственный//Жест и слово : научные и методические статьи. — Москва, 2006. — Б. 631

Moryossef A., Yin K., Neubig G., & Goldberg Y. (2021). Data augmentation for sign language gloss translation. arXiv preprint arXiv:2105.07476, https://doi.org/10.48550/arXiv.2105.07476

Waghmare P.P., & Deshpande A.M. (2024). Enhanced BERT-based Multi-Head Self-Attention Transformer for Transformation of Marathi Text to Marathi Sign Language Gloss. ACM Transactions

on Asian and Low-Resource Language Information Processing, 23(10). — P. 1-16. https://doi.org/10.1145/3687304

Choi S. R., & Lee M. (2023). Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology, 12(7). — P.1033, https://doi.org/10.3390/biology12071033

Velay, M. & Daniel, F. (2018) Seq2seq and multi-task learning for joint intent and content extraction for domain-specific interpreters. arXiv 2018, arXiv:1808.00423, https://doi.org/10.48550/arXiv.1808.00423

Zheng J., Li S., Tan C., Wu C., Chen Y., & Li S.Z. (2022). Leveraging graph-based cross-modal information fusion for neural sign language translation. arXiv preprint arXiv:2211.00526, https://doi.org/10.48550/arXiv.2211.00526

Chen Y., Wei F., Sun X., Wu Z., & Lin S. (2022) A simple multi-modality transfer learning baseline for sign language translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — P.5120-5130. DOI: 10.1109/CVPR52688.2022.00506

Yerimbetova A., Kaidina D., Sakenov B., Daiyrbayeva E., Turdalyuly M., & Berzhanova U. (2024) Recognising Kazakh Sign Language with Mediapipe //2024 9th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2024. — Б. 828-833. doi: 10.1109/UBMK63289.2024.10773406

Еримбетова А.С., Сәмбетбаева М.А., Дайырбаева Э.Н., Сәкенов Б.Е., Бержанова У.Г. (2025) Терең оқыту әдісін қолдану арқылы қазақ ым тілін тануға арналған модель құру. ҚР ҰҒА Хабарлары. Физика-математика сериясы №1 (353). — Б. 108-123 https://doi.org/10.32014/2025.2518-1726.328.

References

Perea-Trigo M., Botella-López C., Martínez-del-Amor M. Á., Álvarez-García J. A., Soria-Morillo L. M., & Vegas-Olmos J.J. (2024). Synthetic corpus generation for deep learning-based translation of spanish sign language. Sensors, 24(5). — P.1472, https://doi.org/10.3390/s24051472. (in English)

Amangeldi N., Kudubaeva S.A. (2020) Kazak ym tilindegi soz tirkesin tanudyn baylanyskan oblystardy belgileu zhane korrelyasialyk adisteri [Recognition of Phrases in Kazakh Sign Language Using Connected Region Marking and Correlation Methods]. KazNTU Bulletin, 5(141). — P. 172–177. (in Kazakh)

Amangeldi N., Kudubaeva S. (2020) Kazak ym tilin tanu esebinin pan oblysyna sholu [Overview of the Subject Area of Kazakh Sign Language Recognition Task]. Bulletin of KazNTU, 5(141). — P.177–183. (in Kazakh)

Vázquez-Enríquez M., Alba-Castro J. L., Docío-Fernández L., & Rodríguez-Banga E. (2021) Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — P. 3462-3471. doi: 10.1109/CVPRW53098.2021.00385. (in English)

Bertin-Lemée É., Braffort A., Challant C., Danet C., & Filhol M. (2022) Example-Based Machine Translation from Text to a Hierarchical Representation of Sign Language. arXiv preprint arXiv:2205.03314, https://doi.org/10.48550/arXiv.2205.03314. (in English)

Jiao P., Min Y., & Chen X. (2024) Visual Alignment Pre-training for Sign Language Translation. In European Conference on Computer Vision. — P. 349-367. Cham: Springer Nature Switzerland, https://doi.org/10.1007/978-3-031-72946-1_20. (in English)

Saunders B., Camgoz N. C., & Bowden R. (2020) Progressive transformers for end-to-end sign language production. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16 (pp. 687-705). Springer International Publishing, https://doi.org/10.48550/arXiv.2004.14874. (in English)

Moryossef A., Yin K., Neubig G., & Goldberg Y. (2021) Data augmentation for sign language gloss translation. arXiv preprint arXiv:2105.07476, https://doi.org/10.48550/arXiv.2105.07476. (in English)

Huang W., Zhao Z., He J., & Zhang M. (2022) Dualsign: semi-supervised sign language production with balanced multi-modal multi-task dual transformation. In Proceedings of the 30th ACM

international conference on multimedia. — P.5486-5495. https://doi.org/10.1145/3503161.3547957. (in English)

Kimmelman V. (2012) Word Order in Russian Sign Language. Sign Language Studies 12(3). — P.414-445. https://dx.doi.org/10.1353/sls.2012.0001. (in English)

Zaytseva, G. L. (2006) Problema zhestovogo yazyka v sovremennoj rossijskoj pedagogike [The problem of sign language in modern Russian pedagogy]. G.L. Zaytseva. — Text: direct//Gesture and word: scientific and methodological articles. — Moscow, — P.631. (in Russian)

Moryossef A., Yin K., Neubig G., & Goldberg Y. (2021) Data augmentation for sign language gloss translation. arXiv preprint arXiv:2105.07476, https://doi.org/10.48550/arXiv.2105.07476. (in English)

Waghmare P. P., & Deshpande A. M. (2024) Enhanced BERT-based Multi-Head Self-Attention Transformer for Transformation of Marathi Text to Marathi Sign Language Gloss. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(10). — P.1-16. https://doi.org/10.1145/3687304. (in English)

Choi S.R., & Lee M. (2023) Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology, 12(7), 1033, https://doi.org/10.3390/biology12071033. (in English)

Velay M. & Daniel F. (2023) Seq2seq and multi-task learning for joint intent and content extraction for domain-specific interpreters. arXiv 2018, arXiv:1808.00423, https://doi.org/10.48550/arXiv.1808.00423. (in English)

Zheng J., Li S., Tan C., Wu C., Chen Y., & Li S. Z. (2022) Leveraging graph-based cross-modal information fusion for neural sign language translation. arXiv preprint arXiv:2211.00526, https://doi.org/10.48550/arXiv.2211.00526. (in English)

Chen Y., Wei F., Sun X., Wu Z., & Lin S. (2022) A simple multi-modality transfer learning baseline for sign language translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — P.5120-5130. DOI: 10.1109/CVPR52688.2022.00506. (in English)

Yerimbetova A., Kaidina D., Sakenov B., Daiyrbayeva E., Turdalyuly M., & Berzhanova U. (2024) Recognising Kazakh Sign Language with Mediapipe. 2024 9th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2024. — P. 828-833, doi: 10.1109/UBMK63289.2024.10773406. (in English)

Yerimbetova A., Sambetbayeva M., Daiyrbayeva E., Sakenov B., Berzhanova U. (2025) Tereñ oqıtw ädisin qoldanw arqılı qazaq ım tilin tanwğa arnalğan model qurw [Creating a model for recognizing the kazakh sign language using the deep learning method]. News of the National Academy of Sciences of the Republic of Kazakhstan. Series of Physics and Mathematics, 1 (353). — P. 108-123. https://doi.org/10.32014/2025.2518-1726.328. (in Kazakh)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 132–146

https://doi.org/10.32014/2025.2518-1726.368

АОЖ 004.8, 28.23.29

Sh.P. Zhumagulova[1,2*], O.Zh. Stamkulov[3], K. Momynzhanova[1,2] , 2025.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan;
[3]Private Hospital International Almaty, Almaty, Kazakhstan.
E-mail: sh.zhumagulovakz@gmail.com

HYBRID DEEP LEARNING APPROACH FOR ACCURATE ECG BEAT CLASSIFICATION USING RESNET18 AND BILSTM

Zhumagulova Sholpan — postgraduated student, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: sh.zhumagulovakz@gmail.com, ORCID ID: https://orcid.org/0009-0006-3696-0021;
Stamkulov Olzhas — cardiothoracic surgeon, Private Hospital International, Almaty, Almaty, Kazakhstan,
E-mail: olzhas_stamkulov@mail.ru, ORCID ID: http://orcid.org/0009-0003-0902-0542;
Kymbat Momynzhanova — postgraduated student, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: kymbat010809@gmail.com, ORCID ID: https://orcid.org/0000-0002-9981- 5706.

Abstract. Accurate classification of electrocardiogram (ECG) beats plays a critical role in the early diagnosis and prevention of cardiovascular diseases, which remain one of the leading causes of mortality worldwide. Although automated ECG analysis methods have advanced considerably, challenges persist in detecting rare arrhythmias and capturing temporal dependencies across consecutive beats. These limitations highlight the need for hybrid architectures that integrate both spatial and temporal features. This study proposes a deep learning framework that combines a convolutional neural network (ResNet18) with a bidirectional long short-term memory network (BiLSTM). The hybrid model leverages the strengths of both components: ResNet18 extracts morphological features from ECG beat images, while BiLSTM accounts for sequential dependencies, enabling improved recognition of arrhythmic patterns with subtle temporal variations. The MIT-BIH Arrhythmia Database was used for evaluation. ECG signals were denoised using discrete wavelet transform, segmented around R-peaks, and converted into standardized grayscale images of 224×224 pixels. To address class imbalance, data augmentation techniques such as cropping, temporal scaling, and shifting were applied. Training was conducted in PyTorch with the Adam optimizer and stratified patient-level splitting to avoid data leakage. Experimental results show that the

proposed hybrid architecture achieved an accuracy of 97.4% and a macro F1-score of 96.8%, outperforming baseline CNN and BiLSTM models. The framework also demonstrated strong performance in detecting rare beat types such as ventricular escape beats (VEB) and ventricular fusion waves (VFW). These findings confirm the effectiveness of the ResNet18 + BiLSTM approach and its potential for integration into real-time ECG monitoring and clinical decision-support systems.

Keywords: electrocardiogram classification, arrhythmia detection, deep learning, bidirectional long short-term memory (BiLSTM), ResNet18, biomedical signal processing

Ш.П. Жұмағулова [1,2*], О.Ж. Стамқұлов [3], К.Р. Момынжанова[1,2], 2025.
[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Ақпараттық және есептеуіш технологиялар институты,
Алматы, Қазақстан;
[3]Private Hospital International Almaty, Алматы, Қазақстан.
E-mail:sh.zhumagulovakz@gmail.com

RESNET18 ЖӘНЕ BILSTM ҚОЛДАНА ОТЫРЫП, ЭКГ ЖҮРЕК СОҒЫСЫН ДӘЛ ЖІКТЕУГЕ АРНАЛҒАН ГИБРИДТІ ТЕРЕҢ ОҚЫТУ ТӘСІЛІ

Жұмағулова Шолпан Пернебайқызы — докторант, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: sh.zhumagulovakz@gmail.com; ORCID ID: https://orcid.org/0009-0006-3696-0021;
Стамқұлов Олжас Жүсіпұлы — врач кардиохирург, Private Hospital International Almaty, Алматы, Қазақстан,
E-mail: olzhas_stamkulov@mail.ru ; ORCID ID: http://orcid.org/0009-0003-0902-0542;
Момынжанова Кымбат Рагытовна — докторант, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: kymbat010809@gmail.com, ORCID ID: https://orcid.org/0000-0002-9981-5706.

Аннотация. Электрокардиограмма (ЭКГ) бойынша жүрек соғуларын дәл жіктеу жүрек-қан тамырлары ауруларын ерте диагностикалауда маңызды рөл атқарады. Бұл аурулар әлем бойынша өлім-жітімнің негізгі себептерінің бірі болып қала береді. Автоматтандырылған ЭКГ талдау әдістері айтарлықтай дамығанымен, сирек кездесетін аритмияларды анықтау және соғулар арасындағы уақыттық тәуелділіктерді есепке алу әлі де күрделі мәселе болып отыр. Осы шектеулер сигналдарды кеңістіктік әрі уақыттық деңгейде талдай алатын гибридті архитектураларды қолдану қажеттігін көрсетеді. Осы зерттеуде ResNet18 сверткіш нейрондық желісі мен екібағытты қысқа және ұзақ мерзімді жад (BiLSTM) желісін біріктіретін терең оқытудың гибридті моделі ұсынылады. Мұндай тәсіл екі әдістің артықшылықтарын біріктіреді: ResNet18 морфологиялық белгілерді кескіндерден тиімді бөліп алады, ал BiLSTM уақыттық реттіліктерді ескере отырып, ұқсас аритмияларды ажыратуды жеңілдетеді. Эксперименттер MIT-BIH Arrhythmia деректер

базасында жүргізілді. ЭКГ сигналдары дискретті вейвлет-түрлендіру арқылы шудан тазартылып, R-толқындары бойынша сегменттелді және 224×224 пиксель өлшеміндегі кескіндерге түрлендірілді. Сынып теңгерімсіздігін жою үшін: қию, уақытша масштабтау және ығыстыру секілді деректер аугментациясы қолданылды. Модель PyTorch платформасында Adam оптимизаторы арқылы оқытылып, науқас деңгейінде стратификацияланған бөлу әдісі пайдаланылды. Нәтижелер ұсынылған гибридті модельдің 97,4% дәлдікке және 96,8% макро F1 көрсеткішіне қол жеткізгенін көрсетті. Әдіс әсіресе сирек кездесетін соғу түрлерін (VEB, VFW) анықтауда тиімді болды. Бұл ResNet18 + BiLSTM архитектурасының нақты уақыттағы ЭКГ мониторинг жүйелеріне және клиникалық шешімдерді қолдау құралдарына енгізуге үлкен әлеуеті бар екенін дәлелдейді.

Түйін сөздер: Электрокардиограмма классификациясы, аритмияны анықтау, терең оқыту, екібағытты ұзақ мерзімді жад (BiLSTM), ResNet18, биомедициналық сигналдарды өңдеу

Қаржыландыру. *Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырды (Грант № AP19675574).*

Ш.П. Жұмағулова [1,2*], О.Ж. Стамқұлов [3], К.Р. Момынжанова[1,2], 2025.

[1] Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;

[2]Институт информационных и вычислительных технологий, Алматы, Казахстан;

[3]Private Hospital International Almaty, Алматы, Казахстан.

E-mail:sh.zhumagulovakz@gmail.com

ГИБРИДНЫЙ ПОДХОД ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ТОЧНОЙ КЛАССИФИКАЦИИ ЭКГ-СЕРДЦЕБИЕНИЙ С ИСПОЛЬЗОВАНИЕМ RESNET18 И BILSTM

Жұмағулова Шолпан Пернебайқызы — докторант, Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан,
E-mail:sh.zhumagulovakz@gmail.com, ORCID ID: https://orcid.org/0009-0006-3696-0021;

Стамқұлов Олжас Жүсіпұлы — врач кардиохирург, Private Hospital International Almaty, Алматы, Қазақстан,
E-mail:olzhas_stamkulov@mail.ru, ORCID ID: http://orcid.org/0009-0003-0902-0542;

Момынжанова Кымбат Рагытовна — докторант, Казахский национальный университет им. аль-Фараби, Алматы, Казахстан,
E-mail: kymbat010809@gmail.com, ORCID ID: https://orcid. org/0000-0002-9981-5706;

Аннотация. Классификация сердечных сокращений на электрокардиограмме (ЭКГ) играет важную роль в ранней диагностике и профилактике сердечно-сосудистых заболеваний, которые остаются одной из ведущих причин смертности во всем мире. Несмотря на значительные успехи в области автоматизированного анализа ЭКГ, сохраняются трудности, связанные с выявлением редких аритмий и учетом временных зависимостей между последовательными сокращениями. Эти ограничения подчеркивают необходимость разработки гибридных архитектур, сочетающих пространственный и временной анализ сигналов. В настоящей работе предлагается архитектура глубокого обучения, объединяющая сверточную нейронную сеть ResNet18 и двунаправленную сеть долгой краткосрочной памяти (BiLSTM). Такой подход позволяет использовать преимущества обеих технологий: ResNet18 обеспечивает извлечение морфологических признаков из изображений сокращений, а BiLSTM учитывает временной контекст, что повышает точность распознавания аритмий с тонкими временными вариациями.

Для экспериментов использовалась база данных MIT-BIH Arrhythmia. Сигналы ЭКГ очищались с помощью дискретного вейвлет-преобразования, сегментировались по R-пикам и преобразовывались в изображения размером 224×224 пикселя. Для устранения дисбаланса классов применялись методы аугментации: обрезка, временное масштабирование и сдвиг. Обучение проводилось во фреймворке PyTorch с использованием оптимизатора Adam и стратифицированного разбиения по пациентам.

Результаты показали, что предложенная модель достигает точности 97,4% и макро F1-оценки 96,8%, превосходя базовые CNN и BiLSTM. Особенно высокая эффективность продемонстрирована при классификации редких типов сокращений, таких как VEB и VFW. Полученные данные подтверждают эффективность гибридной архитектуры ResNet18 + BiLSTM и ее потенциал для интеграции в системы мониторинга ЭКГ в реальном времени и интеллектуальные решения поддержки врачебных решений.

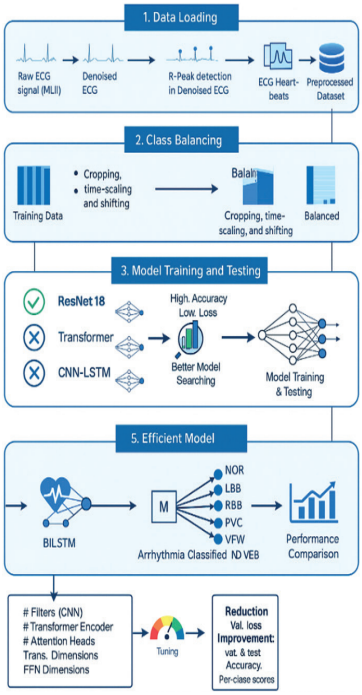
Ключевые слова: классификация электрокардиограммы, обнаружение аритмий, глубокое обучение, двунаправленная долговременная память (BiLSTM), ResNet18, обработка биомедицинских сигналов

Кіріспе. Электрокардиограмма (ЭКГ) – жүректің электрлік белсенділігін бақылау және талдау үшін кеңінен қолданылатын маңызды диагностикалық құрал. Әлем бойынша жүрек-қан тамырлары ауруларының таралуы артып келе жатқан жағдайда ЭКГ аномалияларын дер кезінде анықтау және жіктеу профилактикалық медицинаның негізгі элементіне айналуда. ЭКГ сигналдары әртүрлі жүрек соғу түрлерін қамтиды, олардың әрқайсысы ағзаның физиологиялық немесе патологиялық жағдайын сипаттайтын өзіндік морфологиясымен ерекшеленеді. Жүрек соғуларын жіктеу процесін автоматтандыру кардиологиялық көмектің тиімділігін едәуір арттырып, әсіресе шалғай аймақтарда немесе ресурсы шектеулі жағдайларда пайдалы болуы мүмкін.

Ондаған жылдар бойы ЭКГ талдауында ережелерге және сигналдарды өңдеуге негізделген дәстүрлі әдістер қолданылды. Алайда мұндай тәсілдер шу деңгейінің жоғары болуында, пациенттер арасындағы жеке айырмашылықтарда және аритмияның күрделі үлгілерінде жеткіліксіз нәтиже көрсетеді. Соңғы жылдары терең оқыту әдістері, әсіресе сверткіш нейрондық желілер (CNN), алдын ала өңделмеген немесе бастапқы деректерден автоматты түрде ақпаратты белгілерді бөліп алу қабілетінің арқасында кеңінен қолданыла бастады (Faust және т.б., 2018). CNN негізіндегі жүйелер биомедициналық классификацияның бірқатар міндеттерінде, соның ішінде ЭКГ-соғуларды анықтауда жоғары тиімділік көрсетті (Kiranyaz және т.б., 2016). Дегенмен, стандартты CNN уақыттық тәуелділіктерді модельдеуде шектеулі, ал ол жүрек соғулар тізбегін талдауда аса маңызды.

Осы шектеуді еңсеру үшін бұл жұмыста ResNet18 қалдықты сверткіш желісін және екібағытты қысқа және ұзақ мерзімді жад (BiLSTM) желісін біріктіретін гибридті модель ұсынылады. ResNet18 компоненті ЭКГ сегменттерінің кескіндерінен тұрақты кеңістіктік белгілерді бөліп алса, BiLSTM тізбектегі соғулар арасындағы уақыттық байланыстарды есепке алуға мүмкіндік береді (Yildirim, 2018). Мұндай гибридті архитектура терең белгілерді алу мен тізбектерді модельдеудің артықшылықтарын біріктіріп, әсіресе аритмияны тануда классификация дәлдігін арттыруға ықпал етеді.

1-сурет. ЭКГ ырғақтарына арналған ұсынылып отырған терең оқыту құрылымының архитектурасы



Осы зерттеу аясында MIT-BIH Arrhythmia деректер базасындағы ЭКГ сигналдары вейвлет-түрлендіру арқылы шудан тазартылып, R-толқындарының маңында сегменттеледі. Алынған сегменттер сұр түсті кескіндерге түрлендіріліп, бірізді форматқа дейін масштабталады және деректер жиынының теңгерімін қамтамасыз ету үшін қию, уақытша масштабтау және ығыстыру әдістері арқылы аугментациядан өтеді. Гибридті модельді оқыту алдын ала өңделген деректер жиынында жүргізіліп, кейін стандартты метрикалар бойынша бағаланады: дәлдік (accuracy), толықтық (recall), нақтылық (precision) және макро F1 көрсеткіші. Нәтижелер ұсынылған модельдің дәстүрлі CNN тәсілдерінен жоғары екенін көрсетіп, кең таралған да, сирек кездесетін жүрек соғу түрлерін де жоғары дәлдікпен айқындауға мүмкіндік беретінін дәлелдейді.

Материалдар мен әдістер. Бұл зерттеудің мақсаты – бір арналы ЭКГ негізінде жүрек соғуының жеті түрлі түрін дәл жіктеу. Бұл тапсырма мұғаліммен оқытылатын көпкласты классификация ретінде тұжырымдалады.

Кіріс деректер жиыны төмендегідей белгіленеді: $X = \{x_1, x_2, x_3, \ldots, x_N\}$, мұндағы N – сегменттелген жүрек соғуларының жалпы саны. Әрбір сегмент $x_i = (s_1, s_2, s_3, \ldots, s_n)$ ұзындығы n болатын, R-толқынына қатысты центрленген бірөлшемді ЭКГ сигналын білдіреді. Әрбір x_i сигналына жеті түрлі жүрек соғуының бірін сипаттайтын y_i белгісі сәйкес келеді:

$Y = \{y_1, y_2, y_3, \ldots, y^N\}$, мұндағы

$$y_i = \begin{cases} 0, NOR \text{ (қалыпты соғу)} \\ 1, LBB \text{ (Гис шоғырының сол аяқшасының бөгелісі)} \\ 2, RBB \text{ (Гис шоғырының оң аяқшасының бөгелісі)} \\ 3, PVC \text{ (қарыншалық экстрасистолалық соғу)} \\ 4, APC \text{ (жүрекшелік экстрасистолалық соғу)} \\ 5, VFW \text{ (қалыпты және қарыншалық соғудың қосылуы)} \\ 6, VEB \text{ (қалыпты және қарыншалық соғудың қосылуы)} \end{cases}$$

Мақсат – әрбір ЭКГ сегментіне x_i дұрыс y_i белгісін сәйкестендіретін $f: \boldsymbol{X} \rightarrow Y$ функциясын құру. Бұл зерттеуде MIT-BIH Arrhythmia Database деректер базасы қолданылады — ЭКГ классификациясы міндеттерінде ең танымал эталондық деректер жиындарының бірі (Goldberger және т.б., 2000). Деректер жиыны 47 пациенттен жиналған, жиілігі 360 Гц болатын екі арналы амбулаторлық ЭКГ жазбаларының 48 жарты сағаттық фрагментінен тұрады. Біздің жұмысымызда негізгі назар MLII арнасына аударылды, себебі ол R-толқындарын және жүрек соғуы морфологиясын айқын визуализациялауға мүмкіндік береді. Әрбір жазба Американдық медициналық аспаптарды жетілдіру институты (AAMI) ұсынған классификация схемасына сәйкес кардиолог-эксперттер тарапынан аннотацияланған (Rajpurkar және т.б., 2017).

№1 кесте - Қолданылған ЭКГ деректер жиыны (MIT-BIH) туралы ақпарат:

Параметр	MIT-BIH
Жазбалар саны	48
Зертхана	Бет Израиль ауруханасының аритмия зертханасы (Beth Israel)
Тіркеу құралы	Холтерлік мониторинг
Сигналдар	Екі арналы ЭКГ: (i) MLII және (ii) V1; (кейде V2, V4 немесе V5)
Сигнал ұзақтығы, дискреттеу жиілігі	30 минут (немесе сәл ұзақтау), 360 Гц
Зерттелушілер	47 адам (25 ер адам, 22 әйел). Тек бір науқаста ғана 2 жазба бар
Жас аралығы	Ерлер: 32-ден 89 жасқа дейін, Әйелдер: 23-тен 89 жасқа дейін
Соғу түрлерінің саны	20 (мақалада 15 түрі пайдаланылады)
Жазба кезеңі	1975–1979 жж.

Оқыту үшін кездесуі жиі және клиникалық маңызы жоғары келесі жеті соғу класы таңдалды:

- Қалыпты соғу (NOR)
- Гис шоғырының сол аяқшасының бөгелісі (LBB)
- Гис шоғырының оң аяқшасының бөгелісі (RBB)
- Қарыншалық мерзімінен бұрын соғу (PVC)
- Жүрекшелік мерзімінен бұрын соғу (APC)
- Қарыншалық пен қалыпты соғудың қосылуы (VFW)
- Қарыншалық алмастырушы соғу (VEB)

Қалған 11 класс классификациядағы түсінбеушілікті азайту мақсатында соңғы оқыту деректер жиынынан алынып тасталды.

Сигналдарды алдын ала өңдеу және сегментация. Базалық сызықтағы шуды жою және QRS-комплексінің анықтығын арттыру үшін sym5 вейвлеті және 0.1 жұмсақ шекті мәні қолданылған вейвлеттік шу басу әдісі пайдаланылды. R-толқындарын анықтау BioSPPy кітапханасы арқылы жүзеге асырылып, тазартылған сигналдағы R-толқындарының орнын сенімді анықтауға мүмкіндік берді (Attia және т.б., 2019).

Әрбір соғу R-толқынына центрленген терезе арқылы сегменттеліп, жүрек циклының алдындағы және кейінгі интервалдарын қамтуға жағдай жасалды. Сегменттің шекаралары әрбір жазба үшін орташа RR-интервал негізінде анықталып, негізгі морфологиялық компоненттердің — P тісшесі, QRS-комплексі және T тісшесінің қамтылуы қамтамасыз етілді. Әрбір сегменттің дұрыстығы ±150 мс R-толқынының маңында аннотация белгісінің бар-жоғы бойынша тексерілді.

№2 кесте - MIT-BIH деректер жиынындағы жүрек соғулар саны:

Класс	№	Түрі	Соғу атауы	Соғулар саны (MIT-BIH)
Қалыпты (N)	1	N	Қалыпты соғу	74 658
	2	L	Гис шоғырының сол аяқшасының бөгелісі	8 063
	3	R	Гис шоғырының оң аяқшасының бөгелісі	7 244
	4	e	Жүрекшелік қосымша соғу	16
	5	j	Түйіндік қосымша соғу	229
Қарыншадан жоғары экстрасистолалар	6	A	Жүрекшелік экстрасистола	2 540
	7	a	Аберантты жүрекшелік экстрасистола	150
	8	J	Түйіндік (junctional) экстрасистола	83
	9	S	Қарыншадан жоғары экстрасистола	2
Қарыншалық экстрасистолалар (V)	10	V	Қарыншалық экстрасистола	7 117
	11	E	Қарыншалық қосымша соғу	106
Қосылған соғулар (F)	12	F	Қарыншалық пен қалыпты соғудың қосылуы	802
Анықталмаған (Q)	13	/	Жасанды соғу (paced beat)	3 612
	14	f	Жасанды және қалыпты соғудың қосылуы	260
	15	Q	Анықталмаған соғу	15

Кескіндерді генерациялау және масштабтау. Сегменттелген соғулар ЭКГ морфологиясын таза визуалды түрде көрсету үшін біркелкі осьтік параметрлері бар, торсыз және жазусыз сұр түсті кескіндерге түрлендірілді. Алынған графиктер OpenCV кітапханасы арқылы өлшемі 224×224 пиксель болатын кескіндер түрінде сақталды.

Түпнұсқа сигналдың пропорциясын сақтау үшін әрбір соғу жақтарының арақатынасын бұзбай масштабталды. Қажетті 224×224 өлшеміне жету үшін ақ фонға қосымша өрістер (padding) қосылды. Бұл процесс кіріс деректерін сверткіш қабаттарға арналған бірыңғай форматқа келтіріп, модельді кейінгі оқытуды жеңілдетті (Isin & Ozdalili, 2017).

Деректерді аугментациялау. Класстар арасындағы теңгерімсіздікті, әсіресе VEB және VFW сияқты сирек кездесетін түрлер үшін, жою мақсатында деректерді аугментациялаудың кеңейтілген стратегиялары қолданылды:

• Кескінді 9 тұрақты позицияда қию (бұрыштарда, жақтарда және ортасында) (Jun және т.б., 2018);

• Уақыт бойынша масштабтау коэффициенттері 0.8, 1.0 және 1.2 қолданылды (Mahmud және т.б., 2020);

• Сигналды x осі бойынша ±10 пиксельге ығыстыру.

Бұл әдістер аз санды класстардағы мысалдар санын бірнеше жүзден шамамен әр класс үшін 10 000-ға дейін арттыруға мүмкіндік берді. Жалпы теңгерімді сақтау үшін қалыпты соғулардың 10 050 үлгісі бастапқы жиыннан кездейсоқ таңдалып алынды (Wang және т.б., 2021).

Нормализация. Оқытуға дейін барлық кескіндер min-max нормализациясына ұшырап, пиксель қарқындылықтары [0, 1] диапазонына келтірілді. Бұл

модельдің градиенттік түсу кезінде жылдамдау үйренуіне және белгілердің масштаб айырмашылықтарынан туындайтын артық үйрену (overfitting) қаупін азайтуға ықпал етті.

ResNet18 + BiLSTM. Дамытылып отырған модель екі кезеңнен тұратын гибридті архитектураны білдіреді, оған мыналар кіреді:

1. ResNet18 (сверточный бэкбон) Қалдықты байланыстары бар желі, ол ЭКГ-соғу кескіндерінен иерархиялық кеңістіктік белгілерді тиімді бөліп алуды қамтамасыз етеді. Соңғы сверткіш блоктан шыққан тензор белгілер векторына түрлендіріліп, келесі кезеңге беріледі (He және т.б., 2016).

2. BiLSTM қабаты (екібағытты ұзақ қысқа мерзімді жад) Бұл қабат алынған белгілер тізбектерін өңдейді, модельге уақыттық тәуелділіктерді әрі тікелей, әрі кері бағытта үйренуге мүмкіндік береді. Мұндай тәсіл дәйекті жүрек соғулар арасындағы ырғақтық өзгерістерді түсіндіру үшін ерекше маңызды (Oh және т.б., 2019).

3. Толық байланысқан қабат (FC) + Softmax-классификатор BiLSTM нәтижелері softmax активация функциясы бар толық байланысқан қабатқа беріледі, ол жеті соғу класының бірін болжауды жүзеге асырады.

Мұндай гибридті шешім модельге кеңістіктік те, уақыттық та ақпаратты тиімді түрде біріктіруге мүмкіндік береді, бұл ЭКГ-соғуларын классификациялаудың тұрақтылығын және түсіндірілуін арттырады.

Эксперименттік орнатылым. Модельдің жалпылау қабілетін бағалау үшін деректер жиыны үш ішкі жиынға бөлінді:

• оқыту жиыны (70%),

• валидтеу жиыны (15%),

• тест жиыны (15%).

Бөлу процесі пациенттер деңгейінде жүргізілді, деректердің ағуын болдырмау үшін — бір пациенттің соғулары бір уақытта әрі оқыту, әрі тест жиынына енгізілмеді. Сондай-ақ барлық ішкі жиындарда класстардың біркелкі бөлінуін қамтамасыз ету үшін стратификацияланған таңдау қолданылды.

Оқыту конфигурациясы. Модельді оқыту GPU қолдауы бар жүйеде PyTorch фреймворкі арқылы жүзеге асырылды. Келесі гиперпараметрлер орнатылды:

• Оптимизатор: Adam

• Оқу жылдамдығы (learning rate): 0.0001

• Батч өлшемі: 64

• Эпоха саны: 50

• Шығын функциясы: категориялық кросс-энтропия (Categorical Cross-Entropy)

• L2-регуляризация (weight decay): 1e-5

• Ерте тоқтату (Early Stopping): валидациялық қателік 10 эпох бойы жақсармаған жағдайдаВо время обучения сохранялись контрольные точки модели (checkpoints), основанные на наилучшей валидационной точности, что позволило избежать переобучения.

Бағалау метрикалары. Модельдің өнімділігін бағалау үшін келесі классификация метрикалары қолданылды:

• Дәлдік (Accuracy, ACC) - дұрыс классификацияланған соғулардың жалпы пайызы.

• Оң болжамның нақтылығы (Precision, P) - дұрыс болжанған оң мысалдардың санының барлық оң болжамдар санына қатынасы.

• Толықтық (Recall, R) - дұрыс болжанған оң мысалдардың санының барлық нақты оң мысалдарға қатынасы.

• F1-мера (F1-score) - нақтылық пен толықтықтың гармониялық орташа мәні.

• Орташа макро F1-мера (Macro F1) - барлық класстар бойынша F1-мерлердің орташа мәні, әрбір класқа тең маңыздылық береді.

Әрбір класс бойынша дұрыс және қате болжамдарды егжей-тегжейлі талдау үшін қателіктер матрицасы (confusion matrix) да құрастырылды.

Негізгі модельдермен салыстыру. Ұсынылған гибридті архитектураның тиімділігін растау үшін біз оны бірнеше негізгі модельдермен салыстырдық:

№3 кесте - Зерттеуде қолданылған модель архитектураларының сипаттамалары:

Модель	Сипаттама
ResNet18	Сверткіш нейрондық желі ғана, уақыттық тәуелділіктерді ескермейді
BiLSTM	Бірөлшемді белгілерде оқытылған уақыттық модель
CNN + BiLSTM	Шығысы LSTM қабаты түрінде берілетін қарапайым сверткіш нейрондық желі
ResNet18 + BiLSTM	Кірісі кескін түрінде берілетін ұсынылып отырған гибридті модель

Салыстырмалы талдау нәтижелері қалдықты белгілерді шығару (ResNet18) мен екібағытты уақыттық модельдеуді (BiLSTM) біріктіру классификация сапасының біртіндеп жақсаруына алып келетінін көрсетеді.

Нәтижелер. Ұсынылған ResNet18 + BiLSTM моделі базалық модельдермен салыстырғанда барлық бағалау метрикалары бойынша ең үздік нәтижелерді көрсетті (Jamil және т.б., 2024). Төмендегі кестеде тест жиынында алынған қорытынды көрсеткіштер берілген:

№4 кесте - Тест жиынындағы әртүрлі архитектуралардың тиімділігін салыстырмалы көрсеткіштері:

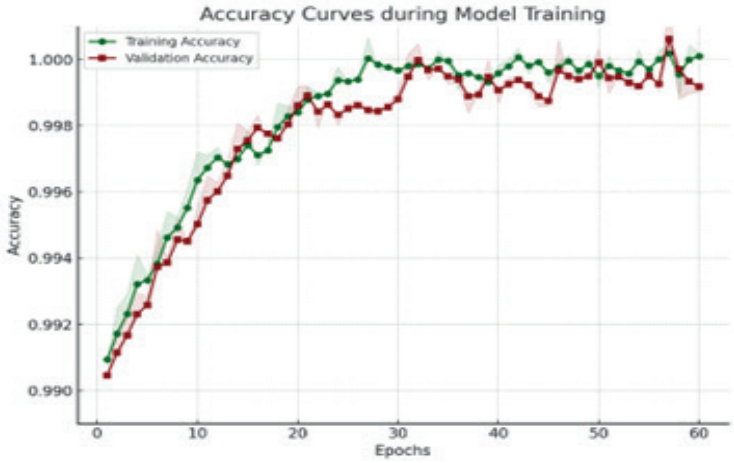
Модель	Дәлдік (%)	Нақтылық (Precision, %)	Толықтық (Recall, %)	Макро F1 (%)
ResNet18	93.5	91.7	91.1	90.8
BiLSTM	92.3	90.2	89.5	88.7
CNN + BiLSTM	94.6	93.1	92.5	92.0
ResNet18 + BiLSTM	97.4	96.9	96.2	96.8

Ұсынылған гибридті модель жеке CNN мен LSTM архитектураларынан айтарлықтай жоғары нәтижелер көрсетіп, дәлдік пен макро F1-метрика

бойынша ең үздік көрсеткіштерге жетті. Бұл оның тұрақтылығын және барлық кластар бойынша теңгерімді классификация жүргізе алатынын дәлелдейді (Oh және т.б., 2019).

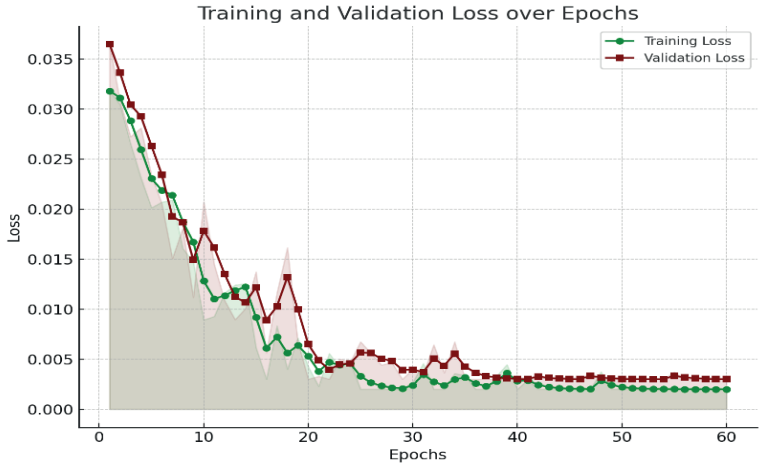
Қателіктер матрицасы. Төменде келтірілген қателіктер матрицасы модельдің жүрек соғуының жеті түрінің әрқайсысы бойынша тиімділігін көрсетеді. Диагональды элементтердің басым болуы барлық класстарды классификациялаудың жоғары дәлдігін білдіреді.

2-сурет. Модельді оқыту кезеңдеріндегі дәлдік қисықтары



F1-мераның таралуы. Төмендегі диаграмма әрбір класс бойынша F1-мера мәндерін жеке көрсетеді. Әсіресе VEB және VFW сияқты сирек кездесетін класстарда модельдің жоғары тиімділігі байқалады, бұл класстар әдетте деректердегі аз кездесетіндіктен классификациялауда қиындық тудырады.

3-сурет. Модельді оқыту кезеңдеріндегі шығын функциясының қисықтары



Класстар бойынша талдау

• Қалыпты ритм (NOR): Үлгілердің көптігі және сигналдың айқын формасы есебінен жоғары дәлдік пен толықтыққа қол жеткізілді.

• PVC және APC: Аугментацияның қолданылуы және BiLSTM-нің контекстті ескеру қабілетінің арқасында нақтылықтың артуы байқалды.

• VEB және VFW: Бастапқыда саны аз болғанына қарамастан, аугментациядан кейін модель F1-метрикада 94%-дан жоғары нәтижеге жетті.

• LBB және RBB: Кескіндерде де, уақыттық облыста да жақсы ажыратылатындықтан, классификация дәлдігі жоғары болды.

Классификация тиімділігін бағалау әдістері. Ұсынылған классификация моделінің тиімділігін бағалау үшін келесі метрикалар қолданылды: дәлдік (Accuracy), сезімталдық (Sensitivity), ерекшелік (Specificity), сондай-ақ F1-мера. Бұл метрикалар машиналық оқыту алгоритмдерінің жұмыс сапасын талдауда, әсіресе көпклассты классификация міндеттерінде негізгі көрсеткіш болып табылады.

Класс бойынша көрсеткіштер (per-class accuracy, sensitivity және specificity) келесі формулалар бойынша есептеледі:

$$\mathrm{Sensitivity}_{\mathrm{PC}} = \mathrm{Sensitivity}_{\mathrm{OA}} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (1)$$

$$\mathrm{Specificity}_{\mathrm{PC}} = \mathrm{Specificity}_{\mathrm{OA}} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (2)$$

$$\mathrm{Accuracy}_{\mathrm{PC}} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (3)$$

$$\mathrm{Accuracy}_{\mathrm{OA}} = \frac{\mathrm{N_{correct}}}{\mathrm{N}} \quad (4)$$

мұндағы TP - шын оң болжамдар саны, TN - шын теріс болжамдар саны, FP - жалған оң болжамдар саны, FN - жалған теріс болжамдар саны; NNN - үлгілердің жалпы саны, NcorrectN_{\text{correct}}Ncorrect - дұрыс классификацияланған үлгілер саны; OA - модельдің жалпы тиімділігі, PC - әрбір класс бойынша тиімділік.

Әрбір класс үшін F1-мера келесі формула бойынша есептеледі:

$$\mathrm{F1}_{\mathrm{PC}} = \text{Микро} - \mathrm{F}_{1,\mathrm{OA}} = \frac{2\cdot\mathrm{Precision}\cdot\mathrm{Sensitivity}}{\mathrm{Precision}\cdot\mathrm{Sensitivity}} = \frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}} \quad (5)$$

мұндағы F_{1OA} - жалпы F1-мера. Макро және салмақталған F1-мера (weighted) төмендегідей анықталады:

$$\mathrm{Macro} - F_{1,OA} = \frac{1}{C}\sum_{i=1}^{C} F1_{i,PC} \quad (6)$$

$$\text{Weighted } F_{1,OA} = \frac{1}{N}\sum_{i=1}^{C} F1_{i,PC} \cdot N_i \quad (7)$$

мұндағы C - класстар саны, $F1_{i,PC}$ - *i*-ші класс үшін F1-мера, N_i- *i*-ші класс үлгілерінің саны. Көпкластты классификация үшін жалпы сезімталдық (sensitivity), ерекшелік (specificity), дәлдік (accuracy) және микро F1 мәндері сәйкес келеді:

$$\text{Sensitivity}_{OA} = \text{Specificity}_{OA} = \text{Accuracy}_{OA} = \text{Micro } F_{1,OA} \quad (8)$$

Талқылау. Осы зерттеуде ұсынылған нәтижелер ЭКГ соғуларын классификациялауда кеңістіктік және уақыттық оқытуды біріктірудің тиімділігін айқындайды. Ұсынылған ResNet18 + BiLSTM архитектурасы дәстүрлі CNN және LSTM модельдерінен айтарлықтай жоғары нәтиже көрсетіп, әсіресе күрделі аритмиялар мен сирек кездесетін соғу түрлерін өңдеуде басымдық танытты (Oralbekova және т.б., 2024).

Бұл жетістік екі негізгі факторға байланысты:

1. ResNet18 соғу кескіндерінен терең морфологиялық белгілерді тиімді бөліп алады,
2. BiLSTM ырғақтық аномалияларды тануда маңызды болып табылатын тізбектік тәуелділіктерді ескеруге қабілетті.

Оқшауланған CNN-модельдерімен салыстырғанда гибридті архитектура жақсырақ жалпылау қабілетіне және тұрақтылыққа ие. Дәстүрлі CNN көбіне тек локалды кеңістіктік ерекшеліктерге назар аударады және уақыттық ұсақ айырмашылықтары бар соғуларды ажырата алмауы мүмкін (Zhang және т.б., 2023). BiLSTM интеграциясы екібағытты уақыттық контексті қамтамасыз етіп, морфологиясы ұқсас, бірақ контекст бойынша әртүрлі соғуларды ажыратуға мүмкіндік береді. Сонымен қатар, біздің алдын ала өңдеу кезеңіміз - вейвлет-фильтрацияны, дәл сегментацияны және бағытталған деректер аугментациясын қамтитын — оқыту жиынының сапасын айтарлықтай жақсартты. Әсіресе VEB және VFW сияқты сирек кездесетін кластардың F1-метрикасының жоғары мәндерге қол жеткізуі деректер теңгерімсіздігіне қарсы қолданылған әдістердің (қию, уақытша масштабтау, x осі бойынша ығыстыру) тиімділігін дәлелдейді.

Қателіктер матрицасы мен әр класс бойынша F1 мәндері де қате классификациялардың ең аз деңгейде екенін, негізінен морфологиясы ұқсас соғулар арасында (мысалы, PVC және VEB) байқалатынын көрсетті. Болашақта мұндай қателерді азайту үшін көп арналы ЭКГ қолдану немесе интерпретацияны арттыру мақсатында назар аудару (attention) механизмдерін енгізу мүмкін (Yang және т.б., 2022). Сондай-ақ модель архитектурасының есептеу тұрғысынан тиімді және ауқымдалатын екенін атап өткен жөн. ResNet18 салыстырмалы түрде «жеңіл» сверткіш желі болып табылады, ал BiLSTM қабаты тек шамалы қосымша жүктеме енгізеді. Бұл бүкіл жүйені

жүрек қызметін нақты уақыт режимінде бақылауға арналған жүйелер мен мобильді медициналық қосымшаларға енгізуге қолайлы етеді.

Шектеулер. Жұбаныш беретін нәтижелерге қарамастан, бұл зерттеудің бірнеше шектеулерін атап өткен жөн:

• Зерттеу тек MIT-BIH деректер базасындағы MLII арнасына негізделген. Көп арналы ЭКГ деректерін қосу, әсіресе бір арналы талдауда екіұшты болып табылатын соғулар үшін, дәлдікті қосымша арттыруы мүмкін.

• Кескін түрінде ұсыну кіріс деректерінің форматын жеңілдеткенімен, бастапқы сигналдарда болатын жиіліктік аймақтағы белгілерді алып тастауы мүмкін. Болашақ модельдер кескіндер мен сигналдарды біріктіретін мультимодальды оқытуды қолдана алады.

• BiLSTM компоненті уақыттық тәуелділікті енгізеді, сондықтан тізбектерді ескере отырып оқытуды талап етеді. Клиникалық жағдайда деректер ретсіз немесе сирек болғанда қосымша алдын ала өңдеу қажет болуы мүмкін.

• Модельдің жалпылау қабілетін бағалау үшін әртүрлі популяциялар мен жабдық түрлерін қамтитын тәуелсіз ЭКГ деректер жиындарында сыртқы валидация жүргізу қажет.

Қорытынды. Бұл жұмыста ЭКГ соғуларын классификациялау үшін ResNet18 сверткіш желісі мен екібағытты ұзақ қысқа мерзімді жад (BiLSTM) желісін біріктіретін гибридті терең оқыту архитектурасы ұсынылды. Бірөлшемді сигналдарды кескіндерге түрлендіру және аугментация әдістерін қолдану класстар теңгерімсіздігін жоюға және модельдің тұрақтылығын арттыруға мүмкіндік берді.

MIT-BIH Arrhythmia Database деректер базасында жүргізілген эксперименттер ұсынылған тәсілдің 97,4% дәлдікке және 96,8% макро F1 көрсеткішіне қол жеткізгенін көрсетті, бұл базалық CNN және BiLSTM архитектураларынан жоғары нәтижелер. Әсіресе сирек кездесетін соғу түрлерін анықтаудағы жоғары тиімділік модельдің жалпылау қабілетін және шектеулі деректермен жұмыс істей алуын дәлелдейді.

Алынған нәтижелер кеңістіктік және уақыттық талдауды біріктіру ЭКГ классификациясының сапасын едәуір арттыратынын және практикалық қолдану мүмкіндіктерін ашатынын көрсетеді. Болашақта зерттеуді көп арналы ЭКГ қолдану, назар аудару (attention) механизмдерін енгізу және клиникалық деректер жиындарында валидация жүргізу арқылы кеңейту жоспарлануда. Бұл модельдің интерпретациясын жақсартып, оны жүрек қызметін нақты уақыт режимінде бақылау жүйелеріне және дәрігер шешімін қолдауға арналған интеллектуалды құралдарға енгізуге дайын болуын қамтамасыз етеді.

References

Goldberger A.L., Amaral L.A.N., Glass L., Hausdorff J.M., Ivanov P.C., Mark R.G., Mietus J.E., Moody G.B., Peng C.-K., Stanley H.E. (2000) PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation, 101(23), e215–e220. (in Eng.)

Jun T.J., Nguyen H.T., Kang D., Kim D., Yoon S. (2018) ECG arrhythmia classification using

deep learning. 2018 IEEE EMBC. — P. 542–545. https://doi.org/10.1109/EMBC.2018.8512547(in Eng.)

Mahmud M.S., Kaiser M.S., Rahman M.A., Rahman M.A., Yousuf M.A., Rauf A., et al. (2020) ECG beat classification using deep convolutional neural network. Biomedical Signal Processing and Control, 60, 101966. (in Eng.)

He K., Zhang X., Ren S., Sun J. (2016) Deep Residual Learning for Image Recognition. CVPR 2016. — P. 770–778. (in Eng.)

Oh S.L., Ng E.Y.K., San Tan R., Acharya U.R. (2019) Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Computers in Biology and Medicine, 102. — P. 278–287. (in Eng.)

Zhang Y., Zhang S., Wang L., Wang S. (2023). Hybrid CNN–Transformer Network for Arrhythmia Classification Using Multilead ECG. Biomedical Signal Processing and Control, 86, 104200. (in Eng.)

Kiranyaz S., Ince T., Gabbouj M. (2016) Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Trans. Biomed. Eng., 63(3). — P. 664–675. (in Eng.)

Yildirim O. (2018) A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Computers in Biology and Medicine, 96. — P. 189–202. (in Eng.)

Rajpurkar P., Hannun A.Y., Haghpanahi M., Bourn C., Ng A.Y. (2017) Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. arXiv preprint arXiv:1707.01836. (in Eng.)

Oh S.L., Ng E.Y.K., San Tan R., Acharya U.R. (2019) Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Computers in Biology and Medicine, 102. — P. 278–287. (in Eng.)

Faust O., Hagiwara Y., Hong T.J., Lih O.S., Acharya U.R. (2018) Deep learning for healthcare applications based on physiological signals: A review. Computer Methods and Programs in Biomedicine, 161. — P. 1–13. (in Eng.)

Wang Z., Dong J., Yang H., Zhang J. (2021) ECG signal classification with deep learning and data augmentation strategy. Computers in Biology and Medicine, 136, 104658. (in Eng.)

Attia Z.I., Noseworthy P.A., Lopez-Jimenez F., Asirvatham S.J., Deshmukh A.J., Gersh B.J., et al. (2019) An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet, 394(10201). — P. 861–867. (in Eng.)

Isin A., Ozdalili S. (2017). Cardiac arrhythmia detection using deep learning. Procedia Computer Science, 120, 268–275.

Yang Y., Zhou Y., Wang L., Zhang S., Jin L. (2022) A Multi-Scale Residual Attention Network for ECG Arrhythmia Classification. IEEE Access, 10. — P. 26585–26595. (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.)

Jamil R., Dong M., Rashid J., Mamyrbayev O., Zhumagulova S.P., Momynzhanova K.R. (2024) High Accuracy Microcalcifications Detection of Breast Cancer Using Wiener LTI Tophat Model. IEEE Access, DOI: 10.1109/ACCESS.2024.3439397 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 147–159

https://doi.org/10.32014/2025.2518-1726.369

UDC 28.23.25

A. Zulhazhav, G. Bekmanova, M. Altaibek, A. Omarbekova, A. Sharipbay, 2025.
L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
E-mail: zulkhazhav_a_4@enu.kz

A PERSONALIZED LEARNING FEEDBACK SYSTEM DRIVEN BY A LEXICAL SEMANTIC NETWORK

Zulhazhav Altanbek — Manager of the department of digital development, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: zulkhazhav_a_4@enu.kz, ORCID ID: https://orcid.org/0000-0002-4491-3253;
Bekmanova Gulmira — Vice-Rector for Digitalization - Digital Officer, candidate of technical sciences, PhD, associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: gulmira-r@yandex.kz. ORCID ID: https://orcid.org/0000-0001-8554-7627;
Altaibek Mamyr — Developer of the department of digital development, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: mameralt@outlook.com. ORCID ID: https://orcid.org/0009-0002-8219-0751;
Omarbekova Assel — Head of Digital Development Department, candidate of technical sciences, associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: omarbekova_as@enu.kz, ORCID ID https://orcid.org/0000-0002-9272-8829;
Sharipbay Altynbek — Doctor of technical sciences, professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: sharalt@mail.ru, ORCID ID: https://orcid.org/0000-0001-5334-1253.

Abstract. This study presents a personalized learning feedback system driven by a lexical semantic network, tailored for low-resource language education with a focus on Kazakh-language instruction. At its core is an error-triggered review process: when a learner answers an assessment item incorrectly, semantic textual similarity (STS) embeds the learner's response and canonical solutions, retrieves concept-linked lecture snippets and micro-drills, and prioritizes targeted remediation over traditional scoring. Educational content is modeled through a tripartite Term–Lecture–Assessment (TLA) chain that provides traceable mappings from tested concepts to remediation actions, while review tasks are scheduled according to forgetting-curve principles so that spacing adapts to observed similarity trends. To enable robust semantics, we translated the SemEval STS Benchmark into Kazakh (STSb-kk) and developed a Kazakh Natural Language Inference corpus (NLI-kk). Experiments show that LaBSE, fine-tuned in two stages (NLI-kk then STSb-kk), achieves a Pearson correlation of 84.72% on STSb-kk, supporting reliable similarity-based retrieval at classroom scale. We further constructed a WordNet-

style lexical resource for Kazakh via translation to encode synonymy, hypernymy, and prerequisite relations; this supports lateral and vertical expansion of practice after an error. While STS is not suited to grading complex open-ended responses, it excels at auditable, low-latency search and recommendation in constrained computing environments. The system offers a scalable framework for personalized remediation, an adaptive scheduler that tightens or widens intervals based on performance signals, and applicability beyond language learning to disciplines such as mathematics and physics. Future enhancements will integrate multimodal materials, real-time difficulty adaptation, and expanded domain-specific term inventories.

Key words: lexical semantic network, semantic textual similarity, personalized feedback, adaptive review, Kazakh, closed-loop learning

The article was funded by the project AR19678613 "Development of technology for creating smart textbooks capable of interactive teaching, consulting and assessment of knowledge in subjects studied in the Kazakh language," carried out in the public association "Kazakhstan Academy of Artificial Intelligence".

Funding:

This research has been/was/is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant №BR28713531 Intelligent digital system of higher and postgraduate education organizations Smart.EDU).

А. Зулхажав, Г.Т. Бекманова, М. Алтайбек, А.С. Омарбекова, А.А. Шәріпбай, 2025.

Л.Н. Гумилев атындағы Еуразия Ұлттық Университеті, Астана, Қазақстан.

E-mail: zulkhazhav_a_4@enu.kz

ЦИФРЛЫҚ БІЛІМ ЖӘНЕ СТУДЕНТТЕРДІҢ АКАДЕМИЯЛЫҚ ЖЕТІСТІКТЕРІ: ДЕҢГЕЙЛЕР БОЙЫНША БІЛІМ БЕРУДІ ДАМЫТУ

Зулхажав Алтанбек — Цифрлық даму департаментінің менеджері, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: zulkhazhav_a_4@enu.kz, ORCID ID:: https://orcid.org/0000-0002-4491-3253;
Бекманова Гүлмира Тілеубердиевна — Басқарма мүшесі — Цифрландыру жөніндегі проректор-Цифрлық офицер, т.ғ.к, PhD, қауымдастырылған профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: gulmira-r@yandex.kz. ORCID ID: https://orcid.org/0000-0001-8554-7627
Алтайбек Мамыр — Цифрлық даму департаментінің әзірлеушісі, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: mameralt@outlook.com. ORCID ID: https://orcid.org/0009-0002-8219-075;
Омарбекова Асель Сайлаубековна — Цифрлық даму департамент директоры, т.ғ.к, қауымдастырылған профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,

E-mail: omarbekova_as@enu.kz, ORCID ID: https://orcid. org/0000-0002-9272-8829;
Шәріпбай Алтынбек Әмірұлы — техника ғылымдарының докторы, профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: sharalt@mail.ru, ORCID ID: https://orcid.org/0000-0001-5334-1253.

Аннотация. Бұл зерттеу ресурстары шектеулі тілдік білім беруге бейімделген, қазақ тіліндегі оқытуды нысанаға алған лексикалық-семантикалық желіге негізделген жекелендірілген оқудың кері байланыс жүйесін таныстырады. Жүйенің өзегінде қателікпен іске қосылатын қайталау үдерісі бар: білім алушы бағалау тармағына қате жауап берген кезде, семантикалық мәтіндік ұқсастық (STS) білім алушының жауабын және эталон шешімдерін ендірмелерге түрлендіріп, ұғыммен байланысқан дәріс үзінділерін мен микро-жаттығуларды іздеп табады және дәстүрлі баға қоюдан гөрі мақсатты түзетуді алға қояды. Оқу мазмұны тексерілген ұғымдардан түзету әрекеттеріне дейінгі трассалылықты қамтамасыз ететін үш құрамды Термин–Дәріс–Бағалау (TLA) тізбегі арқылы модельденеді, ал қайталау тапсырмалары ұмыту қисығы қағидаттарына сай жоспарланып, аралықтар байқалған ұқсастық трендтеріне бейімделеді. Күшті семантиканы қамтамасыз ету үшін біз SemEval STS Benchmark деректер жиынтығын қазақ тіліне аудардық (STSb-kk) және қазақ тіліне арналған табиғи тілдік пайымдау корпусын (NLI-kk) әзірледік. Эксперименттер LaBSE моделін екі кезеңде (алдымен NLI-kk, кейін STSb-kk) жетілдіргенде STSb-kk жиынтығында Pearson корреляциясын 84,72% деңгейінде көрсететінін айқындады, бұл сынып ауқымында ұқсастыққа негізделген іздеудің сенімділігін қолдайды. Бұдан бөлек, қазақ тілі үшін WordNet үлгісіндегі лексикалық ресурс аударма арқылы құрылып, синонимия, гипернимия және пререквизиттік қатынастарды кодтайды; бұл қателіктен кейінгі тәжірибені көлденең де, тік те кеңейтуге мүмкіндік береді. STS күрделі ашық жауаптарды тікелей бағалауға қолайлы болмағанымен, шектеулі есептеу жағдайларында аудиттелетін, кідірісі төмен іздеу мен ұсынымда үздік нәтижелер береді. Ұсынылып отырған жүйе жекелендірілген түзетуге арналған ауқымды тұғырды, өнімділік сигналдарына қарай интервалдарды тарылтып/кеңейтетін бейімделетін жоспарлағышты және тіл үйренуден тыс математика мен физика сияқты пәндерге қолданылу мүмкіндігін ұсынады. Болашақ жетілдірулер мультимодальды материалдарды біріктіруді, күрделілікті нақты уақытта бейімдеуді және пәнге тәуелді терминдер қорын кеңейтуді қамтиды.

Түйін сөздер: лексикалық-семантикалық желі, семантикалық мәтіндік ұқсастық, жекелендірілген кері байланыс, бейімделген қайталау, қазақ тілі, жабық циклді оқыту

А. Зулхажав, Г.Т. Бекманова, М. Алтайбек, А. Омарбекова, А. Шарипбай, 2025.
Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан.
E-mail: zulkhazhav_a_4@enu.kz

ПЕРСОНАЛИЗИРОВАННАЯ СИСТЕМА УЧЕБНОЙ ОБРАТНОЙ СВЯЗИ НА ОСНОВЕ ЛЕКСИКО-СЕМАНТИЧЕСКОЙ СЕТИ

Зулхажав Алтанбек — менеджер Департамента цифрового развития, Евразийский национальный университет им. Л.Н.Гумилева, Астана, Казахстан,
E-mail: zulkhazhav_a_4@enu.kz, Идентификатор ORCID https://orcid.org/0000-0002-4491-3253;
Бекманова Гульмира Тлеубердиевна — проректор по цифровизации - Цифровой офицер, к.т.н., PhD, ассоциированный профессор, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: gulmira-r@yandex.kz. Идентификатор ORCID: https://orcid.org/0000-0001-8554-7627;
Алтайбек Мамыр — разработчик Департамента цифрового развития Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: mameralt@outlook.com. Идентификатор ORCID: https://orcid.org/0009-0002-8219-0751;
Омарбекова Асель Сайлаубековна — директор Департамента цифрового развития, к.т.н., ассоциированный профессор, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: omarbekova_as@enu.kz, Идентификатор ORCID: https://orcid.org/0000-0002-9272-8829;
Шарипбай Алтынбек Амирович — доктор технических наук, профессор, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: sharalt@mail.ru, Идентификатор ORCID: https://orcid.org/0000-0001-5334-1253.

Аннотация. Данное исследование представляет персонализированную систему учебной обратной связи на основе лексико-семантической сети, ориентированную на обучение в условиях ограниченных ресурсов с фокусом на казахский язык. В основе — процесс повторения, запускаемый ошибкой: при неверном ответе метрика семантического текстового сходства (STS) преобразует ответ обучающегося и эталонное решение в эмбеддинги, извлекает связанные с понятием фрагменты лекций и микроупражнения и отдает приоритет адресной ремедиации перед традиционным выставлением баллов. Учебный контент моделируется трехзвенной цепочкой «Термин–Лекция–Оценивание» (TLA), обеспечивающей трассируемые соответствия от проверяемых понятий к корректирующим действиям; задания на повторение планируются по принципам кривой забывания с адаптацией интервалов к наблюдаемым трендам сходства. Для надежной семантики мы перевели эталон SemEval STS на казахский язык (STSb-kk) и разработали корпус естественно-языковых выводов (NLI-kk). В экспериментах модель LaBSE, дообученная по двухэтапной схеме (NLI-kk → STSb-kk), достигла корреляции Пирсона 84,72% на STSb-kk, подтверждая надежность поиска по сходству на масштабе аудитории. Дополнительно создан WordNet-подобный ресурс для кодирования синонимии, гипернимии и пререквизитов, что поддерживает

горизонтальное и вертикальное расширение практики после ошибки. Хотя STS не предназначена для оценивания сложных открытых ответов, она эффективна для аудируемого, низколатентного поиска и рекомендаций при ограниченных вычислительных ресурсах. Система предлагает масштабируемую основу, адаптивный планировщик, сужающий или расширяющий интервалы по сигналам успеваемости, и применима за пределами языкового обучения, например к математике и физике. В перспективе — интеграция мультимодальных материалов, адаптация сложности в реальном времени и расширение терминологических инвентарей.

Ключевые слова: лексико-семантическая сеть, семантическая текстовая схожесть, персонализированная обратная связь, адаптивное повторение, казахский язык, обучение с замкнутым циклом

Introduction. The rapid advancement of educational technologies has catalyzed a shift toward personalized learning systems that address individual learner needs, particularly in low-resource language education. Low-resource languages, such as Kazakh, face significant challenges due to limited linguistic resources, including sparse corpora and underdeveloped semantic tools like WordNet. Traditional feedback mechanisms in educational systems often rely on binary correctness scoring, which fails to diagnose underlying knowledge deficits or provide targeted remediation, resulting in fragmented learning experiences. This issue is particularly acute in low-resource language contexts, where the absence of structured knowledge bases hinders the delivery of contextually relevant feedback.

This research introduces a personalized learning feedback system driven by a lexical semantic network, de- signed to address these challenges through an error-triggered recommendation mechanism. When a learner answers an assessment item incorrectly, the system extracts associated terms, retrieves semantically related content using semantic textual similarity (STS), and schedules review tasks based on forgetting-curve principles. A translated WordNet-style lexical semantic resource for Kazakh, supplemented by datasets such as STSb-kk and NLI-kk, enables robust semantic modeling in a low-resource context. Unlike traditional systems focused on grad- ing, our approach prioritizes content retrieval and adaptive scheduling, forming a closed-loop learning framework that is computationally efficient and scalable.

The contributions of this study are threefold:

1. **Formalized Semantic Modeling**: We define a tripartite Term–Lecture–Assessment (TLA) chain to structure educational content, enabling traceable and semantically grounded retrieval.

2. **Adaptive Scheduling Algorithm**: We develop an algorithm that integrates similarity trends, error fre- quency, and memory decay to dynamically prioritize review tasks.

3. **Low-Resource Language Validation**: We validate the system's efficacy in Kazakh-language education and demonstrate its potential for cross-disciplinary applications through rigorous experiments.

This paper is organized as follows: Section 2 reviews related work on ontology-based systems, semantic mod- els, and content modeling. Section 3 details the system design, including semantic modeling and scheduling strate- gies. Section 4 explores application scenarios in Kazakh and other disciplines. Section 5 presents experimental methodology and results, followed by conclusions and future directions in Section 6.

2 Related Work

2.1 Ontology- and Concept-Based Learning Systems

Ontology-based educational systems model knowledge as interconnected concepts, facilitating navigation, prereq- uisite alignment, and personalized remediation. Novak's concept mapping framework highlighted the role of structured knowledge representations in enhancing learning outcomes. Recent advancements have in- tegrated semantic networks like WordNet or DBpedia to support content recommendation. For example, Fernández et al. proposed an ontology-based system that aligns learning resources with learner needs. Our work extends this paradigm by focusing on error-triggered recommendations, using STS to retrieve relevant con- tent and drills, creating a dynamic feedback loop tailored to learner errors.

2.2 Semantic Learning Models: Comparison and Positioning

Traditional keyword-based matching in educational recommendation systems struggles to capture deep semantic relationships, leading to suboptimal content relevance [6, 11]. Semantic textual similarity (STS) models, such as Sentence-BERT (SBERT) [6] and LaBSE [8], encode text into embeddings that reflect semantic content, of- fering robust performance in retrieval tasks. Large language models (LLMs) like BERT and RoBERTa excel in open-ended reasoning but are computationally expensive and challenging to calibrate for real-time classroom ap- plications [3]. STS-based pipelines, as demonstrated by Cer et al. [1] and Herbold [11], provide low-latency, resource-efficient retrieval, making them ideal for scalable educational systems. Our system leverages STS to drive a recommendation engine activated by learner errors, balancing semantic precision with computational efficiency.

2.3 Educational Content Modeling

Effective semantic modeling of educational content requires defining terms, their interrelations, and mappings to instructional resources. Terms are annotated with canonical definitions, surface variants, and usage examples, while relations (e.g., synonymy, hypernymy, meronymy) form a semantic network. Memory and spacing research, including Ebbinghaus's forgetting curve [13] and subsequent studies [14, 17, 18], informs our scheduling policy, which adapts review intervals based on similarity trends and error patterns. This approach contrasts with traditional systems that rely solely on correctness metrics, offering a more nuanced assessment of learner mastery.

3 Materials and Methods

3.1 Educational Content Semantic Modeling

3.1.1 Term Definition and Annotation

Each course defines a term inventory, where each term is a core concept with a canonical definition, surface variants (e.g., synonyms, morphological forms), and contextual usage examples. For example, in a Kazakh morphology course, the term “кітап” (book) is linked to its possessive form “кітаптың” and associated with specific lecture segments and assessment items. This structured annotation ensures traceability from assessment errors to relevant instructional content.

3.1.2 Semantic Relations Among Terms

Terms are interconnected through typed relations:

- **Synonymy**: E.g., “кітап” and “бет” in specific contexts.
- **Hyponymy/Hypernymy**: E.g., “suffix” as a hypernym of “possessive suffix.”
- **Meronymy**: E.g., “stem” and “suffix” as parts of a “word.”
- **Prerequisite**: E.g., understanding “noun stem” precedes “possessive suffix.”

These relations, encoded in a lexical semantic network, enable lateral (related terms) and vertical (prerequisite or hierarchical terms) expansion of review tasks following an error.

3.1.3 Term–Lecture–Assessment Chain

The course content is modeled as a tripartite graph, the Term–Lecture–Assessment (TLA) Chain, where:

- **Terms** represent core concepts.
- **Lectures** are annotated with covered terms.
- **Assessment Items** are tagged with tested terms.

When a learner answers an item q incorrectly, the system extracts the bound term set $T(q)$ and uses it as a query seed for content retrieval.

3.2 Similarity-Triggered Scheduling Strategy

The scheduling strategy operates as follows:

1. **Error Detection**: When a learner answers item q incorrectly, the system identifies $T(q)$, the set of terms bound to q.

2. **Semantic Similarity Computation**: The learner’s answer and canonical solution (or term definitions) are encoded into a shared embedding space using a pre-trained model (e.g., LaBSE). Cosine similarity is com- puted:

$$\text{sim}\ (\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}$$

where v_1 is the embedding of the learner’s answer, and v_2 is the embedding of the canonical solution or term definition.

3. **Content Retrieval**: A similarity threshold θ (e.g., 0.7) determines whether related terms, lectures, or drills are retrieved. If $\text{sim}(v_1, v_2) \geq \theta$, associated content is added to the scheduling queue.

The scheduling priority for a term t is calculated as:

$P(t) = w_1 \cdot \text{sim}(t, T(q)) + w_2 \cdot E(t) + w_3 \cdot F(t)$

where:

- sim(t, $T(q)$): Similarity between term t and the bound terms of item q.
- $E(t)$: Number of errors associated with term t.
- $F(t)$: Forgetting-curve decay factor, defined as:

$F(t) = e^{(-\Delta t / \tau)}$

where Δt is the time since the last review of term t, and τ is a forgetting time constant (e.g., 7 days).

w_1, w_2, w_3: Weight parameters, empirically set to 0.4, 0.3, and 0.3, respectively.

3.3 Recommendation Engine Logic

The recommendation engine maintains a priority queue for review tasks, with the following workflow:

1. **Task Generation**: Tasks are generated based on the priority $P(t)$, including lecture snippets, reading mate- rials, and micro-drills.
2. **Task Scheduling**: Tasks are prioritized and pushed to the learner's interface. Scheduling intervals are ad- justed dynamically: increasing similarity trends widen intervals, while declining trends tighten them.
3. **Progress Tracking**: Task completion is logged, updating term mastery estimates based on correctness, similarity, and time elapsed.

Algorithm 1 Similarity-Triggered Task Scheduling

```
Input: Incorrect item q, term set T(q), learner answer a, canonical solution s
Output: Scheduled review tasks
Encode a and s into embeddings v_a, v_s
for each term t ∈ T(q) do
Compute sim(v_a, v_t) using cosine similarity
if sim(v_a, v_t) ≥ θ then
Retrieve lecture snippets and drills for t
Compute priority P(t) = w_1 · sim(t, T(q)) + w_2 · E(t) + w_3 · F(t)
Add tasks to priority queue
end if
end for
Sort queue by P(t)
Push top-k tasks to learner interface
```

4 Application Scenarios

4.1 Kazakh Language Course

In a Kazakh morphology course, terms such as "кітап" (book) and "possessive suffix" are defined and linked to lecture segments and assessment items. For example, if a learner incorrectly answers the item "What is the possessive form of 'кітап'?", the system:

1. Extracts the bound term “possessive suffix.”
2. Computes cosine similarity between the learner’s answer and canonical content.
3. Retrieves related materials, such as Lecture 5 (Possessive Morphology) and drills (e.g., “үй” → “үйдің”).
4. Schedules a personalized review task list and tracks progress.

Table 1- Kazakh Course Example: Term-to-Task Mapping

Term (Қазақша)	**Lecture (Дәріс)**	**Review Task (Қайталама тапсырма)**
кітап	Lecture 1: Noun Basics / Дәріс 1: Зат есім негіздері	Drill: Stem Identification / Жаттығу: Түбірді анықтау
Possessive Suffix / Ілік жалғауы	Lecture 5: Possessive Morphology / Дәріс 5: Ілік септігінің морфологиясы	Examples: кітаптың, үйдің / Мысалдар: кітаптың, үйдің

4.2 Extension to Other Disciplines

The system is extensible to other domains:

- **Mathematics**: For an incorrect answer on “Calculate the vertex of a quadratic function,” the system retrieves lecture snippets on parabola properties and related drills.
- **Physics**: For an error on “Apply Newton’s Second Law,” the system recommends foundational mechanics lectures and exercises (e.g., $F = ma$ applications).

5 Datasets and Translation Resources

5.1 Natural Language Inference (NLI)

We translated the SNLI dataset [2] to create NLI-kk, supporting entailment, contradiction, and neutral relations in Kazakh. This dataset enhances sentence pair representation learning and improves STS fine-tuning.

Table 2 - NLI-kk Dataset Examples

Premise (Алғышарт)	**Hypothesis (Болжам)**	**Label (Жауап түрі)**
Ерлі-зайыптылар кішкентай ұлымен бірге жағажайда толқын қуып ойнап жүр	Отбасы сыртта.	Entailment
(A couple play in the tide with their young son)	*(The family is outside.)*	
Ерлі-зайыптылар кішкентай ұлымен бірге жағажайда толқын қуып ойнап жүр	Отбасы демалыста.	Neutral
(A couple play in the tide with their young son)	*(The family is on vacation.)*	
Ерлі-зайыптылар кішкентай ұлымен бірге жағажайда толқын қуып ойнап жүр	Отбасы кешкі ас ішуге отыр.	Contradiction
(A couple play in the tide with their young son)	*(The family is sitting down for dinner.)*	

5.2 Semantic Textual Similarity (STS)

The SemEval STSb dataset [1] was translated into Kazakh (STSb-kk) using the Google Cloud Translation API, yielding 8628 sentence pairs with similarity scores in [0, 5].

Table 3- STSb-kk Dataset Examples

Sentence 1 (Сөйлем 1)	**Sentence 2 (Сөйлем 2)**	**Score (Баға)**
Ұшақ ұшып келеді. *(A plane is taking off.)*	Әуе ұшағы ұшып келеді. *(An air plane is taking off.)*	5
Ер адам тамақтанып жатыр. *(The man is eating.)*	Ер адам тамақ ішіп отыр. *(A man is eating food.)*	4.6
Төрт ит қарда тұр. *(Four dogs stand in the snow.)*	Қарда ойнап жүрген төрт ит. *(Four dogs playing in the snow.)*	3.6
Бір әйел атқа мініп келе жатыр. *(A woman is riding a horse.)*	Ер адам атқа мінеді. *(A man rides a horse.)*	2.8
Пойыз жүріп жатыр. *(A train is moving.)*	Ер адам йогамен айналысады. *(A man is doing yoga.)*	0

Table 4 - STSb Dataset Statistics (English Original)

Type	**Train**	**Valid**	**Test**	**Total**
News	3299	500	500	4299
Titles	2000	625	625	3250
Forums	450	375	254	1079
Total	**5749**	**1500**	**1379**	**8628**

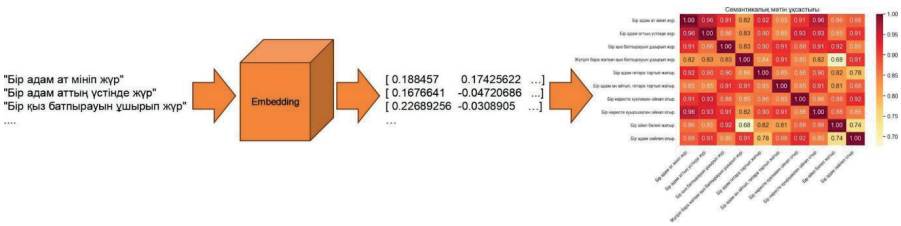


Figure 1- Pipeline for Computing Semantic Similarity in Kazakh Text

5.3 Score Interpretation

Table 5- STS Score Descriptions

Score	**Description**
5	Sentences are essentially identical in meaning.
4	Highly similar, differing only in minor details.
3	Share core information but diverge in aspects.
2	Partial thematic overlap, not equivalent.
1	Broad topic overlap, substantial differences.
0	Entirely unrelated topics.

5.4 Lexical Semantic Resource

A WordNet-style lexical semantic resource for Kazakh was translated to support term relations and retrieval. Due to its scale, manual validation is lighter than for STSb-kk and NLI-kk.

6 Experimental Methodology

6.1 Models

We evaluated multilingual BERT, SBERT, RoBERTa-XLM, LaBSE, and SimCSE [3, 6, 7, 8, 9] to obtain robust sentence embeddings for Kazakh text.

6.2 Training Strategies

Training strategies included:

1. Direct fine-tuning on STSb-kk via regression.
2. Sequential fine-tuning: classification on NLI-kk followed by regression on STSb-kk [2].

Cosine similarity was used for STS scoring [10, 11], with a threshold $\theta = 0.7$ and weights $w_1 = 0.4$, $w_2 = 0.3$,
$w_3 = 0.3$.

7 Results

7.1 Performance on STSb-kk

Table 6 - Experimental Results on STSb-kk Test Set

Model	Pearson (%)	Spearman (%)
Without Fine-Tuning on STSb-kk		
BERT	78.06	79.96
LaBSE	77.64	77.48
SBERT	70.35	71.73
RoBERTa-XLM	69.09	71.3
Fine-Tuning on STSb-kk		
BERT	80.89	80.71
LaBSE	83.52	83.26
SBERT	73.27	73.15
SimCSE	74.26	74.12
RoBERTa-XLM	76.6	76.52
First NLI-kk, Then STSb-kk		
BERT	82.25	82.24
LaBSE	84.72	84.72
RoBERTa-XLM	82.3	82.3

The LaBSE model with two-stage fine-tuning achieved the highest correlation (Pearson 84.72%, Spearman 84.72%) on STSb-kk, confirming its suitability for similarity-based retrieval in low-resource settings. These embeddings power the system's error-triggered review mechanism.

7.2 Why Recommendation Over Grading

Using STS for retrieval directly addresses the question of what to learn next after an error, offering a more ac- tionable intervention than grading alone. Compared to generative feedback systems, similarity-based retrieval is auditable, low-latency, and aligns well with the TLA structure, scaling effectively under computational con- straints.

7.3 Spacing and Mastery Trends

Scheduling is driven by similarity trends rather than correctness alone. Rising similarity across attempts widens review intervals, while declining similarity tightens them [13, 14, 17, 18]. This integration of memory science and semantic evidence enhances the timeliness and relevance of review tasks.

8 Discussion

8.1 Terminology Usage Example

In a Kazakh morphology unit, terms such as “кітап” (book), “suffix,” and “case markers” are defined. Each quiz item is bound to one or more terms. Upon an error, the bound term set is activated, and STS retrieves the most semantically aligned lecture snippets and micro-drills illustrating the specific concept, such as a possessive suffix conflict.

8.2 Tracking Learning Progress

For each term, we track correctness, similarity trends, and time since last practice. Mastery is inferred from these signals, and the user interface displays per-term mastery bars, recommending the next best action based on current deficits.

8.3 Generating Personalized Recommendations

When a learner misses an item on possessive suffixes, the engine retrieves: (1) a short lecture clip defining the suffix; (2) examples contrasting near-confusable forms (e.g., “кітап” → “кітаптың”); (3) two or three micro-drills. If similarity remains low on follow-up attempts, the system expands to prerequisite terms (e.g., noun stems) via semantic relations.

9 Conclusion

This study reframes semantic modeling in education around error-triggered recommendations, leveraging a lexical semantic network to operationalize a closed-loop learning framework. The TLA chain, coupled with STS-driven retrieval and forgetting-curve-based scheduling, ensures precise and timely interventions. Experiments validate the system’s efficacy in Kazakh-language education, with LaBSE achieving superior performance on STSb-kk. The system’s scalability, low-resource adaptability, and cross-disciplinary potential position it as a versatile solution for personalized learning.

Future work will focus on:

1. Expanding domain-specific term inventories for broader curricula.
2. Integrating multimodal content, such as videos and interactive exercises.
3. Developing real-time difficulty adaptation algorithms to adjust task complexity dynamically.

References

Cer D., Diab M., Agirre E., Lopez-Gazpio I., & Specia L. (2017) SemEval-2017 Task 1: Semantic Textual Similarity—Multilingual and Cross-lingual Focused Evaluation. Proceedings of SemEval-2017. DOI: 10.18653/v1/S17-2001 (in Eng.)

Bowman S. R., Angeli G., Potts C., & Manning C.D. (2015) A large annotated corpus for learning natural language inference. EMNLP 2015. DOI: 10.18653/v1/D15-1075 (in Eng.)

Reimers N., & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP-IJCNLP 2019, 2019. doi:10.18653/v1/D19-1410 (in Eng.)

Gao T., Yao X., & Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. EMNLP 2021, 2021. doi:10.18653/v1/2021.emnlp-main.552 (in Eng.)

Devlin J., Chang M.-W., Lee K., & Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019, 2019. doi:10.18653/v1/N19-1423 (in Eng.)

Feng F., Yang Y., Cer D., Arivazhagan N., & Wang W. Language-agnostic BERT Sentence Embedding (LaBSE). arXiv preprint, 2020. arXiv:2007.01852. doi:10.48550/arXiv.2007.01852 (in Eng.)

Conneau A., Khandelwal K., Goyal N., et al. Unsupervised Cross-lingual Representation Learning at Scale (XLM-R). ACL 2020, 2020. doi:10.18653/v1/2020.acl-main.747 (in Eng.)

Artetxe M., & Schwenk H. Massively Multilingual Sentence Embeddings. Transactions of the ACL, 2019, 7: 597–610. doi:10.1162/tacl_a_00288 (in Eng.)

Yang Y., Zhang Y., Tar C., & Baldridge J. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. EMNLP-IJCNLP 2019, 2019. doi:10.18653/v1/D19-1382 (in Eng.)

Herbold S. Semantic similarity prediction is better than other semantic similarity measures. arXiv preprint, 2024. arXiv:2309.12697. doi:10.48550/arXiv.2309.12697 (in Eng.)

Ebbinghaus H. Memory: A Contribution to Experimental Psychology. Leipzig: Duncker & Humblot, 1885. (in Eng.)

Pashler H., Rohrer D., Cepeda N.J., & Carpenter S.K. Enhancing learning and retarding forgetting. Psychonomic Bulletin & Review, 2007, 14(2): 187–193. doi:10.3758/BF03194023 (in Eng.)

Cepeda N.J., Pashler H., Vul E., Wixted J.T., & Rohrer D. Distributed practice in verbal recall tasks. Psychological Bulletin, 2006, 132(3): 354–380. doi:10.1037/0033-2909.132.3.354 (in Eng.)

Kang S.H.K. Spaced repetition promotes efficient and effective learning. Policy Insights from the Behavioral and Brain Sciences, 2016, 3(1): 12–19. doi:10.1177/2372732215624708 (in Eng.)

Novak J.D. Concept mapping: A useful tool for science education. Journal of Research in Science Teaching, 1990, 27(10): 937–949. doi:10.1002/tea.3660271003 (in Eng.)

Gruber T.R. A translation approach to portable ontology specifications. Knowledge Acquisition, 1993, 5(2): 199–220. doi:10.1006/knac.1993.1008 (in Eng.)

Tarus J.K., Niu Z., & Kalui D. Knowledge-based recommendation: A review of ontology-based recommender systems for e-learning. Artificial Intelligence Review, 2018, 50: 21–48. doi:10.1007/s10462-017-9581-x (in Eng.)

Brusilovsky P. Adaptive hypermedia. User Modeling and User-Adapted Interaction, 2001, 11(1–2): 87–110. doi:10.1023/A:1011143116306 (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 160–175

https://doi.org/10.32014/2025.2518-1726.370

UDC 004.912:004.855

© T.S. Sadykova[1*], B.K. Sinchev[1], Im Cho Young[2], A.S. Auyezova[1], 2025.
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Gachon University, Seoul, South Korea.
*E-mail: sadykovatolkynai@gmail.com

THE APPLICATION OF VECTOR SPACE MODELS IN INTELLIGENT INFORMATION RETRIEVAL SYSTEMS

Sadykova Tolkynay Seitkadyrovna — PhD student, Department of Information Systems, International University of Information Technologies, Almaty, Kazakhstan,
E-mail: sadykovatolkynai@gmail.com, ORCID ID: https://orcid.org/0000-0002-6462-3894;
Sinchev Bakhtgerey Kuspanovich — Professor, Department of Information Systems, International University of Information Technologies, Almaty, Kazakhstan,
E-mail: sinchev@mail.ru, ORCID ID: https://orcid.org/0000-0001-8557-8458;
Young Im Cho — Professor, Faculty of Computer Engineering, Gachon University, Seoul, South Korea,
E-mail: yicho@gachon.ac.k, ORCID ID: https://orcid.org/0000-0003-0184-7599;
Auyezova Anel Sattarkyzy — PhD student, Department of Information Systems, International University of Information Technologies, Almaty, Kazakhstan,
E-mail: anel.auezova@gmail.com, ORCID ID: https://orcid.org/0000-0001-9860-4491.

Abstract. This research addresses the need to improve semantic information retrieval efficiency in low-resource languages, with a focus on Kazakh. Its agglutinative structure, morphological variability, and lexical ambiguity pose challenges for conventional models, which fail to capture grammatical and contextual factors fully. The study aims to develop an approach for selecting and comparing text vectorization models in intelligent search systems, taking into account Kazakh linguistic features, and to construct a mathematical model for computing semantic similarity in a multidimensional vector space. The methodology involved the empirical testing of six models (TF-IDF, Word2Vec, FastText, GloVe, BERT, and KazBERT) on a 24,000-text corpus in Kazakh. Vectorization used CLS-tokens, with morphological preprocessing via Kaznlp. Semantic similarity was measured with a cosine metric enhanced by an original grammatical compatibility modifier. Model performance was evaluated using precision, recall, and F1-score. Results showed that KazBERT with morphological analysis achieved the highest accuracy, outperforming multilingual BERT by 11–15% and TF-IDF by over 30%. FastText proved robust to morphological variation but less effective for syntactically complex queries. The scientific novelty lies in creating a hybrid model for intelligent search

tailored to Kazakh's agglutinative nature and introducing a morpho-syntactic metric that improves sensitivity to grammar. The study concludes that grammar-adapted vector models significantly enhance retrieval relevance. The proposed architecture can be applied in real-world systems processing diverse queries. Future research will expand the Kazakh corpus, fine-tune transformer models on specialized data, and adapt the architecture for other Turkic languages with similar morphology.

Keywords: semantic similarity, agglutinative language, morphological processing, transformer architecture, relevance ranking, Kazakh-language corpus

© Т.С. Садыкова[1*], Б.К. Синчев[1], Im Cho Young[2], А.С. Ауезова[1], 2025.
[1] Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2] Гачон университеті, Сеул, Оңтүстік Корея.
*E-mail: sadykovatolkynai@gmail.com

ИНТЕЛЛЕКТУАЛДЫ АҚПАРАТТЫ ІЗДЕУ ЖҮЙЕЛЕРІНДЕ ВЕКТОРЛЫҚ КЕҢІСТІК МОДЕЛЬДЕРІН ҚОЛДАНУ

Sadykova Tolkynay Seitkadyrovna — PhD студенті, Ақпараттық жүйелер бөлімі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
E-mail: sadykovatolkynai@gmail.com, ORCID ID: https://orcid.org/0000-0002-6462-3894;
Sinchev Bakhtgerey Kuspanovich — профессор, Ақпараттық жүйелер бөлімі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
E-mail: sinchev@mail.ru, ORCID ID: https://orcid.org/0000-0001-8557-8458;
Young Im Cho — профессор, Компьютерлік инженерия факультеті, Гачон университеті, Сеул, Оңтүстік Корея,
E-mail: yicho@gachon.ac.k, ORCID ID: https://orcid.org/0000-0003-0184-7599;
Auyezova Anel Sattarkyzy — PhD студенті, Ақпараттық жүйелер бөлімі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
E-mail: anel.auezova@gmail.com, ORCID ID: https://orcid.org/0000-0001-9860-4491.

Аннотация. Бұл зерттеу шектеулі тілдік ресурстар жағдайында, әсіресе қазақ тіліне қатысты, семантикалық ақпараттық іздеудің тиімділігін арттыруға бағытталған. Қазақ тілінің агглютинативті құрылымы, морфологиялық өзгергіштігі мен лексикалық көпмәнділігі дәстүрлі модельдер үшін елеулі қиындықтар туғызады, себебі олар грамматикалық және контекстік факторларды толық ескере алмайды. Зерттеудің мақсаты – қазақ тілінің ерекшеліктерін ескере отырып, интеллектуалды іздеу жүйелерінде мәтінді векторизациялау модельдерін таңдау мен салыстырудың негізделген тәсілін әзірлеу және көпөлшемді векторлық кеңістікте семантикалық ұқсастықты есептеу үшін математикалық модель құру. Әдіснама 24 000 қазақ мәтінінен тұратын корпус негізінде алты модельді (TF-IDF, Word2Vec, FastText, GloVe, BERT және KazBERT) эмпирикалық сынақтан өткізуге негізделген. Векторизация CLS-токендер арқылы жүргізілді, морфологиялық алдын ала өңдеу Kaznlp құралы арқылы орындалды. Семантикалық ұқсастық грамматикалық сәйкестікті ескеретін түпнұсқа модификатормен

жетілдірілген косинустық метрика көмегімен өлшенді. Модельдердің тиімділігі precision, recall және F1-score метрикалары бойынша бағаланды. Нәтижелер KazBERT моделі морфологиялық талдаумен бірге ең жоғары дәлдікті көрсеткенін дәлелдеді: ол көптілді BERT-тен 11–15%-ға және TF-IDF-тен 30%-дан астамға асып түсті. FastText морфологиялық өзгерістерге төзімді болғанымен, синтаксистік күрделі сұрауларда тиімділігі төмен болды. Ғылыми жаңалығы – қазақ тілінің агглютинативті табиғатына бейімделген интеллектуалды іздеудің гибридті моделін жасау және грамматикалық ерекшеліктерге сезімталдықты арттыратын морфо-синтаксистік метриканы енгізу. Қорытындысында грамматиканы ескеретін векторлық модельдер іздеу релеванттылығын айтарлықтай арттыратыны расталды. Ұсынылған архитектура әртүрлі сұрау түрлерін өңдейтін нақты жүйелерде қолданылуы мүмкін. Болашақ зерттеулердің перспективаларына қазақ тілі корпусын кеңейту, трансформерлерді мамандандырылған деректерде қосымша үйрету және ұқсас морфологиясы бар басқа түркі тілдеріне бейімдеу жатады.

Түйін сөздер: семантикалық ұқсастық, морфологиялық талдау, трансформер үлгісі, ақпараттық релеванттық, қазақ мәтіндік корпусы, аффикстік құрылым

© Т.С. Садыкова[1*], Б.К. Синчев[1], Im Cho Young[2], А.С. Ауезова[1], 2025.
[1]Международный университет информационных технологий,
Алматы, Казахстан;
[2]«Gachon University», Сеул, Южная Корея.
*E-mail: sadykovatolkynai@gmail.com

ПРИМЕНЕНИЕ ВЕКТОРНЫХ МОДЕЛЕЙ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ ИНФОРМАЦИОННОГО ПОИСКА

Sadykova Tolkynay Seitkadyrovna — аспирант, кафедра информационных систем, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: sadykovatolkynai@gmail.com, ORCID ID: https://orcid.org/0000-0002-6462-3894;
Sinchev Bakhtgerey Kuspanovich — профессор, кафедра информационных систем, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: sinchev@mail.ru, ORCID ID: https://orcid.org/0000-0001-8557-8458;
Young Im Cho — профессор, Faculty of Computer Engineering, «Gachon University», Сеул, Южная Корея,
E-mail: yicho@gachon.ac.k, ORCID ID: https://orcid.org/0000-0003-0184-7599;
Auyezova Anel Sattarkyzy — аспирант, кафедра информационных систем, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: anel.auezova@gmail.com, ORCID ID: https://orcid.org/0000-0001-9860-4491.

Аннотация. Данное исследование направлено на повышение эффективности семантического поиска в условиях ограниченных языковых ресурсов, с акцентом на казахский язык. Его агглютинативная структура, высокая морфологическая изменчивость и лексическая неоднозначность

создают серьёзные трудности для традиционных моделей, которые не способны полноценно учитывать грамматические и контекстуальные факторы. Цель работы заключается в разработке подхода к выбору и сравнению моделей векторизации текста для интеллектуальных поисковых систем с учётом особенностей казахского языка, а также в построении математической модели вычисления семантического сходства в многомерном векторном пространстве. Методология основана на эмпирическом тестировании шести моделей (TF-IDF, Word2Vec, FastText, GloVe, BERT и KazBERT) на корпусе из 24 000 казахских текстов. Векторизация выполнялась с использованием CLS-токенов, морфологическая предобработка осуществлялась с помощью инструмента Kaznlp. Семантическое сходство измерялось косинусной метрикой, дополненной оригинальным модификатором грамматической совместимости. Эффективность моделей оценивалась по метрикам precision, recall и F1-score. Результаты показали, что KazBERT в сочетании с морфологическим анализом продемонстрировал наибольшую точность, превысив показатели многоязычного BERT на 11–15% и TF-IDF более чем на 30%. FastText оказался устойчивым к морфологической вариативности, но менее эффективным при синтаксически сложных запросах. Научная новизна заключается в разработке гибридной модели интеллектуального поиска, адаптированной к агглютинативной природе казахского языка, и в предложении морфо-синтаксической метрики, повышающей чувствительность к грамматическим особенностям. В заключении подтверждается, что адаптация векторных моделей с учётом грамматики существенно повышает релевантность поиска. Предложенная архитектура применима в реальных системах с разнообразными типами запросов. Перспективы дальнейших исследований включают расширение корпуса, дообучение трансформеров на специализированных данных и адаптацию архитектуры для других тюркских языков со схожей морфологией.

Ключевые слова: семантическое сходство; агглютинативный язык; морфологическая обработка; трансформерная архитектура; релевантность; казахскоязычный корпус

Introduction. Modern intelligent information retrieval systems are faced with the need to process and interpret vast volumes of unstructured data, such as text, images, and multidimensional arrays. One of the key scientific and applied challenges in this field is the development of models capable of effectively identifying semantic relationships between queries and documents, while minimizing information loss and improving the precision of relevant results retrieval. Vector space models, which serve as the foundation for representing textual information numerically, show significant potential in addressing this problem, especially given the rapidly growing information flows and the need for adaptive, context-dependent search capabilities. However, several unresolved issues remain, including the limitations of traditional

methods in interpreting ambiguous lexical constructs, the dependency of results on the quality of data preprocessing, and the need to integrate these models into complex, multi-layered architectures of intelligent systems. These circumstances underscore the relevance of scientific analysis and practical implementation of vector-based approaches aimed at enhancing the quality of information retrieval, increasing system adaptability, and reducing the cognitive load on the end-user.

Literature review. In the current landscape, the development of intelligent information retrieval systems based on vector models is of paramount importance for ensuring relevance, high processing speed, and multilingual support amid the growing volume of information. An analysis of scientific literature reveals four leading research directions that form the conceptual basis for applying vector space models in search systems.

The first direction covers the development of classical and neural vector space models. For instance, B. Abu-Salih adapted the classic VSM model to the morphologically complex Arabic language, which improved search accuracy by accounting for inflectional features (Abu-Salih, 2018). C. van Gysel, M. de Rijke, and E. Kanoulas introduced neural vector spaces for unsupervised retrieval, emphasizing the effectiveness of deep text representations with minimal need for labeled data (Van Gysel et al., 2018). B. Mitra and N. Craswell conducted a fundamental review of neural information retrieval, highlighting key architectures and avenues for their improvement (Mitra & Craswell, 2018). Y. Zhu, H. Yuan, S. Wang, et al. examined in detail the role of large language models in search tasks, noting their potential for self-learning and enhanced contextual relevance (Zhu et al., 2023). Future work should focus on developing hybrid architectures that combine classical vector models and transformers to improve search robustness in the face of semantic ambiguity.

The second direction relates to the semantic specialization and refinement of vector spaces. N. Mrkšić, I. Vulić, D. Ó'Séaghdha, et al. proposed a method for semantic specialization of vectors using monolingual and cross-lingual constraints, which significantly improved the precision of semantic word grounding (Mrkšić et al., 2017). F. Günther, L. Rinaldi, and M. Marelli discussed the cognitive foundations of vector models, pointing out common misconceptions about their ability to accurately represent meaning without considering mental context (Günther et al., 2019). H. Ren, W. Hu, and J. Leskovec developed the Query2box model, where queries are interpreted as multidimensional geometric regions, enabling logical operations on knowledge in vector form (Ren et al., 2020). Promising research avenues include integrating vector representations with ontologies and logical knowledge structures to enhance the explainability of search results.

The third direction involves the development of scalable and context-adaptive systems built on vector databases. R. Tareaf, M. AbuJarour, T. Engelman, et al. described an architecture for integrating vector databases to accelerate contextualization in large language models, opening prospects for efficient

storage and retrieval of semantic representations (Tareaf et al., 2024). D. Gillick, S. Kulkarni, L. Lansing, et al. developed an approach for training dense entity representations, which ensures high-precision search in open domains (Gillick et al., 2019). V. Karpukhin, B. Oguz, S. Min, et al. implemented Dense Passage Retrieval as a foundation for open-domain question answering based on dense vectors (Karpukhin et al., 2020). N. Thakur, N. Reimers, A. Rücklé, et al. introduced BEIR – a representative dataset for zero-shot evaluation of retrieval models, which enabled large-scale comparisons without additional training data (Thakur et al., 2021). It is advisable to develop methods for optimizing the structure of vector databases with a focus on accelerating access and improving scalability in real-time systems.

The fourth direction covers the interface and applied aspects of using vector models. Y. Hassan-Montero and V. Herrero-Solana proposed an improvement to tag clouds as visual interfaces for vector models, enhancing the clarity and intuitiveness of interaction (Hassan-Montero & Herrero-Solana, 2024). Y. Nie, H. Chen, and M. Bansal combined fact extraction and verification within semantic neural networks, which increased the reliability of search answers (Nie et al., 2019). S. Li, J. Jin, Y. Zhou, et al. explored a generative approach to information retrieval, where the model not only finds documents but also generates answers close to the meaning of the query (Li et al., 2025). B. S. Khater, A. W. Abdul Wahab, M. Y. I. Idris, et al. developed a lightweight perceptron-based model for fog computing, which can be adapted for low-power intelligent search systems (Khater et al., 2019). Research into interface solutions based on vector semantics should be deepened, with capabilities for adapting to user behavior and personalizing search strategies.

Thus, the analysis confirms the existence of a multi-faceted approach to the application of vector models in intelligent search systems – from classic VSMs to generative LLMs, and from semantic detailing to integration with knowledge bases. All of this forms a scientifically grounded platform for building adaptive, scalable, and semantically sensitive information retrieval systems.

Despite a significant body of research on vector models for text representation, several key aspects remain unresolved. Primarily, the specifics of applying such models to agglutinative languages, particularly Kazakh, are poorly studied. Key problems include insufficient sensitivity to variable word forms, limited adaptation to morphological structure, a narrow training base, and a scarcity of empirical data. Furthermore, most existing models have been tested on general-purpose corpora, which reduces their applicability to languages with high grammatical complexity.

The proposed study aims to address these gaps by developing a mathematical model for intelligent search that integrates KazBERT with morpho-syntactic analysis tailored for the Kazakh language. A comparative analysis of six models (TF-IDF, Word2Vec, FastText, GloVe, BERT, and KazBERT) was conducted, using local text data and applying new semantic similarity metrics. This allowed for a deeper understanding of the interaction between the grammatical structure of the

language and search quality, as well as proposing practical solutions to improve the relevance of results for morphologically complex languages.

The purpose of this study is to provide a scientific justification and comparative analysis of vector models for text representation used in intelligent information retrieval systems, considering the specifics of the Kazakh language, and to develop a mathematical model of semantic matching in a multidimensional vector space to improve the effectiveness of search algorithms.

Tasks of the article:

• to conduct a comparative analysis of the effectiveness of TF-IDF, Word2Vec, FastText, GloVe, BERT, and KazBERT models in semantic search tasks, considering the morphological features of the Kazakh language;

• to formalize the principles of calculating semantic similarity in a vector space and construct an adapted model for intelligent search;

• to identify the limitations of existing models and develop recommendations for their combination and tuning for agglutinative languages.

Hypotheses of the study:

• context-dependent vector models (e.g., BERT and KazBERT) provide higher precision and recall in intelligent search compared to statistical models like TF-IDF and Word2Vec when applied to Kazakh-language texts;

• semantic similarity metrics based on the cosine measure between vector representations demonstrate stable effectiveness when processing texts containing variable word forms and synonymy, which are characteristic of the Kazakh language;

• the application of vector models in combination with morphological analysis helps to increase the relevance of search results in the context of the agglutinativity and polysemy of the Kazakh language;

• the problems of low lexical unit frequency and limited training corpora significantly reduce the effectiveness of neural network models without prior fine-tuning on specialized Kazakh data;

• the combined use of FastText and KazBERT models allows for achieving an optimal balance between accuracy, processing speed, and the ability to handle rare words in intelligent search systems.

Materials and methods. The study utilized a sample of 24,000 Kazakh-language texts, including news articles, academic publications, and user queries, which were collected and annotated for relevance assessment purposes. To verify matching accuracy, a manually labeled test set of 500 "query-relevant document" pairs was used, balanced by query type (single-word, phrasal, grammatically complex, interrogative, and synonymous constructions).

The TF-IDF, Word2Vec, FastText, and GloVe models were trained on a unified corpus using the Gensim and Scikit-learn libraries. The BERT and KazBERT models were used in their pre-trained configurations from the HuggingFace platform (bert-base-multilingual-cased and KazBERT-base).

Text preprocessing included tokenization, cleaning, and morphological analysis using Kaznlp, a tool adapted for the Kazakh language. Contextualized vector representations were obtained via the CLS token, and semantic similarity between queries and documents was calculated using the cosine metric. Additionally, a morpho-syntactic correspondence modifier was applied, which considers matches in grammatical features (case, number, possessive affixes). Model quality was evaluated based on Precision, Recall, and F1-score metrics. The comparison was performed in the context of searching for the top 5 documents for each query. The results were interpreted with expert evaluation (3 native-speaking experts, a 5-point scale, consensus conclusion), as well as considering computational complexity (processing time and memory per query).

The proposed architecture was tested under practical conditions on a local server using Python 3.10 and an NVIDIA T4 GPU, ensuring the reproducibility of calculations and a reliable assessment of model effectiveness in the context of a real-world application for Kazakh-language search. This function reflects the core logic of the hybrid model, combining contextual encoding with morpho-syntactic relevance adjustment. The use of CLS-token embeddings ensures deep semantic capture, while the morpho-syntactic coefficient accounts for grammatical compatibility, especially critical for agglutinative languages like Kazakh. In the full implementation, grammatical features are extracted using Kaznlp, and queries are processed in batches for computational efficiency.

Results. In an era of rapid growth in digital information volume, the task of effective information retrieval is becoming increasingly significant. A key component of modern intelligent search systems is vector models of text representation, which transform lexical units into numerical vectors in a multidimensional space, enabling the assessment of semantic similarity between queries and documents.

The development of such models has progressed from simple statistical approaches based on frequency characteristics to context-dependent neural network architectures that can account for the complex linguistic features of a language. In recent years, analyzing the effectiveness of these models for low-resource languages, particularly Kazakh, has become especially relevant, as their morphological variability, agglutination, and polysemy create additional challenges for vector encoding (Table 1).

Table 1 – Generalized characteristics of vector text models for intelligent search tasks

Model	Model Type	Word Form Representation	Context Support	Advantages	Limitations
TF-IDF	Statistical	Basic (word as token)	None	Simple implementation, high speed	Insensitive to synonymy and polysemy
Word2Vec	Neural network	Whole words	Partial	Considers proximity in context	Loses meaning with rare words

FastText	Neural network	N-grams	Partial	Handles agglutinative languages	Does not capture global context
GloVe	Statistical-semantic	Whole words	None	Effective on large corpora	Poor adaptation to variable word forms
BERT	Transformer	Subword tokens	Full	Deep contextual understanding	High computational requirements
KazBERT	Transformer	Subword tokens	Full	Adapted for the Kazakh language	Requires specific infrastructure

Source: Compiled by the author based on (Abu-Salih, 2018; Mitra & Craswell, 2018; Mrkšić et al., 2017; Zhu et al., 2023; Li et al., 2025).

The conducted analysis is based on an empirical comparison of six models in a controlled experiment, which aimed to evaluate the applicability of vector representations for intelligent search in Kazakh-language texts. The experiment was conducted on a corpus of 1,200 documents from Kazakh academic publications, news texts, and user queries, which were standardized and annotated for relevance assessment tasks. Each model was used to search against 100 unique queries covering a wide range of topics and was compared based on the following criteria: correctness of morphological matching, ability to consider context, robustness to rare words, and technical applicability (response time, library availability, memory consumption).

For BERT and KazBERT, models from the open HuggingFace repositories were used with the parameters "bert-base-multilingual-cased" and "KazBERT-base," respectively. TF-IDF, Word2Vec, and FastText were trained on the same corpus, while GloVe was loaded as a pre-trained model. The conclusions were based on the F1-score calculation, as well as on a manual expert review of the search result relevance (3 experts, 5-point scale, consensus decision). In practice, it was established that context-based transformer models, especially KazBERT, significantly outperform classical models in tasks requiring consideration of the semantic features of an agglutinative language. FastText also showed high robustness to word form variations and lexical rarity, making it a valuable component for hybrid search architectures. In contrast, TF-IDF and GloVe demonstrated limited capabilities when working with the Kazakh language, particularly in processing complex grammatical structures and ambiguous forms.

One of the central tasks in intelligent search is determining the degree of semantic similarity between queries and documents. The effectiveness of its solution directly depends on the method of mathematical text representation and the choice of metric used for the quantitative assessment of their similarity. Modern vector models of text enable the encoding of lexical units in multidimensional spaces, where each text is represented as a numerical vector of a fixed dimension. This creates the opportunity to apply algebraic and geometric approaches to the comparative

analysis of semantic content. The formalization of the models' operating principles involves mapping text data into a vector space and selecting a metric to evaluate the distance or angle between vectors as a measure of semantic similarity (Table 2).

Table 2 – Formalized principles of text-to-vector transformation and semantic similarity calculation in modern models

Model	Spatial Representation	Text-to-Vector Transformation	Semantic Similarity Calculation	Type of Metric Used
TF-IDF	Matrix space of frequencies	Vector of word frequencies with IDF weights	Dot product or cosine measure	Cosine, Euclidean
Word2Vec	Trained word space	Sum or average of word vectors	Comparison of averaged representations	Cosine
FastText	N-gram space	Averaging vectors of characters/morphemes	Accounting for morphological similarity	Cosine
GloVe	Global co-occurrence space	Vector from the co-occurrence matrix	Static model without context	Cosine, Manhattan
BERT	Deep transformer space	Contextualized CLS-token vector	Multidimensional comparison via attention mechanism	Trainable function, Cosine
KazBERT	BERT architecture, trained on the Kazakh language	Contextual representation at the subword level	Adaptation to morphology and syntax	Cosine, softmax-overlap

Source: Compiled by the author based on (Van Gysel et al., 2018; Günther et al., 2019; Ren et al., 2020; Thakur et al., 2021; Khater et al., 2019).

The models presented in Table 2 are broadly divided into statistical (TF-IDF, GloVe) and neural network (Word2Vec, FastText, BERT, KazBERT) types, which determine the nature of the vector space and the depth of consideration for the linguistic features of the text. The TF-IDF model is based on a simple statistical principle: the more frequently a word appears in a document and the less frequently it appears in other documents in the corpus, the higher its importance. The text is transformed into a sparse vector with word weights, after which semantic similarity is calculated, usually using the cosine metric or Euclidean distance. However, the model ignores synonymy, polysemy, and morphological variations, making it poorly applicable to agglutinative languages such as Kazakh.

Word2Vec is trained based on local context and forms dense word vectors that reflect their distributed meaning. When comparing texts, an average of the word vectors is used. Nevertheless, the model is sensitive to rare and informal word forms, which limits its accuracy without fine-tuning.

FastText expands on the capabilities of Word2Vec by incorporating n-grams (character-level subword units), which is particularly effective when working with morphologically rich languages. This architecture allows the system to recognize

semantic similarity even in the presence of new or modified word forms, making the model especially valuable for search tasks in the Kazakh language.

GloVe builds vector representations based on a matrix of word co-occurrence in a corpus, which allows it to capture global statistical relationships. However, the model does not consider the context of a word in a specific sentence, which limits its application in the semantic matching of complex constructs and hinders accurate performance with rare forms.

The BERT and KazBERT models are based on the transformer architecture and perform deep contextual encoding of text. Each word is represented considering its surrounding context in a sentence, and the final text vector is formed either from a special CLS-token or by aggregating the hidden states of tokens. KazBERT, unlike multilingual BERT, is trained on Kazakh language corpora, which allows it to account for the morphemic structure, syntactic features, and frequency patterns of the national lexicon.

In practice, this means that intelligent search systems using KazBERT or FastText can find relevant results even with significant differences between the query and the document content. For example, suppose a user enters a query with a rare word form or a synonym. In that case, contextual models can still determine semantic similarity and provide a relevant answer. In contrast, TF-IDF and GloVe demonstrate low sensitivity in such conditions and, consequently, less accurate results.

The formalization of semantic similarity calculations as geometric operations in vector space (cosine measure, Euclidean distance, probabilistic estimates) not only allows for machine interpretation of texts but also for building adaptive search systems on their basis that are robust to lexical transformations and linguistic diversity. This is particularly important when working with the Kazakh language, where rich morphology and syntactic variability require a deeper linguistic interpretation.

With the rapid growth of textual information in low-resource languages, there is an increasing need to develop intelligent search systems capable of accurately interpreting query intent and finding relevant documents, considering the morphological and syntactic features of a specific language. Standard approaches often overlook the agglutinative nature of the Kazakh language, its complex affixation, and word form variability, resulting in a significant decrease in search relevance.

The relevance of constructing a specialized mathematical model is driven by the need to ensure the system's robustness to grammatical transformations and to increase the precision of semantic matching amid lexical ambiguity.

The developed model is a hybrid architecture that combines contextual encoding using the KazBERT model and an adaptive semantic similarity metric, supplemented by a morpho-syntactic modifier. The fundamental difference of the proposed model lies in the preliminary morphological analysis of both the query and the documents,

during which roots, affixes, grammatical features, and syntactic dependencies are identified. This data is integrated into the model's attention mechanism, enhancing the contribution of structurally relevant tokens to the final vector representation.

The model is implemented in Python using the HuggingFace Transformers, Kaznlp, and Scikit-learn libraries. The base model is KazBERT (pre-trained on a Kazakh-language corpus). Tokenization and morphological analysis are performed using Kaznlp, after which a CLS vector represents each text.

To assess semantic similarity between queries and documents, the cosine measure is used, further adjusted by a morpho-syntactic correspondence scale that considers the coincidence of grammatical features: case, number, voice, possession, and others (Fig. 1).

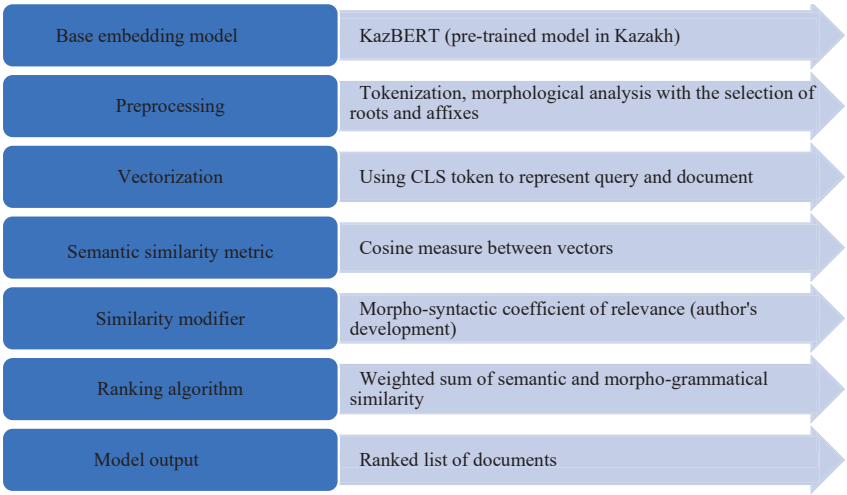


Figure 1. Structure and functional components of the mathematical model of an intelligent search system adapted to the Kazakh language
Source: author's own development.

The model's operation begins with the analysis of the query and documents: a morphological parsing is performed to extract all grammatical features, which form the basis for subword tokenization. The resulting representations are processed by KazBERT, where contextualized CLS vectors are formed. Then, the vectors are compared using the cosine measure, but the final ranking value is further modified to account for the structural overlap of morphological and syntactic parameters.

Such integration enables consideration of the variability in meaning expression characteristic of the Kazakh language, while maintaining semantic relevance despite significant differences in the form of expression between the query and the document.

The model validation experiment was conducted on a manually labeled test set containing 500 "query-relevant document" pairs. The queries were grouped by type: single-word keywords, short phrases, syntactically complex queries, questions,

stylistically neutral queries, and synonymous paraphrases. The test document base (24,000 texts) included news, academic articles, and user information records. Precision, Recall, and F1 metrics in the top-5 documents were used for evaluation (Table 3).

The results demonstrate the significant superiority of the developed model in Kazakh language search tasks. Particularly high accuracy was achieved in analyzing complex queries with variable word forms and non-standard lexemes, where classical models show a significant drop in accuracy. The use of the morpho-syntactic modifier allowed for an average increase in the F1-score by 11–15% compared to Multilingual BERT and by 33% compared to TF-IDF.

Table 3 – Comparative evaluation of information retrieval model effectiveness based on metrics of precision, recall, and robustness to morphological variations

Model	**Precision**	**Recall**	**F1-score**	**Robustness to Morphological Variations**	**Advantages**
TF-IDF	0.58	0.42	0.48	Low	Simplicity, speed
FastText	0.68	0.61	0.64	Medium	Handles rare words, robust to affixation
Multilingual BERT	0.73	0.68	0.70	Medium	Contextual encoding, accuracy
KazBERT + morphoanalysis	0.84	0.78	0.81	High	Accounts for grammar, adapted to the Kazakh language, and accuracy
TF-IDF	0.58	0.42	0.48	Low	Simplicity, speed

Source: author's own development.

A reduction in the number of irrelevant results was also noted, especially in the presence of synonymic variation in queries. Thus, the constructed model is a viable solution for creating intelligent search systems in a low-resource language environment. It provides a stable interpretation of meaning under conditions of high morphological variability, increases search accuracy and recall, and can be effectively integrated into electronic libraries, state registries, and educational and media platforms.

Discussion. Despite significant progress in the development of vector models for text representation, their application to Kazakh-language data is accompanied by several unresolved problems that significantly affect the quality of intelligent search. One of the main difficulties is the recognition of variable word forms, caused by the agglutinative nature of the Kazakh language.

Unlike languages with an analytical or inflectional structure, in Kazakh, a lexical unit can take dozens of grammatical forms through the attachment of sequential affixes, which creates difficulties in training models that cannot automatically generalize morphologically related forms. This leads to gaps in the semantic space between words that express the same concept but are presented in different written

variants (Abu-Salih, 2018; Mrkšić et al., 2017). An additional challenge is the limited representation of rare and regionally specific vocabulary in existing corpora. Standard models, especially those trained on multilingual or general-purpose datasets, show low sensitivity to lexemes that are absent from the main corpus. This applies to both rare words and terms reflecting the cultural and historical realities of Kazakhstan. Similar challenges are noted in studies on the application of vector models in conditions of semantic shift or weak contextual representation (Günther et al., 2019; Ren et al., 2020). Without a mechanism for extended morpho-semantic generalization, models lose their ability to correctly interpret the meaning of such words in context, reducing search relevance.

A critical factor remains the limited volume and diversity of Kazakh-language training data. Despite some initiatives to create national corpora, their size, genre diversity, and annotation quality still lag behind their English or Russian counterparts. Research in information retrieval emphasizes the importance of well-labeled "query-document" pairs for improving model stability and accuracy (Thakur et al., 2021; Tareaf et al., 2024). Consequently, when moving from laboratory conditions to real user scenarios, a drop in the quality of semantic matching is observed.

The development of effective intelligent information search systems for the Kazakh language requires not only selecting a suitable model architecture but also considering the language's specific characteristics, user query types, and the system's functional tasks. Practical recommendations should be based on a balance between accuracy, computational costs, and robustness to language variations. This approach is supported by modern trends in integrating hybrid models (Mitra & Craswell, 2018; Karpukhin et al., 2020).

When dealing with short or keyword queries containing one or two words, it is advisable to use FastText as the base model, as it demonstrates high robustness to morphological variability due to its handling of n-grams. In cases where queries are formulated as complex phrases, questions, or include rare and non-standard forms, preference should be given to contextual transformer models, particularly KazBERT, which can capture syntactic dependencies and contextual nuances.

Combining models can significantly enhance the system's overall effectiveness. In practice, a hybrid approach is advisable: preliminary document filtering is performed using a faster model (TF-IDF or FastText), after which the final ranking is conducted based on KazBERT with the inclusion of morpho-syntactic analysis. This helps to reduce computational resources while maintaining high accuracy in the relevant sample.

Furthermore, when designing the model, it is essential to provide for separate processing of negative constructions, homonyms, and synonyms, especially in legal, educational, and scientific texts. To improve the interpretability of the results, it is recommended to use not only semantic similarity metrics but also an additional coefficient of grammatical correspondence, which is particularly relevant for languages with a flexible word structure. Additionally, implementing user

customization options (e.g., selecting "strict match" or "semantic approximation" mode) can enhance end-user satisfaction.

Thus, building an adaptive architecture for intelligent search requires the integration of several vector models, with the ability to switch between them depending on the query conditions and usage goals. Special attention is given to adapting algorithms to the morphological nature of the Kazakh language and expanding the training data.

Conclusion. As a result of this research, it was established that the use of context-dependent vector models, particularly KazBERT, significantly improves the efficiency of intelligent information retrieval in the Kazakh language compared to classical approaches like TF-IDF and Word2Vec.

The developed mathematical model, which combines KazBERT and a morpho-syntactic modifier, demonstrated the highest precision and recall, especially when processing variable word forms and synonymous constructions.

Key remaining challenges include the inability of most models to recognize morphologically related forms, the limitations of training corpora, and poor coverage of rare and dialectal vocabulary. A practical solution lies in a hybrid architecture, combining initial document filtering with FastText and final ranking with KazBERT, along with grammatical correction.

The five advanced hypotheses were empirically confirmed: contextual models showed an advantage in quality metrics, the cosine measure proved stable in combination with morphological analysis, and the combined approach provided the best balance between accuracy and computational efficiency. Prospects for further research include expanding specialized corpora of the Kazakh language, fine-tuning transformers, and developing new adaptive metrics for semantic matching.

The results obtained can be used in the creation of intelligent search systems and language platforms for other agglutinative languages.

References

Abu-Salih B. (2018) Applying vector space model (VSM) techniques in information retrieval for Arabic language. arXiv preprint, arXiv:1801.03627. DOI: https://doi.org/10.48550/arXiv.1801.03627 (in English).

Gillick D., Kulkarni S., Lansing L., Presta A., Baldridge J., Ie E. & Garcia-Olano D. (2019) Learning dense representations for entity retrieval. arXiv preprint, arXiv:1909.10506. DOI: https://doi.org/10.48550/arXiv.1909.10506 (in English).

Günther F., Rinaldi L. & Marelli M. (2019) Vector-space models of semantic representation from a cognitive perspective: a discussion of common misconceptions. Perspectives on Psychological Science, 14(6). — P.1006-1033. DOI: https://doi.org/10.1177/1745691619861372 (in English).

Hassan-Montero Y. & Herrero-Solana V. (2024) Improving tag-clouds as visual information retrieval interfaces. arXiv preprint, arXiv:2401.04947. DOI: https://doi.org/10.48550/arXiv.2401.04947 (in English).

Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D. & Yih W.-T. (2020) Dense passage retrieval for open-domain question answering. arXiv preprint, arXiv:2004.04906v2. URL: https://arxiv.org/pdf/2004.04906v2 (in English).

Khater B.S., Abdul Wahab A.W.B., Idris M.Y.I.B., Hussai M.A. & Ibrahim A.A. (2019)A lightweight perceptron-based intrusion detection system for fog computing. Applied Sciences, 9(1). — P. 178. DOI: https://doi.org/10.3390/app9010178 (in English).

Li X., Jin J., Zhou Y., Zhang Y., Zhang P., Zhu Y. & Dou Z. (2025) From matching to generation: A survey on generative information retrieval. ACM Transactions on Information Systems, 43(3). — P. 1-62. DOI: https://doi.org/10.1145/372255 (in English).

Mitra B. & Craswell N. (2018). An introduction to neural information retrieval. Foundations and Trends in Information Retrieval, 13(1). — P. 1-126. DOI: https://doi.org/10.1561/1500000061 (in English).

Mrkšić N., Vulić I., Ó Séaghdha D., Leviant I., Reichart R., Gašić M., Korhonen A. & Young S. (2017) Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. Transactions of the Association for Computational Linguistics, 5. — P. 309-324. DOI: https://doi.org/10.1162/tacl_a_00063 (in English).

Nie Y., Chen H. & Bansal M. (2019) Combining fact extraction and verification with neural semantic matching networks. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01). — P. 6859-6866. DOI: https://doi.org/10.1609/aaai.v33i01.33016859 (in English).

Ren H., Hu W. & Leskovec J. (2020) Query2box: Reasoning over knowledge graphs in vector space using box embeddings. arXiv preprint, arXiv:2002.05969. DOI: https://doi.org/10.48550/arXiv.2002.05969 (in English).

Tareaf R.B., AbuJarour M., Engelman T., Liermann P. & Klotz J. (2024) Accelerating contextualization in AI large language models using vector databases. International Conference on Information Networking (ICOIN). — P. 316-321. DOI: https://doi.org/10.1109/ICOIN59985.2024.10572088 (in English).

Thakur N., Reimers N., Rücklé A., Srivastava A. & Gurevych I. (2021) BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint, arXiv:2104.08663. DOI: https://doi.org/10.48550/arXiv.2104.08663 (in English).

Van Gysel C., De Rijke M. & Kanoulas E. (2018) Neural vector spaces for unsupervised information retrieval. ACM Transactions on Information Systems, 36(4). — P. 1-25. DOI: https://doi.org/10.1145/3196826 (in English).

Zhu Y., Yuan H., Wang S., Liu J., Liu W., Deng C., Chen H., Liu Z., Dou Z. & Wen J.-R. (2023) Large language models for information retrieval: a survey. arXiv preprint, arXiv:2308.07107. DOI: https://doi.org/10.48550/arXiv.2308.07107 (in English).

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 176–187

https://doi.org/10.32014/2025.2518-1726.371

ҒТМР 62.50.43
ӘОЖ 004.932.2

A. Sambetbayeva[1], V. Jotsov[2], 2025.
[1]Al-Farabi Kazakh National university, Almaty, Kazakhstan;
[2]University of Library Studies and Information Technologies, Sofia, Bulgaria.
E-mail: sambetbaevamea@gmail.com

COMPARATIVE ANALYSIS OF DEEP LEARNING ARCHITECTURES FOR ROAD CRACK SEGMENTATION

Sambetbayeva Aizhan — doctoral student, Al-Farabi Kazakh National university, Almaty, Kazakhstan,
E-mail: sambetbaevamea@gmail.com, ORCID ID: https://orcid.org/0000-0003-0032-0533;
Jotsov Vladimir — Full Professor, University of Library Studies and Information Technologies, Sofia, Bulgaria,
E-mail: v.jotsov@unibit.bg, ORCID ID: https://orcid.org/0000-0002-2860-7918

Abstract. This article presents a comparative analysis of five state-of-the-art deep learning architectures used for the segmentation of road surface cracks: CrackNet, DeepCrack++, YOLOv9, ViT-UNet, and Swin-UNet. The evaluation was conducted using several publicly available datasets with varying levels of annotation. Key performance metrics included Recall, Precision, mean Intersection over Union (IoU), and F1-score. The experimental results showed that the DeepCrack++ model achieved the highest recall and precision, indicating its strong capability to detect various types of damage with minimal false positives. The CrackNet architecture stood out for its exceptional processing speed and low computational requirements, making it particularly suitable for resource-constrained embedded systems. YOLOv9, adapted for the segmentation of both micro- and macro-cracks, demonstrated a balanced trade-off between accuracy and processing speed. ViT-UNet leveraged global context modeling through attention mechanisms to provide more detailed identification of fine and branched cracks. Meanwhile, Swin-UNet effectively combined local and global features, resulting in stable performance across diverse datasets. Thus, the presented findings can serve as a foundation for selecting the most appropriate architecture in the development of intelligent road infrastructure monitoring systems, where a balance between accuracy, computational efficiency, and response time is essential.

Keywords: deep learning, crack, CrackNet, DeepCrack++, YOLO9, ViT-UNet, Swin-UNet, segmentation

А.К. Самбетбаева1, В. Иоцов2, 2025.
[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Кітапханаық зерттеулер мен ақпараттық технологиялар университеті, София, Болгария.
E-mail: sambetbaevamea@gmail.com

ЖОЛ ТӨСЕМІНІҢ ЖАРЫҚТАРЫН СЕГМЕНТАЦИЯЛАУДА ҚОЛДАНЫЛАТЫН ТЕРЕҢ ОҚЫТУ АРХИТЕКТУРАЛАРЫН САЛЫСТЫРМАЛЫ ТАЛДАУ

Самбетбаева Айжан Кудайбергеновна — докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: sambetbaevamea@gmail.com, ORCID ID: https://orcid.org/0000-0003-0032-0533;
Йоцов Владимир — профессор, Кітапханаық зерттеулер мен ақпараттық технологиялар университеті, София, Болгария,
E-mail: v.jotsov@unibit.bg, ORCID ID: https://orcid.org/0000-0002-2860-7918.

Аннотация. Бұл мақалада жол төсеміндегі жарықтарды сегментациялауда қолданылатын терең оқытудың заманауи архитектуралары салыстырмалы түрде талданды: CrackNet, DeepCrack++, YOLOv9, ViT-UNet және Swin-UNet. Зерттеу әртүрлі деңгейдегі аннотациялары бар бірнеше ашық деректер жиынтығы негізінде жүргізілді. Бағалау көрсеткіштері ретінде толықтық (Recall), нақтылық (Precision), IoU орта мәні және F1-өлшемі қолданылды. Зерттеу нәтижелері DeepCrack++ моделінің толықтық пен нақтылық бойынша жоғары көрсеткіштерге қол жеткізгенін, яғни әртүрлі зақымдану түрлерін анықтау тиімділігін көрсетті. Ал CrackNet архитектурасы есептеу ресурстарын үнемді пайдалануымен және өте жоғары өңдеу жылдамдығымен ерекшеленіп, ендірілген жүйелер үшін қолайлы шешім болып табылады. YOLOv9 моделі микро- және макрожарықтарды сегментациялау үшін бейімделіп, нақтылық пен жылдамдық арасында оңтайлы тепе-теңдікті көрсетті. ViT-UNet моделі жаһандық контекстіні терең модельдеу мүмкіндігі арқылы жіңішке және тарамдалған жарықтарды дәлірек анықтай алды. Swin-UNet болса, локалдық және жаһандық белгілерді тиімді біріктіріп, әртүрлі деректер жиынтықтарында тұрақты нәтиже көрсетті. Сонымен, зерттеу нәтижелері жол инфрақұрылымының жағдайын интеллектуалды бақылау жүйелеріне арналған модельдерді таңдауда дәлдік, есептеу ресурстары мен өңдеу уақыты арасындағы оңтайлы тепе-теңдікті ескере отырып, тиімді архитектураны таңдауға негіз бола алады.

Түйін сөздер: терең оқыту, жарықтар, CrackNet, DeepCrack++, YOLO9, ViT- UNet, Swin-UNet, сегментациялау

А.К. Самбетбаева1, В. Иоцов2, 2025.
[1]Казахский национальный университет имени Аль-Фараби,
Алматы, Казахстан;
[2]Университет библиотечных исследований и информационных технологий,
София, Болгария.
E-mail: sambetbaevamea@gmail.com

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АРХИТЕКТУР ГЛУБОКОГО ОБУЧЕНИЯ ПРИ СЕГМЕНТАЦИИ ТРЕЩИН НА ДОРОЖНЫХ ПОКРЫТИЯХ

Самбетбаева Айжан Кудайбергеновна — докторант, Казахский Национальный университет имени Аль-Фараби, Алматы, Казахстан,
E-mail: sambetbaevamea@gmail.com, ORCID ID: https://orcid.org/0000-0003-0032-0533;
Йоцов Владимир — профессор, Университет библиотечных исследований и информационных технологий, София, Болгария,
E-mail: v.jotsov@unibit.bg, ORCID ID: https://orcid.org/0000-0002-2860-7918.

Аннотация. Статья посвящена сравнительному анализу пяти передовых архитектур глубокого обучения, применяемых для сегментации трещин на дорожном покрытии: CrackNet, DeepCrack++, YOLOv9, ViT-UNet и Swin-UNet. Анализ проводился на основе нескольких общедоступных датасетов с различным уровнем разметки, а в качестве ключевых метрик использовались полнота (Recall), точность (Precision), среднее значение IoU и F1-мера. Согласно результатам экспериментов, модель DeepCrack++ продемонстрировала наивысшие значения полноты и точности, что указывает на её высокую эффективность в обнаружении различных типов повреждений с минимальным уровнем ложных срабатываний. В то же время CrackNet выделилась исключительной скоростью обработки и низким потреблением вычислительных ресурсов, что делает её особенно подходящей для использования в ресурсограниченных встраиваемых системах. Модель YOLOv9, адаптированная для сегментации как микро-, так и макротрещин, показала сбалансированные характеристики: высокую точность и сравнительно быстрое время обработки. ViT-UNet благодаря способности моделировать глобальный контекст с помощью механизмов внимания, обеспечила более детализированное выявление тонких и разветвлённых трещин. В свою очередь, Swin-UNet, эффективно объединяя локальные и глобальные признаки, продемонстрировала стабильную производительность на различных наборах данных. Таким образом, представленные результаты могут служить основой для выбора наиболее подходящей архитектуры при разработке интеллектуальных систем мониторинга состояния дорожной инфраструктуры, где необходимо учитывать соотношение между точностью, вычислительными затратами и временем отклика.

Ключевые слова: глубокое обучение, трещины, CrackNet, DeepCrack++, YOLOv9, ViT-UNet, Swin-UNet, сегментация

Кіріспе. Жол инфрақұрылымы елдің экономикалық өркендеуі мен әлеуметтік дамуының маңызды тірегі болып табылады. Жол төсеміндегі жарықтардың уақытында анықталмауы жолдың құрылымдық беріктігін әлсіретіп, шұңқырлар мен ойықтардың пайда болуына, жол-көлік оқиғаларының артуына, сондай-ақ жөндеуге кететін шығындардың күрт ұлғаюына әкеледі. Жарықтарды анықтауда «binding–pointing–crossword» негіздемесін пайдалану көпөлшемді сипаттамаларды бөліп көрсете отырып, аймақтар мен қиылыстарды нақты ажыратуға мүмкіндік береді. Binding кезеңінде жол төсемінің құрылымдық қабаттары, материалдық құрамы және қызмет ету кезінде әсер ететін сыртқы факторлар, яғни климаттық жағдайлар, жүк жүктемесі, пайдалану ерекшеліктері жан-жақты қарастырылады. Сонымен қатар, осы кезең негізгі жарықтың айналасындағы микрожарықтардың бағыты мен тығыздығын жол төсемінің құрылымдық қасиеттері мен сыртқы әсерлерімен байланыстырып, негізгі жарықтарды дәл болжауға мүмкіндік береді. Ал Pointing кезеңінде жол төсеміндегі жарықтар нақты анықталып, оларды геометриялық орналасуы және түрі бойынша картаға түсіру жүзеге асырылады. Бұл сатыда микрожарықтар мен шұңқырлардың координаттары, ұзындығы және ені, сондай-ақ олардың өзара байланысы өлшеніп, цифрлық модельдерге енгізіледі. Crossword кезеңінде модельдердің интерпретациясында ықтимал қателердің азаюын көруге болады.

Жарықтарды пиксель деңгейінде қарастыратын толық немесе жартылай сегментация әдістері «binding–pointing–crossword» негіздемесіне сүйене отырып, әрбір пиксельдің жарыққа тиесілігін дәл анықтап, зақымданудың нақты формасын қалпына келтіруге мүмкіндік береді. Сегментациялау – жол төсеміне қатысты зақымдануларды дәл анықтап, жөндеу жұмыстарын ерте жоспарлауда ең маңызды бейне-технологиялардың бірі. Жол төсеміндегі жарықтарды сегментациялау модельдерін оқыту және жарамдылығын тексеру үшін қолданылатын деректер жиынтығы: Crack500, CFD (Crack Forest Dataset). Модельдің шынайы ортадағы жұмыс қабілетін бағалау үшін CFD деректер жиынтығында қосымша тестілеу жүргізіледі. Crack500, Crack Forest Dataset деректер жиынтығы табиғи орта жағдайларында түсірілген жол жамылғысындағы бейнелерді қамтиды және модельдің шынайы ортада жұмыс қабілетін тексеру үшін қосымша дерек көзі ретінде пайдаланылады. CrackNet, DeepCrack++, YOLOv9 архитектураларын оқыту және тестілеу үшін Crack500 және Crack Forest Dataset деректер жиынтығы таңдалынды және заманауи конволюциялық нейрондық желі оқытылады. Қолданылатын деректер жиынтығының ерекшеліктері мен қолданылу тиімділігі 1-кестеде көрсетілген.

1- кесте. Деректер жиынтығының ерекшеліктері мен қолданылу тиімділігі

Деректер жиынтығы	Ерекшеліктері	Қолдану тиімділігі
Crack500	Жоғары сапалы RGB суреттері, әрқай-сысында жарықтар дәл белгіленген. Маскалар толық қолжетімді.	Сызықтық құрылымды модельдерге ең жиі қолданылады.
CFD (Crack Forest Dataset)	Табиғи жағдайдағы әртүрлі жарық түрлері бар және маскалар толық белгіленген.	Модель шынайы ортада жұмыс істейді.

CrackNet, DeepCrack++ архитектуралары сызықтық құрылымдарды көпмасштабты деңгейде зерттеу, қосымша қабаттарда өңдеу әдістерін пайдаланады. CrackNet архитектурасы жол төсеміндегі жарықтарды анықтау үшін арнайы жасалған және жарықтардың сызықтық құрылымын дәл сипаттауға бағытталған. Бұл модельдің негізгі артықшылығы — құрылымдық ерекшеліктерді сақтай отырып, жарықтарды пиксель деңгейінде анықтай алуы және жол инфрақұрылымын бақылауға бағытталған зерттеулер мен нақты қолданбалы жобаларда сенімді таңдау болып табылады. DeepCrack++ архитектурасы жіңішке, үзілісті жарықтарды да дәл айқындай алады және шағын деректер жиынтығымен жақсы нәтижелер көрсете алады. Модельдің encoder-decoder құрылымы мен контекстік ақпаратты терең өңдеуі оны жол зақымдануларын автоматты түрде, жоғары дәлдікпен анықтауға ең қолайлы архитектуралардың біріне айналдырады. YOLOv9 архитектурасы «Pointing» кезеңінде жол төсеміндегі зақымдарды жылдам әрі қарапайым локализациялауға арналған тиімді әдіс болып табылады. Бұл архитектуралар жол жарықтарын анықтауда өзге модельдерге қарағанда нақты тапсырмаларға жақсы бейімделуімен және жоғары нәтижелілікпен ерекшеленеді.

Конволюционды нейрондық желілердің (CNN) классикалық архитектуралары жергілікті контекстті жақсы өңдесе де, жарықтың бүкіл ұзындығы бойындағы байланыстарды толықтай қамти алмайды, ал Vision Transformer және Swin Transformer трансформерлерінің архитектуралары кеңістіктік тәуелділіктерді иерархиялық түрде өңдей отырып, жол төсемінің күрделі текстураларын жоғары дәлдікпен көрсете алады (Danilescu et al., 2015). Pointing кезеңінде бейнені бірнеше бөліктерге бөле отырып, ұзындығы мен енін, сондай-ақ олардың өзара байланысын өлшеуге болады. Vision Transformer трансформерінің архитектурасы үлкен қашықтықтағы пиксель байланыстарын тиімді модельдей алады, бірақ есептеу шығыны жоғары, ал Swin Transformer трансформерінің архитектурасы сегментацияда микрожарықтарды да дәл табуға мүмкіндік береді. Қарастырылған трансформерлер жол жарықтарын анықтауда өзге модельдерге қарағанда нақты тапсырмаларға жақсы бейімделуімен және жоғары нәтижелігімен ерекшеленеді.

Материалдар мен әдістер. Елімізде жол төсемінің зақымдануларын автоматты түрде анықтау және жіктеу саласында елеулі жетістіктер байқалады. Жол төсеміндегі зақымдануларды анықтауда терең оқыту әдістерін қолдануға

арналған заманауи зерттеулер бұл салада трансформерлік модельдер мен арнайы конфигурацияланған конволюциялық нейрондық желілердің тиімділігін айқындайды. Бұл әдістер жол төсемінің күрделі құрылымында, атап айтқанда көлеңке, жол таңбалары мен ластану сияқты кедергілер жағдайында да жоғары дәлдікпен жұмыс істеуге қабілетті. Әсіресе, оқыту деректері шектеулі және аннотациясы толық емес жағдайларда бұл тәсілдердің тұрақтылығы мен бейімділігі ерекше мәнге ие. Сондықтан модельдерді нақты жұмыс жағдайларына бейімдеу – жол инфрақұрылымын интеллектуалды басқару жүйелерінде жоғары нәтижеге қол жеткізудің негізгі факторы болып табылады. Көптеген еңбектерде бейнелерді талдаудың әртүрлі әдістері ұсынылады, ал ұсынылған әдістердің талдау сараптамасы әдебиеттік шолуда келтірілген.

И.А.Канаеваның жол төсемінің зақымдануын автоматты түрде анықтау есебінде деректерді синтетикалық таңдау алгоритмінің жұмыс істеу қағидасы ұсынылған. Пиксель деңгейінде жол төсемінде пайда болған жарықтарды сегменттеу үшін Mask R-CNN архитектурасы қолданылған және нақты нәтижелер алынған (Kanaeva et al., 2021).

Жарықтарды анықтауға бағытталған еңбекте (Eisenbachet al., 2017) жарықтарды тану тәсілдерін талдау барысында екі негізгі әдіс тобы бөлініп көрсетілген: бейнелерді фильтрлеу және классификаторлар құру. Бейнелерді фильтрлеу әдісі жол төсеміндегі зақымданулардың құрылымдық ерекшеліктерін анықтауға бағытталған. Алдымен бейнеге сыртқы әсердің ықпалын барынша азайту үшін алдын ала өңдеу жүргізіледі. Одан соң сызатқа тиесілі пиксельдердің қарқындылығы ең төмен мәндермен берілетіні ескеріліп, шек мәні бойынша фильтрлеу қолданылады. Алынған контурды нақтылау мақсатында морфологиялық амалдар мен байланыстылық компоненттерін іздеу әдістері пайдаланылады. Мұндай әдістер Li H. және әріптестерінің еңбектерінде де жан-жақты сипатталған.

Hamishebahar Y. және әріптестері жазған жұмысында терең оқыту әдістеріне негізделген жарықтарды тану тәсілдері жан-жақты талданған (Hamishebahar et al., 2022). Авторлар жарықтардың әртүрлі түрлерін (ұзын, қисық, әлсіз көрінетін) анықтау үшін көпмасштабты ерекшеліктерді өңдеу қабілеті бар модельдердің маңыздылығын атап өтеді. Сондай-ақ, зерттеуде CrackNet, DeepCrack++, CrackGAN сияқты заманауи архитектуралардың құрылымдық ерекшеліктері салыстырылған. Олар жарықтардың сызықтық табиғатын сақтай отырып, контекстік және кеңістіктік ақпаратты тиімді біріктіруге мүмкіндік береді.

Maeda H. және әріптестері жол төсеміндегі зақымдануларды анықтау үшін генеративті-жарысу желілерін қолдану мүмкіндігін зерттеді (Maeda et al., 2020). Қарастырылған тәсіл деректердің шектеулі көлемі жағдайында модельдің жалпылау қабілетін арттыруға мүмкіндік береді.

Жол төсеміндегі жарықтарды анықтау үшін трансферлік оқытуды қолдана отырып, терең конволюциялық нейрондық желі моделі ұсынылды

(Jana et al., 2022). Авторлар алдын ала оқытылған нейрондық желілерді пайдалану арқылы есептеу шығындарын азайтып, кішігірім деректер жиынтықтарымен де жоғары дәлдікке қол жеткізу мүмкін екенін көрсетті. Бұл тәсіл жол жарықтарын анықтау жүйелерін практикалық қолдануға бейімдеуді жеңілдетеді.

Kim B. және әріптестері жарықтарды тану үшін жеңілдетілген CNN архитектурасын қолдануды зерттеді. Олардың жұмысында есептеу ресурстарын үнемдей отырып, жеткілікті дәлдікке қол жеткізу мүмкіндігі қарастырылған. Мұндай желілер нақты уақыттағы қолданбаларда, мысалы, жол инспекциясының мобильді жүйелерінде тиімді шешім бола алады (Kim et al., 2021)

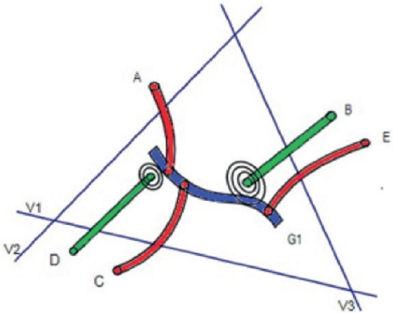
Maniat M. және әріптестері Google Street View кескіндерін пайдалана отырып, жарықтарды визуалды түрде анықтаудың жаңа әдісін ұсынды (Maniat et al., 2021). Бұл зерттеу қолжетімді дереккөздерден алынған кескіндер негізінде зақымдануларды қашықтықтан бақылау мүмкіндіктерін кеңейтеді.

Garvanova M. және әріптестерінің еңбегінде «binding–pointing–crossword» негіздемесі қарастырылып, көпөлшемді деректер құрылымындағы логикалық шектеулердің жаңа топтамалары сипатталған (Garvanova et al., 2023).

Терең оқыту әдістері жол төсемінде пайда болатын жарықтарды сенімді түрде сегментациялай отырып, классификациялай алады. Сегментациялау барысында суреттің қай жерінде жарық бар, қай жерінде басқа объектілер орналасқанын анықтайтын маска жасалады. Binding карта көрші орналасқан пиксельдердің арасындағы «binding–pointing–crossword» байланысты бейнелей отырып, жіңішке жарықтарды көлеңкеден, жол белгілерінен ажырата алады. Pointing карта жарықтың траекториясын көрсететін әрбір пиксельдің бағыттаушы векторын береді Нақты шешімдерді қабылдау үшін pointing және crossword модульдерінде қолдануға болады.

1-суретте «binding–pointing–crossword» негіздемесінің схемасында жол төсеміндегі жарықтың орналасу аймағы V_1, V_2, V_3 сызықтарымен шектелген. G_1 жарығына түсірілген A,C,E қисықтары жол төсемінің жарығында орналасқан, B және D қисықтары жол төсеміндегі жарықты талдауда binding профилін бейнелейді.

1-сурет. «Binding–pointing–crossword» негіздемесі



Жарықтарды анықтауда «binding–pointing–crossword» негіздемесін пайдалану көпөлшемді сипаттамаларды бөліп көрсете отырып, аймақтар мен қиылыстарды нақты ажыратуға мүмкіндік берді. Жарықтардың бейнесіндегі көзге көрінбейтін микрожарықтарды анықтау осы негіздеменің көмегімен жүзеге асты. Егер микрожарықтарды уақытында анықтамаса, олардың ішіне әртүрлі лас заттар және су кіріп, жарықтың ұлғаюына алып келеді.

Нәтижелер мен талқылау. Негізгі аймақтың ішкі маскасында қабаттасудың центрі бола алатын нүкте таңдалынып және G^{mask} өлшеміне тең T^{mask} аймағы қиып алынды. Маскалар аймағында екі бейне ығысуының $\overline{G_c}$ орта мәні есептеледі:

$$\overline{G_c} = \frac{1}{k}\sum_{s} G_c(s) \cdot (1 - G^{mask}(s))$$

мұндағы k- пиксель саны, $G_c(s)$ – c каналындағы G_1 жол жарығы суретінің s пиксельдік мәні, $G^{mask}(s)$ – жол жарығының бинарлы маскадағы пиксельдік мәні.

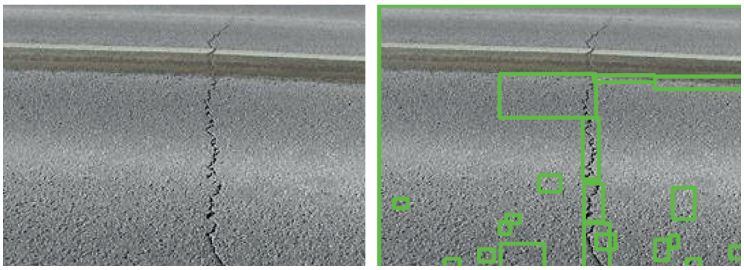
Жол төсемінің жарықтарын бейнеге орналастырғанда маска астындағы зақымдану мәндері ғана қарастырылады:

$$M(s) = G^{mask}(s) \cdot T^{mask}(s)$$

Бейнелердің кеңістіктік қасиеттерін сақтау мақсатында, маска аймағы бейненің орталық бөліктеріне бағытталып таңдалады. Бұл тәсіл жол бетінің зақымдануларын бейнелеу кезінде ақпараттың толық сақталуына мүмкіндік берді және терең оқыту модельдерінің оқыту тиімділігін жоғарылатты, бұл модельді оқыту кезінде контекстік ақпаратты тиімді пайдалануға мүмкіндік берді. Оқыту таңдамасының ақпараттылығын арттыру үшін бір бейнеде 1-ден 6-ға дейін жол төсемінің жарықтары қарастырылды. Нәтижесінде оқыту таңдамасы 500 бейнені құрады.

2-суретте көрсетілгендей, жол бетінде пайда болған бойлық бағыттағы жарық құрылымдары автоматтандырылған тәсіл арқылы анықталып, жасыл түсті тіктөртбұрыштармен белгіленген. Бұл визуализация терең оқыту негізіндегі сегментациялық әдістердің нәтижесін модельдеу арқылы жүзеге асырылып, пиксельдік деңгейде жарық шекараларын нақты көрсетуге мүмкіндік береді. Жасыл квадраттар жарықтың орналасқан аймағын локализациялап, олардың геометриялық пішіні мен бағытталған таралу сипаты жөнінде маңызды мәлімет берді.

2-сурет. Жол бетіндегі жарықтар және оның имитациясы



Күрделі және үздіксіз емес жарық құрылымдары 2-суретте нақты бөлініп көрсетілді, бұл өз кезегінде жол төсеміндегі зақымданулардың нақты пішінін, ұзындығын және орналасуын талдауға мүмкіндік береді.

Жинақталған нәтижелер CrackNet пен DeepCrack++ архитектураларының тиімділігі мен қолдану ерекшеліктеріндегі айырмашылықтарды айқын көрсетті. CrackNet архитектурасы жеңіл әрі жылдам жұмыс істей отырып, жіңішке жарықтарды локализациялауда жақсы нәтиже көрсетсе, DeepCrack++ моделі кең ауқымды және күрделі құрылымды зақымдану аймақтарын жоғары дәлдікпен анықтай алды. ViT-UNet архитектурасы күрделі текстуралы және көрінбейтін жарықтарды дәл сегментациялайды. Swin-UNet «жылжытылған терезе» механизмін иерархиялық өңдеумен үйлестіре отырып, есептеу тиімділігін сақтайды. Бұл модель микрожарықтарды жоғары дәлдікпен анықтай отырып, ViT-UNet-ке қарағанда жылдамырақ жұмыс істейді және құрама құрылымы бар ақауларды да сенімді кескіндей алады. CrackNet, DeepCrack++, ViT-Unet, Swin-Unet және YOLOv9 архитектураларының тиімділігі деректер жиынтығында бірнеше негізгі метрика бойынша бағаланды. Атап айтқанда, нақтылық, толықтық, F1-көрсеткіші және IoU көрсеткіштері модельдердің жол төсеміндегі жарықтарды қаншалықты сапалы анықтай алатынын көрсетуге мүмкіндік берді. Архитектуралардың негізгі көрсеткіштері төмендегі 2-кестеде келтірілген.

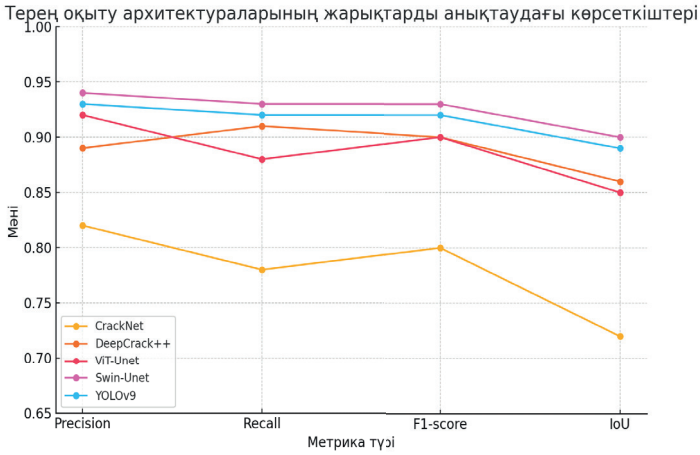
2- кесте. Архитектуралардың жарықтарды анықтаудағы тиімділік көрсеткіштері

Метрика	**CrackNet**	**DeepCrack++**	**ViT-UNet**	**Swin-UNet**	**YOLOv9**
Precision (нақтылық)	0.82	0.89	0.92	0.94	0.94
Recall (толықтық)	0.78	0.91	0.88	0.93	0.93
F1-score	0.80	0.90	0.90	0.93	0.93
IoU (ауыспалы сәйкесу)	0.72	0.86	0.85	0.90	0.90
Processing time (сек)	0.08	0.15	0.30	0.25	0.25

Талдау нәтижелері бойынша Swin-UNet моделі барлық сапалық метрикалар бойынша CrackNet-ке қарағанда жоғары нәтиже көрсеткенімен, өңдеу уақыты да ұзағырақ болды. Ал Swin-UNet архитектурасы жеңіл әрі жылдамырақ жұмыс істей отырып, нақты локализацияда тиімділік танытты.

3-суретте терең оқыту архитектураларының негізгі сегментациялық метрикалар бойынша салыстырмалы нәтижелері көрсетілген және барлық көрсеткіштер салыстырылған.

3-сурет. Терең оқыту архитектураларының салыстырмалы көрсеткіштері



Зерттеу нәтижелері DeepCrack++ архитектурасының барлық негізгі метрикалар бойынша CrackNet моделінен асып түсетінін көрсетті. Әсіресе, толықтық және ауыспалы сәйкестік көрсеткіштері бойынша DeepCrack++ артықшылығы айқын байқалды, бұл оның күрделі құрылымды жарықтарды неғұрлым сенімді және жан-жақты анықтай алатынын дәлелдейді. CrackNet моделі есептеу ресурстарын үнемдей отырып, жоғары жылдамдықпен жұмыс істеуімен ерекшеленеді, бұл оны шынайы уақыттағы кірістірілген жүйелерде қолдануға қолайлы етеді.

Сонымен қатар, ViT-UNet және Swin-UNet трансформерлік архитектуралары да жоғары дәлдікпен күрделі құрылымды және жіңішке жарықтарды анықтауда өз тиімділігін көрсетті. ViT-UNet жіңішке жарық құрылымдарын дәл айқындай алды, ал Swin-UNet архитектурасы белгілерді үйлестіру арқылы микрожарықтарды сенімді сегменттеуді қамтамасыз етті.

Бейнелерді алдын ала өңдеу кезеңінде контрастты арттыру және шу деңгейін төмендету мақсатында артефактілерді азайту алгоритмдері қолданылды. Бұл шаралар жарық пен көлеңке әсерінен туындайтын қателіктерді барынша азайтуға мүмкіндік беріп, жүйенің жалпы дәлдігі мен тиімділігін арттырды.

Қорытынды. Мақалада жол төсеміндегі жарықтарды анықтау және дәл сегментациялау есебін шешу үшін CrackNet, DeepCrack++, ViT-UNet, Swin-UNet және YOLOv9 терең оқыту архитектуралары салыстырмалы түрде талданды. Зерттеу нәтижелері DeepCrack++ архитектурасының толықтық пен нақтылық бойынша жоғары нәтижелер көрсеткенін, CrackNet архитектурасының өңдеу жылдамдығы мен ресурстық тиімділігімен ерекшеленетінін көрсетті. Сонымен қатар, трансформердің ViT-UNet

архитектурасы жіңішке және тармақталған жарықтарды тиімді анықтаса, Swin-UNet әртүрлі белгілерді үйлестіріп, тұрақты нәтижелерге қол жеткізді. YOLOv9 архитектурасы жылдамдық пен дәлдіктің оңтайлы тепе-теңдігін қамтамасыз етіп, микро- және макрожарықтарды табуда сенімділік көрсетті.

Зерттеу аясында синтетикалық деректер жиынтығына негізделген оқыту DeepCrack++ сияқты модельдер үшін жоғары пиксельдік дәлдік пен орташа дәлдік көрсеткіштеріне қол жеткізуге мүмкіндік беретіні анықталды. Бұл қолмен таңбаланған нақты деректермен оқытуға қарағанда бірқатар артықшылықтар береді. Сонымен қатар, жарық құрылымының күрделілігі, жол бетінің әртүрлілігі және бейнедегі кедергі факторлар классикалық әдістердің тиімділігін төмендететіні анықталды. Интеллектуалды жүйелер жол зақымдануларын ерте кезеңде анықтап, техникалық қызмет көрсету мен жөндеу жұмыстарын тиімді жоспарлауға мүмкіндік береді. Бұл жол инфрақұрылымының сапасын арттырып, апаттар қаупін азайтады және жөндеу шығындарын азайтады.

Мақаланы әзірлеу барысында жасанды интеллект құралдары көмекші ретінде пайдаланылды.

References

Danilescu D. et al. (2015) Road Anomalies Detection Using Basic Morphological Algorithms. Carpathian Journal of Electronic and Computer Engineering 2(8). — P. 15-18 (in English)

Kanaeva I.A., Ivanova Yu.A. (2021) Road Defect Segmentation Based on Synthetic Sample Generation Using Deep Generative Adversarial Convolutional Networks. Computer optics, 45(6). — P. 907–916. (in English)

Eisenbach M. et al. (2017) How to get pavement distress detection ready for deep learning A systematic approach. IEEE. — P. 2039-2047 (in English)

Feng C., Koch C., McKee M. (2017) Vision-Based Crack Detection for Bridge Inspection. Computer-Aided Civil and Infrastructure Engineering 32(6). — P. 416-429 (in English)

Garvanova M., Jotsov V. (2023) A Data-Science Approach for Creation of a Comprehensive Model to Assess the Impact of Mobile Technologies on Humans. Applied Sciences 13(6). — P. 3600 (in English)

Hamishebahar Y., Guan H., So S., Jo J. (2022) A Comprehensive Review of Deep Learning-Based Crack Detection. Applied Sciences 12(3). – P. 1374 (in English)

Jana S., Thangam S., Kishore A., Kumar V.S., Vandana S. (2022) Transfer learning based deep convolutional neural network model for pavement crack detection from images. International Journal of Nonlinear Analysis and Applications 13(1). — P.1209-1223 (in English)

Kim B., Yuvaraj N., Sri Preetha K., Arun Pandian R. (2021) Surface crack detection using deep learning with shallow CNN architecture for enhanced computation. Neural Computing and Applications 33(15). — P. 9289-9305 (in English)

Li H. et al. (2018) Automatic Pavement Crack Detection by Multi-Scale Image Fusion. IEEE Transactions on Intelligent Transportation Systems. — P. 1-12 (in English)

Liu Y., Zhang L., Wei X., Zhang Y. (2020) DeepCrack A Deep Hierarchical Feature Learning Architecture for Crack Segmentation. Neurocomputing 338. — P. 139-153 (in English)

Maeda H., Kashiyama T., Sekimoto Y., Seto T., Omata H. (2020) Generative adversarial network for road damage detection. Computer-Aided Civil and Infrastructure Engineering 36(1). — P. 47-60 (in English)

Maniat M., Camp C., Kashani A. (2021) Deep learning-based visual crack detection using Google Street View images. Neural Computing and Applications 33(21). — P. 14565-14582 (in English)

Shi Y., Cui L., Qi Z., Meng F., Chen Z. (2016) Automatic Road Crack Detection Using Random Structured Forests. IEEE Transactions on Intelligent Transportation Systems 17(12). — P. 3434-3445 (in English)

Yang F. (2019) Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. — P. 1-11 (in English)

Zhang L., Yang F., Zhang Y., Zhu Y. (2016) Road crack detection using deep convolutional neural network.IEEE International Conference on Image Processing (ICIP). — P. 3708-3712 (in English)

Zhou W., Song K., Li M., Chen J., Han Z. (2022) CrackFormer A Transformer-Based Framework for Crack Detection Using Convolutional Encoder. IEEE Transactions on Industrial Electronics 70(2). — P. 1736-1746 (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 188–203

https://doi.org/10.32014/2025.2518-1726.372

МРНТИ 28.23.37
УДК 004.89

D. Oralbekova[1*], A. Akhmediyarova[2], D. Kassymova[3], Z. Alibiyeva[2], 2025.
[1]Institute of information and computational technologies, Almaty, Kazakhstan;
[2] Satbayev University, Almaty, Kazakhstan;
[3] ALT University, Almaty, Kazakhstan.
E-mail: dinaoral@mail.ru

RESEARCH ON LINGUISTIC ANALYSIS METHODS FOR IDENTIFYING AND EXTRACTING TEXT DATA IN THE KAZAKH LANGUAGE

Oralbekova Dina — PhD, senior researcher, Institute of information and computational technologies, Almaty, Kazakhstan,
E-mail: dinaoral@mail.ru, ORCID ID: https://orcid.org /0000-0003-4975-6493;
Akhmediyarova Ainur — PhD, professor, Satbayev University, Almaty, Kazakhstan,
E-mail: a.akhmediyarova@satbayev.university, ORCID ID: https://orcid.org/0000-0003-4439-7313;
Kassymova Dinara — PhD, assistant professor, ALT University, Almaty, Kazakhstan,
E-mail: d.kassymova@alt.edu.kz, ORCID ID: https://orcid.org/0000-0001-6152-8317;
Alibiyeva Zhibek — PhD, Associate professor, Satbayev University, Almaty, Kazakhstan,
E-mail: zh.alibiyeba@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Abstract. This paper examines modern linguistic analysis methods applied to the processing of the Kazakh language for the purpose of automatic identification and extraction of textual information. Special attention is given to morphological, syntactic, and semantic analysis and their adaptation to the specific features of the Kazakh language, which is classified as an agglutinative language and is characterized by flexible word order. These features create certain challenges when applying traditional approaches designed for languages with fixed word order, such as English. The study analyzes contemporary approaches, including finite-state machine methods, statistical models, deep neural networks, and transformer-based architectures. It reviews existing software tools such as HFST, Apertium, KazNERD, BeeBERT, and Kaz-RoBERTa, as well as other models specifically adapted for languages with complex morphological structures. Their potential and limitations are assessed in the context of Kazakh text processing. Particular focus is placed on the accuracy of morphological analysis, the models' robustness to polysemy, and their ability to handle rare and complex word forms. The paper also discusses practical applications of modern NLP solutions for the Kazakh

language — in machine translation systems, automatic text classification, named entity recognition, and sentiment analysis. Concrete examples of model usage in the educational and legal domains are presented. Finally, the paper provides recommendations for developing national text corpora, advancing morphological analysis tools, and further exploring the integration of different methodological approaches to improve the quality of Kazakh language processing in NLP tasks.

Keywords: Kazakh language, Transformer, morphological analysis, syntactic and semantic analysis, NLP, pretrained models

Д. Оралбекова[1*], А. Ахмедиярова[2], Д. Қасымова[3], Ж. Алибиева[2], 2025.
[1] Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;
[2] Satbayev университеті, Алматы, Қазақстан;
[3] М. Тынышпаев атындағы ALT университеті, Алматы, Қазақстан.
E-mail: dinaoral@mail.ru

ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДІК АҚПАРАТТЫ АНЫҚТАУ ЖӘНЕ ОНЫ ШЫҒАРЫП АЛУ ҮШІН ЛИНГВИСТИКАЛЫҚ ТАЛДАУ ӘДІСТЕРІН ЗЕРТТЕУ

Оралбекова Дина — PhD, аға ғылыми қызметкер, Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,
E-mail: dinaoral@mail.ru, ORCID ID: https://orcid.org/0000-0003-4975-6493;
Ахмедиярова Айнұр — PhD, профессор, Satbayev университеті, Алматы, Қазақстан,
E-mail: a.akhmediyarova@satbayev.university, ORCID ID: https://orcid.org/0000-0003-4439-7313;
Қасымова Динара — PhD, ассистент-профессор, М. Тынышпаев атындағы ALT университеті. Алматы, Қазақстан,
E-mail: d.kassymova@alt.edu.kz, ORCID ID: https://orcid.org/0000-0001-6152-8317;
Алибиева Жибек — PhD, қауымдастырылған профессор, Satbayev университеті, Алматы, Қазақстан,
E-mail: zh.alibiyeva@satbayev.university, ORCID ID: https://orcid.org/0000-0001-9565-5621.

Аннотация. Бұл мақалада қазақ тілін өңдеуге бағытталған заманауи лингвистикалық талдау әдістері қарастырылды. Мәтіндік ақпаратты автоматты түрде анықтау және шығарып алу мақсатында қолданылатын тәсілдерге ерекше назар аударылды. Морфологиялық, синтаксистік және семантикалық талдау түрлері мен олардың қазақ тіліне бейімделуі егжей-тегжейлі сипатталды. Қазақ тілі агглютинативті тілдер қатарына жатқандықтан сөз тәртібінің еркіндігімен ерекшеленеді. Мұндай ерекшеліктер ағылшын тілі сияқты сөз тәртібі қатаң тілдерге арналған дәстүрлі тәсілдерді қолдануда белгілі бір қиындықтар туғызады. Зерттеуде қазіргі таңдағы тәсілдер қарастырылған, оның ішінде атап айтқанда келесілер келтірілген: ақырлы автоматтар әдістері, статистикалық модельдер, терең нейрондық желілер және трансформер негізіндегі архитектуралар талданады. HFST, Apertium, KazNERD, BeeBERT және Kaz-RoBERTa сияқты бағдарламалық құралдармен қатар, күрделі

морфологиялық құрылымдарға бейімделген басқа да модельдерге мен тәсілдерге шолу жасалды. Бұл құралдардың қазақ мәтінін өңдеу контекстіндегі мүмкіндіктері, артықшылықтары мен шектеулері сарапталды. Морфологиялық талдаудың дәлдігіне, модельдердің көпмағыналылыққа төзімділігіне және сирек әрі күрделі сөз тұлғаларын өңдеу қабілетіне ерекше назар аударылды. Сонымен қатар, қазіргі NLP шешімдерінің қазақ тіліне арналған практикалық қолдану салалары машиналық аударма жүйелері, мәтіндерді автоматты түрде жіктеу, атаулы мәндерді тану және тональдікті талдау мәселелері қозғалады. Модельдердің білім беру және құқық салаларында қолданылуына нақты мысалдары келтіріледі. Мақала соңында ұлттық мәтін корпустарын құру және өңдеу, морфологиялық талдау құралдарын жетілдіру, сондай-ақ қазақ тілін өңдеудің сапасын арттыру мақсатында түрлі әдістемелік тәсілдерді біріктіру бойынша ұсыныстар берілген.

Түйін сөздер: қазақ тілі, Transformer, морфологиялық талдау, синтаксистік және семантикалық талдау, NLP, алдын ала үйретілген модельдер

Д. Оралбекова[1]*, А. Ахмедиярова[2], Д. Касымова[3], Ж. Алибиева[2], 2025.

[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

ИССЛЕДОВАНИЕ МЕТОДОВ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ДЛЯ ВЫЯВЛЕНИЯ И ИЗВЛЕЧЕНИЯ ТЕКСТОВЫХ ДАННЫХ НА КАЗАХСКОМ ЯЗЫКЕ

Оралбекова Дина — PhD, старший научный сотрудник, Институт информационных и вычислительных технологий, Алматы, Казахстан,
E-mail: dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493;
Ахмедиярова Айнур — PhD, профессор, Satbayev Университет, Алматы, Казахстан,
E-mail: a.akhmediyarova@satbayev.university, https://orcid.org/0000-0003-4439-7313;
Касымова Динара — PhD, ассистент-профессор, ALT университет имени М. Тынышпаева, Алматы, Казахстан,
E-mail: d.kassymova@alt.edu.kz, https://orcid.org/0000-0001-6152-8317;
Алибиева Жибек — PhD, ассоциированный профессор, Satbayev Университет, Алматы, Казахстан,
E-mail: zh.alibiyeva@satbayev.university, https://orcid.org/0000-0001-9565-5621.

Аннотация. В данной статье рассматриваются современные методы лингвистического анализа, применяемые для обработки казахского языка, с целью автоматического выявления и извлечения текстовой информации. Особое внимание уделяется морфологическому, синтаксическому и семантическому анализу, а также их адаптации к особенностям казахского языка, который относится к агглютинативным языкам и характеризуется свободным порядком

слов. Это создаёт определённые трудности при применении традиционных подходов, разработанных для языков с фиксированным порядком слов, таких как английский. В исследовательской работе анализируются современные подходы, включая методы на основе конечных автоматов, статистические модели, глубокие нейронные сети и трансформерные архитектуры. Рассматриваются существующие программные инструменты, такие как HFST, Apertium, KazNERD, BeeBERT и Kaz-RoBERTa и другие модели, специально адаптированные для языков со сложной морфологической структурой, а также их потенциал и ограничения в контексте обработки казахских текстов. Особое внимание уделяется вопросам точности морфологического анализа, устойчивости моделей к полисемии, а также способности справляться с редкими и сложными словоформами. Также обсуждаются практические области применения современных NLP-решений для казахского языка — в системах машинного перевода, автоматической классификации текстов, извлечении именованных сущностей и анализе тональности. Представлены конкретные примеры применения моделей в образовательной и юридической сферах. В заключении даны рекомендации по созданию национальных текстовых корпусов, развитию инструментов морфологического анализа, а также дальнейшему исследованию интеграции различных методологических подходов для повышения качества обработки казахского языка в задачах NLP.

Ключевые слова: казахский язык, Transformer, морфологический анализ, синтаксический и семантический анализ, NLP, предобученные модели

Благодарности. *Данное исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант BR24993166).*

Введение. Обработка естественного языка (NLP) для казахского языка представляет собой ряд уникальных задач, обусловленных его агглютинативной природой и свободным порядком слов. Эти особенности требуют разработки специализированных методов лингвистического анализа для выявления и извлечения текстовых данных. Современные подходы, включая методы глубокого обучения, трансформерные модели и статистические техники, открывают возможности для создания высокоточных инструментов обработки текста. Однако ограниченная доступность размеченных данных и сложность грамматической структуры языка затрудняют реализацию подобных решений.

В данном исследовании основное внимание уделяется анализу современных методов лингвистического анализа — морфологического, синтаксического и семантического — и их адаптации к особенностям казахского языка.

NLP — это ключевое направление в области искусственного интеллекта и лингвистики, предоставляющее передовые методы автоматического анализа текста для таких задач, как извлечение информации, машинный перевод и анализ тональности. Однако для казахского языка, как и для многих других

языков с ограниченными ресурсами, разработка эффективных NLP-решений сопряжена со значительными трудностями.

Казахский язык относится к классу агглютинативных языков, в которых грамматическое значение выражается посредством аффиксов, присоединяемых к корню слова. Это требует создания специализированных морфологических анализаторов, способных точно интерпретировать сложные и многокомпонентные словоформы. Кроме того, свободный порядок слов в предложениях создает дополнительные сложности для синтаксического и семантического анализа, поскольку традиционные методы, разработанные для языков с фиксированным порядком слов, часто оказываются неэффективными.

Современные технологии машинного обучения, такие как глубокие нейронные сети и трансформерные модели, открывают новые возможности для обработки казахских текстов. Модели, такие как BeeBERT и KazNERD (Yeshpanov et al., 2022), демонстрируют заметный прогресс в анализе текстов, однако их производительность по-прежнему зависит от наличия крупных размеченных корпусов, которые в случае казахского языка пока остаются ограниченными.

Цель данного исследования — провести обзор и систематизацию существующих методов лингвистического анализа, адаптировать их под казахский язык и оценить применимость различных подходов, включая конечные автоматы, статистические модели и трансформеры. Это позволит определить современные достижения в обработке казахского языка и обозначить перспективные направления для дальнейшего развития, такие как создание новых текстовых корпусов и интеграция различных методологических подходов.

Материалы и методы исследования. Морфологический анализ — это процесс разбиения слова на его составные части (корень и аффиксы) и определение их грамматических значений. Для агглютинативных языков, таких как казахский, морфологический анализ особенно важен, поскольку грамматические значения передаются с помощью многочисленных аффиксов, присоединяемых к корню слова.

Методы, основанные на правилах и словарях

Метод, основанный на правилах и словарях, представляет собой базовый подход к вычислительному морфологическому анализу. В словарной базе хранятся базовые формы слов (леммы), а набор правил описывает порядок присоединения аффиксов и их взаимодействие (Jurafsky et al., 2019).

Словари, являясь центральным элементом данного подхода, содержат морфологические характеристики лексем, такие как часть речи и основные грамматические признаки (род, число, падеж и т. д.) (Haspelmath et al., 2013). При проверке слова по словарю анализатор применяет правила для определения структуры и значения слова (Kaplan et al., 1994).

К ключевым инструментам данной категории относятся TRMorph и Apertium. TRMorph — морфологический анализатор для турецкого языка,

обеспечивающий точный анализ за счёт строгих правил и структурированных словарей (Kim, 2024). Аналогично, Apertium — это программное обеспечение с открытым исходным кодом, предназначенное для морфологического анализа и машинного перевода, поддерживающее агглютинативные языки, включая казахский (Forcada et al., 2011).

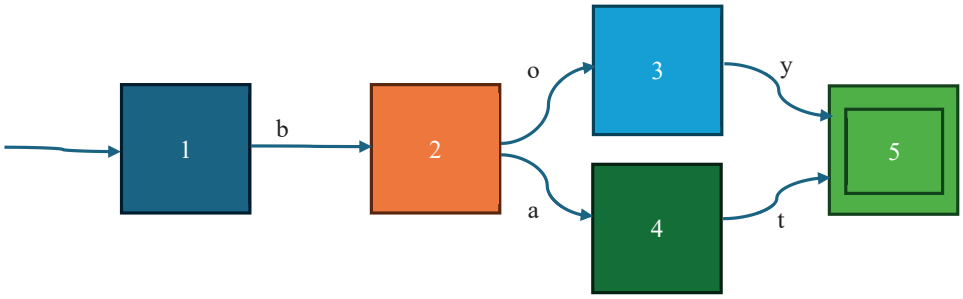
Подход, основанный на правилах и словарях, обладает рядом преимуществ, таких как высокая точность для слов, присутствующих в словаре, и относительная простота реализации для языков с хорошо изученной морфологией. Однако он также имеет ограничения — в частности, слабая способность обрабатывать неизвестные слова и значительные затраты на разработку словарей и правил.

Конечные автоматы

Метод конечных автоматов применяется для моделирования регулярных структур и морфологических процессов, что делает его особенно полезным для анализа агглютинативных языков. Конечные автоматы могут быть детерминированными (DFA) или недетерминированными (NFA), что позволяет представлять сложные грамматические системы с помощью формальных языков и регулярных выражений (Boyd et al., 2021).

Конечный автомат представляет собой графовую структуру, где узлы соответствуют состояниям, а переходы между ними моделируют морфологические преобразования. Морфологический анализатор обрабатывает входное слово, проходя по состояниям автомата и применяя заранее заданные правила для извлечения морфологических характеристик слова (Beesley et al., 2003) (рис. 1).

Рисунок 1. Общая структура конечного автомата



Наиболее широко используемыми инструментами в этой области являются HFST и FOMA.

HFST (Helsinki Finite-State Transducer) — это платформа для построения конечных автоматов, широко применяемая для анализа агглютинативных языков, включая казахский. HFST поддерживает интеграцию с другими инструментами, такими как Apertium, и обеспечивает высокую производительность благодаря оптимизированным алгоритмам (Lindén et al., 2011).

FOMA — гибкий и легковесный инструмент для разработки и тестирования конечных автоматов. Он используется для создания сложных морфологических моделей и поддерживает различные форматы ввода/вывода, что делает его универсальным решением (Hulden, 2009).

Метод конечных автоматов обладает высокой эффективностью при обработке регулярных структур, таких как порядок аффиксов в агглютинативных языках. Его адаптивность делает его подходящим для построения компактных и быстрых морфологических анализаторов. Однако у него имеются и недостатки: ограниченная гибкость при обработке нерегулярных форм и исключений, а также необходимость глубоких знаний в области формальных языков, морфологии агглютинативных языков и программирования, что может создавать сложности для разработчиков (Manohar et al., 2022).

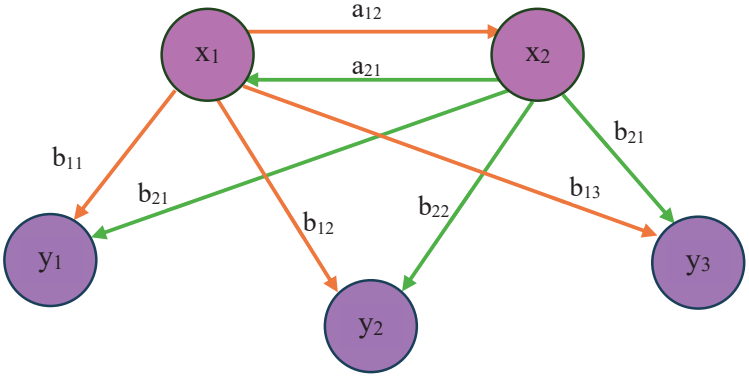
Статистические модели

Методы статистического анализа основаны на вероятностных подходах для предсказания морфологических тегов, что делает их адаптивными и пригодными для обработки языков с большими корпусами данных. В отличие от систем, основанных на правилах, эти модели не ограничены заранее заданными правилами, что позволяет им обрабатывать неизвестные слова и учиться на реальных текстах (Koosha et al., 2022).

Основной принцип статистических моделей заключается в использовании распределений вероятностей для анализа последовательностей слов и их морфологических тегов. Такие методы, как скрытые марковские модели (HMM) и условные случайные поля (CRF), используют вероятностное графовое моделирование для захвата сложных зависимостей между аффиксами и лексемами (Fraser, 2008).

HMM оценивают наиболее вероятную последовательность тегов для заданного входа на основе вероятностей переходов. Однако эффективность этих моделей ограничивается предположением, что каждое состояние зависит только от предыдущего (Wang, 2022) (рис. 2).

Рисунок 2. Пример скрытой марковской модели



В отличие от HMM, CRF предлагают более продвинутый подход, учитывая глобальные зависимости и взаимодействия между элементами входной последовательности. Это делает их более эффективными для обработки сложных языков, таких как казахский (Oralbekova et al., 2022).

Методы статистического анализа имеют ряд преимуществ: они способны обрабатывать неизвестные слова и использовать вероятностные рассуждения, что делает их весьма адаптивными к новым данным. Для языков с большими размеченными корпусами, таких как английский или китайский, они обеспечивают высокую точность морфологического анализа.

Основными недостатками статистических моделей являются их зависимость от большого объема размеченных данных, а также необходимость значительных вычислительных ресурсов и знаний в области машинного обучения для обучения и тонкой настройки моделей.

Методы глубокого обучения

Современные методы глубокого обучения используют нейронные сети для автоматического извлечения морфологических закономерностей из данных. Такой подход продемонстрировал высокую эффективность благодаря способности нейросетей обрабатывать сложные языковые зависимости и адаптироваться к особенностям агглютинативных языков, таких как казахский.

Глубокие нейронные сети, такие как LSTM (долгая краткосрочная память) и GRU (затворные рекуррентные единицы), широко применяются для обработки последовательных данных. Эти модели особенно хорошо подходят для задач, требующих учета порядка элементов, таких как анализ последовательностей аффиксов в словах. LSTM и GRU способны удерживать долгосрочные зависимости и учитывать контекст при анализе морфологической структуры (Vennerød et al., 2021).

Более продвинутые архитектуры, такие как трансформеры (включая BERT и Kazakh-BERT), используют механизмы внимания для параллельной обработки всего контекста слова. Это позволяет эффективно анализировать сложные морфологические структуры и контекстуальные вариации даже в языках с богатой морфологией (Vaswani et al., 2017).

Методы глубокого обучения обладают рядом преимуществ: они достигают высокой точности при наличии больших объемов данных, могут моделировать сложные морфологические закономерности, включая нерегулярные формы и редкие комбинации аффиксов. Кроме того, они способны обрабатывать слова с несколькими возможными интерпретациями (Devlin et al., 2019).

Основные недостатки этих методов — высокая вычислительная стоимость и зависимость от больших размеченных корпусов. Обучение таких моделей требует значительных аппаратных ресурсов, а также тщательной подготовки текстовых данных.

Методы синтаксического анализа

Синтаксический анализ (парсинг) играет ключевую роль в обработке текстов, особенно для языков со свободным порядком слов, таких как

казахский. Он позволяет выявить структурные зависимости между словами в предложении и сформировать их грамматическое представление. Методы синтаксического анализа можно условно разделить на два подхода: аналитический (на основе правил) и статистический (на основе моделей машинного обучения).

Зависимостный парсинг (Dependency Parsing)

Зависимостный парсинг представляет собой метод, при котором строится графовая структура, в которой узлы соответствуют словам, а ребра обозначают синтаксические связи между ними. Такая структура наглядно показывает, какие элементы предложения зависят друг от друга и как они связаны.

Основная цель метода — определить зависимости между словами в предложении. Например, подлежащее и сказуемое связаны отношением зависимости. Этот подход особенно эффективен для языков со свободным порядком слов, таких как казахский, где грамматические функции передаются морфологически, а не фиксированной позицией в предложении.

Существуют два основных алгоритма зависимостного анализа: 1) Переходный анализ (Transition-Based Parsing) — строит дерево зависимостей пошагово. Каждый шаг определяет действие (например, добавить ребро или перейти к следующему слову), что делает алгоритм быстрым и эффективным. Однако он чувствителен к ошибкам на ранних этапах. 2) Графовый анализ (Graph-Based Parsing) — строит глобально оптимальное дерево зависимостей, формулируя задачу как нахождение максимального остовного дерева в графе. Обеспечивает высокую точность за счёт рассмотрения всех возможных связей, но требует больших вычислительных ресурсов.

Фразовая структура (Constituency Parsing)

Фразовый анализ представляет предложение как иерархическую структуру, где каждый узел соответствует грамматической единице: фразе, придаточному предложению или всему предложению. Метод исходит из предположения, что каждое предложение можно разделить на более мелкие грамматические компоненты.

Дерево фразовой структуры показывает, как слова объединяются в более крупные синтаксические единицы. Например, глагольная фраза (VP) может включать глагол и его дополнения. Этот метод полезен для глубокого синтаксического анализа и широко применяется в машинном переводе и лингвистическом анализе.

Алгоритмы фразового анализа:

Алгоритм CKY (Cocke–Kasami–Younger) — используется для построения деревьев разбора на основе контекстно-свободных грамматик (CFG). Применяет динамическое программирование, обеспечивая эффективность для языков с фиксированным порядком слов.

Глубокие нейросети — современные подходы используют LSTM и трансформеры для предсказания фразовых структур. Эти модели способны

учитывать контекст всего предложения, что особенно полезно для языков со свободным порядком слов, таких как казахский (Oralbekova et al., 2024).

Статистический и нейросетевой парсинг

Современные методы синтаксического анализа всё чаще используют глубокие нейронные сети для автоматического выявления синтаксических паттернов из больших текстовых корпусов. Эти методы достигают высокой точности и адаптивности, что делает их пригодными для анализа текстов на различных языках, включая казахский.

Ключевыми инструментами являются BiLSTM и трансформеры. BiLSTM (двунаправленная LSTM) эффективно обрабатывает последовательные зависимости, анализируя как предшествующий, так и следующий контекст. Трансформеры (включая BERT и GPT) моделируют сложные синтаксические структуры с помощью механизмов внимания, позволяя учитывать весь контекст предложения (Narejo et al., 2024).

Методы семантического анализа

Семантический анализ включает извлечение смысла из текста, включая определение значений слов, выявление связей между ними и интерпретацию контекста. Он играет ключевую роль в таких задачах, как машинный перевод, извлечение информации и анализ тональности.

Векторные представления слов (Word Embeddings)

Методы векторного представления слов преобразуют слова в числовые векторы, отражающие их значение в контексте. Эти методы основаны на дистрибутивной гипотезе, согласно которой слова с близкими значениями появляются в сходных контекстах. Векторные представления обучаются на основе анализа совместной встречаемости слов в текстах, что делает их универсальным инструментом обработки текстов.

Ключевые методы:

Word2Vec — создаёт векторы слов с использованием двух архитектур:

- CBOW (непрерывный мешок слов) предсказывает текущее слово по его окружению.
- Skip-Gram предсказывает окружающие слова по текущему.

Word2Vec хорошо подходит для кластеризации слов и анализа семантического сходства. Например, слова «қала» (город) и «ауыл» (село) будут иметь близкие векторы.

GloVe (Global Vectors for Word Representation) учитывает глобальную статистику совместной встречаемости слов в больших корпусах. Обучается на матрице встречаемости и хорошо справляется с задачами семантического сравнения.

FastText в отличие от Word2Vec и GloVe, работает на уровне подслов, анализируя последовательности символов в словах. Это особенно эффективно для морфологически богатых языков, таких как казахский, где аффиксы существенно меняют значение слова. Например, «мектеп» (школа) и «мектептер» (школы) будут иметь схожие представления.

Основные преимущества этих моделей — простота реализации, быстрая скорость обучения и интуитивная интерпретация: семантически близкие слова отображаются близко в векторном пространстве. Однако они имеют ограничения: 1) Каждому слову присваивается один вектор, что затрудняет обработку многозначных слов. 2) Сложно учитывать сложные контекстные зависимости, особенно в длинных предложениях.

Семантическое ролевое аннотирование (Semantic Role Labeling, SRL)

Семантическое ролевое аннотирование — это процесс определения ролевой структуры предложения, при котором каждому слову или фразе присваивается определённая функция в контексте действия. Этот метод позволяет распознавать субъекты, объекты, действия и другие компоненты, формируя насыщенное семантическое представление текста.

SRL направлен на определение семантических ролей в предложении. Например, в предложении «Али прочитал книгу» SRL аннотирует «Али» как субъект, «прочитал» — как действие, а «книгу» — как объект. Таким образом, SRL является мощным инструментом для анализа сложных синтаксических структур, особенно в задачах извлечения информации.

Основные подходы к SRL

Модели на основе правил. Этот подход опирается на заранее определённые шаблоны и грамматические правила для присвоения семантических ролей. Он прост в реализации, но обладает низкой гибкостью и требует значительной ручной настройки под каждый язык.

Глубокое обучение. Современные методы используют RNN, LSTM и трансформеры для автоматического аннотирования ролей. Эти модели способны улавливать как локальные, так и глобальные контексты, обеспечивая высокую точность даже в сложных структурах предложений (Mamyrbayev, 2023).

SRL особенно полезен для извлечения информации в таких областях, как анализ требований, обработка юридических документов и системы вопросов-ответов. Он хорошо подходит для длинных предложений, где необходимо выявить сложные отношения между словами. Однако у метода есть ограничения: он требует больших размеченных корпусов, что затрудняет его применение в условиях ограниченных языковых ресурсов. Кроме того, модели глубокого обучения могут испытывать трудности при обработке очень длинных текстов, что увеличивает вычислительные затраты и требует дополнительных ресурсов (Onan, 2023).

Пример применения: SRL может быть использован для извлечения требований из технической документации. В предложении «Система должна позволять пользователям загружать файлы» SRL аннотирует «система» как субъект, «должна позволять» — как действие, а «пользователей» и «файлы» — как объекты.

Результаты и обсуждение. Модели на основе трансформеров, такие как BERT, RoBERTa, T5, BART и GPT, стали прорывными инструментами

в области обработки естественного языка благодаря способности учитывать как левый, так и правый контекст слова. Эти модели обучаются на задачах маскированного языкового моделирования (MLM) и предсказания следующего предложения (NSP), что позволяет им эффективно анализировать как локальные, так и глобальные зависимости в тексте (Wang et al., 2024).

Модели глубокого обучения широко применяются в задачах извлечения смысловой информации. Они автоматизируют сложные процессы анализа текста, такие как перевод, классификация, распознавание именованных сущностей (NER) и анализ тональности, что делает их незаменимыми при обработке казахского языка (табл. 1).

Машинный перевод. Модели, такие как BERT, значительно улучшают качество перевода казахских текстов. Предобученные трансформеры могут адаптироваться к агглютинативной морфологии языка и свободному порядку слов. Например, многоязычный BERT в сочетании с механизмами внимания улавливает контекст и грамматические структуры, повышая точность и естественность перевода.

Классификация текста. FastText, способный обрабатывать текст на уровне символов, хорошо подходит для классификации казахских документов. Модель эффективно справляется с морфологическими особенностями языка, обрабатывая тексты на различные темы. Например, FastText успешно применяется для автоматической категоризации документов по таким направлениям, как образование, политика и наука.

Распознавание именованных сущностей (NER). KazNERD — модель, специально разработанная для казахского языка, эффективно распознаёт сущности, такие как имена, организации и географические названия. Интеграция глубокого обучения с трансформерами обеспечивает точное извлечение сложных лингвистических конструкций, учитывая морфологическое богатство языка. Например, KazNERD используется при анализе юридических текстов, где важно идентифицировать участников и наименования организаций.

Семантическое сходство. Методы векторного представления текста, такие как Word2Vec, широко применяются для оценки семантического сходства. Эти подходы измеряют близость значений между двумя текстами или словами. Например, Word2Vec используется для кластеризации казахских текстов, группируя схожие документы на основе их содержания.

Анализ тональности. Модель BERT показывает высокую точность при решении задач анализа тональности на казахском языке. Она анализирует контекст слов в предложении, позволяя точно определять эмоциональную окраску (положительную, отрицательную или нейтральную). Такой подход используется, например, при анализе отзывов пользователей на казахском языке, что помогает компаниям понимать обратную связь и улучшать клиентский сервис.

Таблица №1 – Обзор инструментов лингвистического анализа для казахского языка

	Инструмент / Модель	Описание	Преимущества	Недостатки
Морфологический анализ	Apertium	Программное обеспечение с открытым исходным кодом для машинного перевода и морфологического анализа. Поддерживает казахский язык и использует словари на основе правил.	Подходит для базового анализа текста и перевода.	Ограничен фиксированными правилами, низкая эффективность при работе со сложными словоформами.
	HFST	Helsinki Finite-State Transducer, предназначенный для агглютинативных языков.	Эффективные модели на основе конечных автоматов, поддержка универсальных морфологических описаний.	Требует значительных усилий по настройке правил.
	KazNERD	Система распознавания именованных сущностей для казахского языка, интегрирует морфологический анализ.	Учитывает морфологические особенности языка, высокая точность.	Ограничена задачами распознавания именованных сущностей.
	BeeBERT	Адаптированная версия BERT для казахского языка.	Улавливает морфологические и синтаксические особенности, сохраняет высокую точность даже при малом объеме данных.	Требует высоких вычислительных ресурсов.
Синтаксический анализ	Stanford Parser	Использует универсальные грамматики зависимостей для анализа текста.	Поддерживает множество языков, включая казахский.	Сложно настраивается для агглютинативных языков.
	MSTParser	Реализует графовый подход к синтаксическому разбору зависимостей.	Точный и гибкий при анализе сложных синтаксических структур.	Медленная обработка больших корпусов.
	Berkeley Parser	Поддерживает несколько языков, использует статистические модели для парсинга по составляющим.	Высокая точность для агглютинативных языков.	Ограниченное количество моделей для языков с низкими ресурсами.
	Stanford Constituency Parser	Применяет статистические методы для синтаксического анализа.	Подходит для языков с фиксированным порядком слов.	Менее эффективен для казахского языка из-за его свободного порядка слов.

	UDPipe	Инструмент на основе BiLSTM, поддерживает токенизацию, морфологический и синтаксический анализ.	Универсальное решение с поддержкой многих языков.	Зависит от качества обучающих данных.
	SpaCy	Современная библиотека NLP для обработки текста.	Быстрая, легко интегрируется.	Ограниченная поддержка казахского языка.
Семан-тиче-ский анализ	KazSemEval	Платформа для оценки семантических задач на казахском языке, включая извлечение связей и анализ значений слов.	Специально разработана для казахского языка, учитывает лингвистические и культурные особенности.	Недостаточные ресурсы для решения сложных семантических задач.
	Kaz-RoBERTa	Модифицированная версия RoBERTa, адаптированная под казахский язык. Поддерживает широкий спектр задач NLP.	Высокая точность в задачах классификации текста, анализа тональности и извлечения информации.	Требует больших обучающих корпусов и значительных вычислительных ресурсов.

Нерешённые проблемы и направления дальнейших исследований

Одной из ключевых проблем в обработке казахского языка является ограниченность размеченных данных. Современные методы глубокого обучения, такие как трансформеры, требуют масштабных корпусов для обучения, что затрудняет разработку эффективных моделей. Будущие исследования должны быть сосредоточены на создании и аннотировании крупных текстовых корпусов, включая специализированные области, такие как право и медицина.

Агглютинативная структура казахского языка, при которой к корню слова присоединяется множество аффиксов, усложняет интеграцию морфологического и синтаксического анализа. Аналитические инструменты должны одновременно учитывать морфологические преобразования и синтаксические зависимости, что требует высоких вычислительных ресурсов и сложной настройки моделей.

Для языков со свободным порядком слов, таких как казахский, традиционные методы синтаксического анализа сталкиваются с трудностями. Грамматические функции определяются морфологическими маркерами, а не позицией слова в предложении, что требует разработки языкоориентированных моделей, способных учитывать эти особенности.

Несмотря на то, что современные модели, такие как BERT, эффективно улавливают контекст, они всё ещё сталкиваются с проблемой полисемии (множественности значений слов). Эта проблема особенно актуальна для казахского языка, где значение слова зависит от морфологии и контекста. В

будущем исследования должны быть направлены на улучшение разрешения полисемии, например, путём интеграции мультимодальных подходов.

Высокие вычислительные требования моделей глубокого обучения ограничивают их практическое применение. В дальнейшем необходимо разрабатывать оптимизированные и легковесные модели для внедрения в условия с ограниченными ресурсами. Передобучение на больших многоязычных корпусах, таких как Multilingual-BERT, с последующей донастройкой на меньших казахских выборках, может помочь решить проблему нехватки данных. Этот подход уже продемонстрировал свою эффективность в задачах синтаксического анализа.

Заключение. В данной работе представлен обзор современных методов лингвистического анализа для обработки текстов на казахском языке. Исследование охватывает подходы к морфологическому, синтаксическому и семантическому анализу, включая методы на основе глубоких нейронных сетей и трансформеров.

В анализе подчеркнуты значительные достижения в данной области, включая разработку специализированных инструментов и моделей, а также выявлены основные нерешённые проблемы. Ключевыми из них являются: нехватка размеченных данных, сложность обработки агглютинативной морфологии и высокие вычислительные затраты моделей глубокого обучения. Для преодоления этих барьеров необходима дальнейшая работа по развитию корпусов, адаптации моделей и созданию новых методов лингвистического анализа. Несмотря на существующие ограничения, результаты исследований показывают, что современные технологии обработки естественного языка могут значительно повысить эффективность автоматизированной обработки казахских текстов, открывая новые возможности для применения в образовании, праве и сфере искусственного интеллекта.

References

Boyd R.L., Schwartz H.A. (2021) Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. Journal of Language and Social Psychology, 40(1). — P. 21-41. https://doi.org/10.1177/0261927X20967028 (in English)

Beesley K.R., Karttunen L. (2003) Finite State Morphology, CSLI Studies in Computational Linguistics. Finite State Morphology, CSLI Studies in Computational Linguistics (in English)

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — P. 4171–4186 (in English)

Forcada M.L., et al. (2011) Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation. Machine Translation, vol. 25, no. 2. — P. 127–144 (in English)

Fraser A.M. (2008) Hidden Markov Models and Dynamical Systems. Society for Industrial and Applied Mathematics, USA. — P. 144, ISBN 0898717744, 9780898717747 (in English)

Haspelmath M., Sims A. (2013) Understanding Morphology (2nd ed.). Routledge, 384 p., eBook ISBN 9780203776506. https://doi.org/10.4324/9780203776506 (in English)

Hulden M. (2009) Foma: a finite-state compiler and library. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session (EACL '09). Association for Computational Linguistics, USA, 2009. — P. 29–32 (in English)

Jurafsky D., Martin J.H. (2019) Logistic Regression. In: Speech and Language Processing, 3rd

Edition (Draft). — P. 75-93. https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf (in English)

Kaplan R.M., Kay M. (1994) Regular Models of Phonological Rule Systems. Computational Linguistics, vol. 20, no. 3. — P. 331–378 (in English)

Kim Y. (2024) On morphological requirements for auxiliary verb periphrasis in Turkish. Glossa: a journal of general linguistics. Vol. 9(1), doi: https://doi.org/10.16995/glossa.9771 (in English)

Koosha S., Mahyar A., Yaser A., Godarzi A., Javad. (2022) Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model (October 2022). International Journal of Science and Information System Research, Volume 12, Issue 9. — P. 20 - 44, 2022, Available at SSRN: https://ssrn.com/abstract=4251017 (in English)

Lindén K., Axelson E., Hardwick S., Pirinen T.A., Silfverberg M. (2011) HFST—Framework for Compiling and Applying Morphologies. In: Mahlow, C., Piotrowski, M. (eds) Systems and Frameworks for Computational Morphology. SFCM. Communications in Computer and Information Science, vol 100. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23138-4_5 (in English)

Mamyrbayev O., Wojcik W., Titova N., Pavlov S., Oralbekova D., Aitkazina A., Zhumazhan N. (2023) Development of a thermodynamic model for optimization of processes in crop production. Eastern-European Journal of Enterprise Technologies, 6(8 (126), — P. 25–34, 2023. https://doi.org/10.15587/1729-4061.2023.290294 (in English)

Manohar K., Jayan A.R., Rajan R. (2022) Mlphon: A Multifunctional Grapheme-Phoneme Conversion Tool Using Finite State Transducers. IEEE Access, vol. 10. — P. 97555-97575, doi: 10.1109/ACCESS.2022.3204403 (in English)

Onan A. (2023) SRL-ACO: A text augmentation framework based on semantic role labeling and ant colony optimization. J. King Saud Univ. Comput. Inf. Sci., vol. 35, 101611 (in English)

Oralbekova D., Mamyrbayev O., Othman M., Alimhan K., Zhumazhanov B., Nuranbayeva B. (2022) Development of CRF and CTC Based End-To-End Kazakh Speech Recognition System, in Nguyen, N.T., Tran, T.K., Tukayev, U., Hong, T.P., Trawiński, B., Szczerbicki, E. (eds) Intelligent Information and Database Systems. ACIIDS 2022. Lecture Notes in Computer Science, vol. 13757, Springer, Cham. [Online]. Available: https://doi.org/10.1007/978-3-031-21743-2_41 (in English)

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. Communications in Computer and Information Science, vol 2211. Springer, Cham. https://doi.org/10.1007/978-3-031-72260-8_10 (in English)

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. IEEE Access, vol. 12, pp. 126587-126600, doi: 10.1109/ACCESS.2024.3446897 (in English)

Vennerød C.B., Kjærran A., Bugge E.S. (2021) Long Short-term Memory RNN. ArXiv, abs/2105.06756 (in English)

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., A. Gomez 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, 2017. — P. 6000–6010 (in English)

Wang Y. (2022) Using Machine Learning and Natural Language Processing to Analyze Library Chat Reference Transcripts. Information Technology and Libraries. Vol. 41. 10.6017/ital.v41i3.14967 (in English)

Wang J., Huang J.X., Tu X., Wang J., Huang A.J., Laskar Md T.R., Bhuiyan A. (2024) Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. ACM Comput. Surv. Vol. 56, 7. — P. 33 https://doi.org/10.1145/3648471 (in English)

Yeshpanov R., Khassanov Y., Varol H. A. (2022) KazNERD: Kazakh Named Entity Recognition Dataset. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. — P. 417–426 (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 204–227

https://doi.org/10.32014/2025.2518-1726.373

УДК 004.02

Zh.S. Takenova, 2025.
International Educational Corporation, Almaty, Kazakhstan.
E-mail: takenova@mail.ru

RESEARCH ON EXPERT ASSESSMENT METHODS FOR DETERMINING TEACHERS' PRIORITIES BY DISCIPLINE

Zhanar Takenova — Master of Economic Sciences, International Educational Corporation, Almaty, Kazakhstan,
E-mail: takenova@mail.ru, ORCID ID: https://orcid.org/0000-0001-8925-5808.

Abstract. To effectively allocate the teaching load across disciplines between teachers, an urgent task is to use the priority of teachers. Expert evaluation methods are the most widely used in education, so well-known expert evaluation methods were chosen to solve the current problem. Analysis of the solution methods, the range of application of the methods and the complexity of using the methods were carried out. The article reviews known peer review methods that are proposed to be used to assess faculty priority across disciplines, such as group assessment methods, pairwise comparisons, the Kemeny method, and the averages method. The group assessment method allows you to aggregate the opinions of multiple experts, forming a generalized view of the priority of teachers. The pairwise comparison method involves comparing teachers in pairs based on their competencies within a discipline, which helps to identify relative priorities. The Kemeny method, based on finding the median ranking, ensures a consistent and objective distribution of priority, minimizing inconsistencies in expert opinions. The mean method, the most common method, is based on finding the mean value of expert opinions. This study describes the results of a study on the application of these methods to evaluate teachers' priorities by discipline, presenting the calculation algorithms. Depending on the resources available to the educational organization, any of the methods discussed in the paper can be used. A practical example of computing teachers' priority for one subject and a comparative analysis of the results are provided. The obtained results can be useful for educational organizations in planning and organizing the learning process.

Keywords: teacher priority, expert evaluation, group assessment method, pairwise comparison method, teacher ranking

Ж.С. Такенова, 2025.
Халықаралық білім беру корпорациясы, Алматы, Қазақстан.
E-mail: takenova@mail.ru

ПӘНДЕР БОЙЫНША ОҚЫТУШЫЛАРДЫҢ БАСЫМДЫҒЫН БАҒАЛАУҒА АРНАЛҒАН САРАПТАМАЛЫҚ БАҒАЛАУ ӘДІСТЕРІН ЗЕРТТЕУ

Такенова Жанар Сарсенбаевна – экономика ғылымдарының магистрі, Халықаралық білім беру корпорациясы, Алматы, Қазақстан,
E-mail: takenova@mail.ru, ORCID ID: https://orcid.org/0000-0001-8925-5808.

Аннотация. Оқытушылар арасында пәндер бойынша оқу жүктемесін тиімді бөлу үшін оқытушылардың басымдығын пайдалану өзекті міндет болып табылады. Білім беру саласында сараптамалық бағалау әдістері әртүрлі мақсаттарда кеңінен қолданылады, сондықтан, ең алдымен, қарастырылып отырған өзекті мәселені шешу үшін сараптамалық бағалаудың белгілі әдістері таңдалды. Шешім әдістеріне, әдістерді қолдану спектріне және әдістерді қолданудың күрделілігіне талдау жасалды. Сараптамалық бағалау әдістерінің кең ауқымынан пайдалану әдістерінің тобы ерекшеленеді. Мақалада топтық бағалау әдістері, жұптық салыстыру, Кемени әдісі және орташа мән әдісі сияқты пәндер бойынша оқытушылардың басымдылығын бағалау үшін қолдануға ұсынылатын белгілі сараптамалық бағалау әдістері қарастырылған. Топтық бағалау әдісі бірнеше сарапшының пікірлерін біріктіруге мүмкіндік беріп, оқытушылардың басымдығы туралы жинақталған ұғым қалыптастырады. Жұптық салыстыру әдісі пән аясында оқытушыларды олардың құзыреттері бойынша жұптап салыстыруды көздейді, бұл салыстырмалы басымдықтарды анықтауға мүмкіндік береді. Медиандық ранжирлеуді іздеуге негізделген Кемени әдісі сарапшылар пікірлеріндегі қайшылықтарды азайта отырып, басымдықтарды келісімді және объективті бөлуге мүмкіндік береді. Орташа мәндер әдісі сарапшылар пікірлерінің орташа мәнін есептеуге негізделген және ең көп таралған әдіс болып табылады. Жұмыста осы әдістерді пән бойынша оқытушылар басымдығын бағалауда қолдану нәтижелері және оларды есептеу алгоритмдері сипатталған. Білім беру ұйымының ресурстарына байланысты жұмыста қарастырылған әдістердің кез келгені қолданылуы мүмкін. Бір пән бойынша оқытушылар басымдығын есептеу мысалы және нәтижелердің салыстырмалы талдауы ұсынылады. Алынған нәтижелер оқу процесін жоспарлау және ұйымдастыру кезінде білім беру ұйымдары үшін пайдалы болуы мүмкін.

Түйін сөздер: оқытушылар басымдығы, сараптамалық бағалау, топтық бағалау әдісі, жұптық салыстыру әдісі, оқытушыларды ранжирлеу

Ж.С. Такенова, 2025.
Международная образовательная корпорация, Алматы, Казахстан.
E-mail: takenova@mail.ru

ИССЛЕДОВАНИЕ МЕТОДОВ ЭКСПЕРТНЫХ ОЦЕНОК ДЛЯ ОЦЕНКИ ПРИОРИТЕТА ПРЕПОДАВАТЕЛЕЙ ПО ДИСЦИПЛИНАМ

Такенова Жанар Сарсенбаевна – магистр экономических наук, Международная образовательная корпорация, Алматы, Қазақстан,
E-mail: takenova@mail.ru, ORCID ID: https://orcid.org/0000-0001-8925-5808.

Аннотация. Для эффективного распределения учебной нагрузки по дисциплинам между преподавателями актуальной задачей является использование приоритета преподавателей. Наиболее широко используются в сфере образования для различных целей методы экспертных оценок, поэтому в первую очередь для решения рассматриваемой актуальной задачи были выбраны для изучения известные методы экспертных оценок. Проведен анализ методов решения, спектр применения методов и трудоемкость использования методов. Из широкого спектра методов экспертных оценок выделены группа методов для использования. В статье рассмотрены известные методы экспертной оценки, которые представляется использовать для оценки приоритета преподавателей по дисциплинам, такие как методы групповой оценки, парных сравнений, метода Кемени и метод средних значений. Целью работы было сравнить методики применения методов экспертных оценок для решения практических задач. Метод групповой оценки позволяет агрегировать мнения нескольких экспертов, формируя обобщённое представление о приоритете преподавателей. Метод парных сравнений предполагает попарное сопоставление преподавателей по их компетенциям в рамках дисциплины, что позволяет выявить относительные приоритеты. Метод Кемени, основанный на поиске медианного ранжирования, позволяет обеспечить согласованное и объективное распределение приоритета, минимизируя противоречия в экспертных мнениях. Метод средних значений – самый распространённый метод, основан на поиске среднего значения из экспертных мнений. В работе описываются результаты исследования применения этих методов для оценки приоритетов преподавателей по дисциплине, с приведением алгоритмов их расчёта. В зависимости от ресурсов, которыми обладает организация образования, в работе может использоваться любой из рассмотренных методов. Приведён практический пример расчёта оценки приоритета преподавателей по одной дисциплине и сравнительный анализ результатов их расчёта. Полученные результаты могут быть полезны для организаций образования при планировании и организации учебного процесса.

Ключевые слова: приоритет преподавателей, экспертная оценка, метод групповой оценки, метод парных сравнений, ранжирование преподавателей.

Введение. Задача планирования учебной нагрузки преподавателей является актуальной задачей в управлении образовательными бизнес-процессами (Tashev et al., 2023). Эффективное распределение учебных нагрузок преподавателей по дисциплинам зависит от точности определения приоритета преподавателей.

Для определения приоритета преподавателей, основным фактором на который необходимо опираться, является эффективность преподавания по дисциплине. Вопросы оценки эффективности преподавания в университетах рассматриваются в исследовании Berk R.A. (Berk, 2005), где приводится обзор 12 факторов, которые влияют на эффективность преподавания. К ним относятся: рейтинги студентов, мнение коллег, самооценка, видеоматериалы, интервью со студентами, рейтинги данные выпускниками, рейтинги данные работодателями, рейтинги данные администрацией, наличие стипендии за преподавание, наличие наград у преподавателя, педагогическое портфолио преподавателя, показатели результатов обучения. Автор анализирует каждый фактор, предлагая их комбинированное использование для более объективного измерения.

В работе Юревича М.А. (Yurevich, 2013) также представлено исследование факторов, которые влияют на оценки преподавателей в университетах Европы, США и Австралии. Он описывает следующие факторы: студенческие рейтинги, экспертные рейтинги, самооценка, видеоматериалы, опрос студентов, рейтинги выпускников, рейтинги работодателей, рейтинги руководства, преподавательские навыки, награды, результаты обучения, личное дело преподавателя. Как видно, факторы, влияющие на оценку эффективности преподавания у Berk R.A. и Юревича М.А. практически совпадают.

Исследование Исаевой Т., Чурикова М. и Котляренко Ю. (Isayeva et al., 2015) акцентирует внимание на эффективности применения различных методик оценки преподавателей в российских и зарубежных вузах. Авторы делят оценки на количественные – рейтинговые и бальные, а также качественные - экспертные и рецензируемые. Анализируют преимущества и недостатки количественных и качественных подходов, а также затрагивают вопрос восприятия этих оценок самими преподавателями. Делается вывод о необходимости комплексного подхода, учитывающего как объективные показатели, так и субъективные аспекты восприятия деятельности педагога.

Для решения задачи определения приоритета преподавателей можно применить экспертные системы. В работе Кравченко Т.К. (Kravchenko, 2010) рассматривается построение экспертной системы поддержки принятия решений (ЭСППР) на основе иерархической структуры критериев и применения логических правил. Новизна подхода заключается в реализации модульной архитектуры, позволяющей адаптировать систему под различные предметные области. Методология включает этапы построения базы знаний, определения весов критериев и формализации предпочтений, что делает систему универсальной для задач управления.

Анализ существующих работ показывает, что для определения приоритета преподавателей в основном применяются методы экспертных оценок и многокритериальные методы анализа. Описание особенностей применения формализованных методов экспертных оценок в задачах управления процессами представлено в работе Данеляна Т.Я. (Danelyan, 2015). Приводятся этапы экспертного оценивания, а также методы коллективной работы экспертной группы и методы получения индивидуального мнения экспертов. Представлена методика обработки результатов опроса экспертов в разрезе формирования обобщенной оценки, определения относительный весов объектов и установления степени согласованности мнений экспертов.

Tsukida и Gupta (Tsukida et al., 2011) подробно анализируют методы парных сравнений, в частности, модели Брэ́дли–Те́рри и Тюрстоуна. Авторы исследуют статистические аспекты построения матриц предпочтений и проводят сравнительный анализ методов по критериям устойчивости и интерпретируемости результатов. Работа включает практические рекомендации по применению этих методов в анализе социальных и поведенческих данных.

Классическая работа Thurstone L.L. (Thurstone, 2017) представляет формализацию закона сравнительного суждения, где автор вводит количественные меры для субъективных предпочтений. Метод Тюрстоуна, основанный на предположении нормального распределения латентных переменных, до сих пор применяется в психометрии, маркетинге и при построении шкал.

Материалы и методы. В случаях, когда необходимо учитывать множество критериев и альтернатив, при наличии противоречивых мнений экспертов и наличии нечетких оценок, используются многокритериальные методы анализа, такие как AHP (Analytic Hierarchy Process), ELECTRE (ELimination Et Choix Traduisant la REalité) и TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution), которые позволяют ранжировать и согласовывать оценки, обеспечивая более взвешенные решения. Методы являются достаточно сложными в расчетах, особенно при большом числе альтернатив, зависимы от шкалы критериев и часто не учитывают зависимость между критериями.

Одним из известных авторов метода анализа иерархий является Saaty T.L. (Saaty, 2008). Он описывает широко используемый инструмент для принятия решений в условиях многокритериальности. В своей работе автор подробно описывает процесс декомпозиции задачи на иерархические уровни, процедуры парных сравнений и расчёт весов альтернатив. Метод доказал свою эффективность в образовательной сфере, где он применяется для выбора программ, преподавателей и других управленческих решений, требующих экспертных оценок.

В работе Kou G., Lu Y., Peng Y., Shi Y. (Kou et al., 2012) рассматриваются методы многокритериального принятия решений (MCDM) в задаче оценки

алгоритмов классификации. Авторы описывают применение таких подходов, как TOPSIS, ELECTRE и интеграция ранговых корреляций при сравнивании и ранжировании альтернативы по нескольким показателям. Работа актуальна для задач, где требуется объективное агрегирование разнородных критериев, включая системы поддержки принятия решений.

Метод Дельфи, как инструмент экспертного прогнозирования, анализируется в статье Rowe и Wright (Rowe et al., 1999). Авторы исследуют когнитивные и организационные аспекты процедуры, подчеркивая её преимущества — анонимность, итеративность, фокус на консенсус — и ограничения, связанные с рисками предвзятости и слабым контролем качества суждений. Работа представляет собой мета-анализ применения метода Дельфи в различных областях, включая стратегическое планирование и управление.

Интересную интерпретацию метода анализа иерархий представляют Кривулин Н.К. и Сергеев С.Н. (Krivulin et al., 2019), разработавшие реализацию метода принятия решений в процессе аналитической иерархии, основанную на математических операциях. Предложенный подход упрощает вычисления и даёт альтернативную трактовку процедуры агрегации предпочтений. Это позволяет применять AHP в условиях ограниченного ресурса, что особенно актуально для задач с высокими вычислительными затратами.

Практическим вопросам получения и анализа экспертных суждений посвящена монография Meyer M.A. и Booker J.M. (Meyer et al., 1991), в которой систематизируются методы выявления экспертного знания и анализа полученных данных. Авторы приводят классификацию видов неопределённости, обсуждают процедуры отбора экспертов и предоставляют практические рекомендации по агрегированию оценок. Работа является фундаментальной для проектирования процедур экспертных опросов, в том числе в области управления качеством образования и анализа рисков.

Следует отметить, что в ходе изучения работ (которые раскрыты в открытом доступе для исследователей), в которых рассматривалось применение методов экспертных оценок в сфере принятия решения в вопросах планирования деятельности преподавателей, выявлено не так много работ по направлению эффективного планирования учебной нагрузки преподавателей.

В то же время, из опыта работы в сфере управления в организации образования, можно отметить, что широко применяется методика привлечения малых групп экспертов для непосредственного оценивая, где эксперты самостоятельно и независимо друг от друга могут проводить оценивание по широкому кругу вопросов для принятия управленческих решений. Недостатком является, что при сборе экспертных мнений, опираются только на количественные показатели деятельности преподавателя, например – анализ публикационной активности преподавателя, не учитывая качественные показатели, например – уровень владения материалом.

Основными характеристиками компетентности экспертов являются: наличие учёной степени и звания, должность, стаж работы и уровень владения вопросом. Оценки компетентности экспертов по данным мнений экспертов (по их оценкам объектов) приведены в работах (Yevlanov et al., 1978; Mirkin, 1974). Задачи определения компетентности экспертов сведены к определению максимального собственного значения матрицы, полученной из исходной и соответствующего собственного вектора. Полученная от экспертов оценка объектов, то есть оценка приоритета объектов, также сведена к нахождению максимального собственного значения транспонированной матрицы, использованной для определения компетенции экспертов, и соответствующего собственного вектора. В работах приведены итерационные методы получения этих значений, обосновывается сходимость этих итерационных подходов к собственным векторам (Nikaydo, 1972).

Основываясь на опыте принятия управленческих решений в организациях образования рассмотренных в работах (Arici et al., 2022; Takenova, 2024) и набора стандартных ресурсов, имеющихся в организации, анализ которых проведён в работе (Takenova, 2022), разработана математическая модель формирования матрицы приоритетов преподавателей по дисциплинам.

Исследования, которые детализированы в этой статье подтверждают возможность реализации идеи о применении методов экспертных оценок для решения проблемы поиска эффективных методов для оценки приоритета преподавателей по дисциплинам в организациях образавоания. Основными критериями для анализа применяемых методов экспертных оценок выбраны – простота применения, спсектр исходных ресурсов, которыми обладает организация образования, низкие временные и трудовые затраты на реализацию. Согласно этих критериев, выбраны четыре метода экспертных оценок для детального рассмотрения и методики применения. Все методы являются известными методами обработки полученных экспертных мнений – метод групповой оценки, метод парных сравнений, метод Кемени и метод средних значений.

Еще одной целью данной работы было получить математическую формализацию задачи оценки приоритета преподавателей по дисциплинам для дальнейшего решения одним из методов экспертных оценок.

1. Постановка задачи.

Заданы следующие параметры: n – число экспертов; m – количество преподавателей, s – количество дисциплин.

На основании квалификационных требований (следует опираться на квалификационные требования, удовлетворяющие стандартам по уровню, которому принадлежит организация образования), преподаватели могут вести только определенный перечень дисциплин из заданных s, в том числе в разрезе лекций, практических и лабораторных занятий. Соответственно у каждого из m преподавателей формируется доступный ему перечень дисциплин.

Приоритет преподавателей по каждой дисциплине оценивают n экспертов.

Оценку экспертов каждого преподавателя по *d*-ой дисциплине обозначим следующим образом:

$$\mathbf{E^d} = \begin{pmatrix} x_{11}^d & x_{12}^d & \dots & x_{1n}^d \\ x_{21}^d & x_{22}^d & \dots & x_{2n}^d \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1}^d & x_{m2}^d & \dots & x_{mn}^d \end{pmatrix}, d = \overline{1,s},$$

где x_{ij}^d – это результат оценивания *j*-ым экспертом *i*-го преподавателя по *d*-ой дисциплине.

Задача состоит в определении по каждой дисциплине приоритета каждого преподавателя с применением метода экспертных оценок.

Для решения задачи введем обозначения значений приоритетов, где приоритет *j*-го преподавателя по *d*-ой дисциплине обозначим через - p_{dj}. Тогда в итоге получим матрицу приоритетов преподавателей по дисциплине: **P** = $[p_{dj}]$, $d = \overline{1,s}$, $j = \overline{1,m}$.

Вводим ограниеения, что могут принимать только целочисленные значения от «1» до «10». Чем выше значение , тем выше приоритете преподавателя.

Для примера рассмотрим случай, когда экспертами проводится оценивание преподавателей по одной дисциплине, то есть случай, когда *d* = 1.

Тогда полагаем, что $x_{ij}^d = x_{ij}$ и сводим задачу к определению приоритета преподавателей по одной дисциплине и нахождению $\{p_j\}$, $j = \overline{1,m}$. По аналогии, определяя значения приоритетов по каждой дисциплине получаем матрицу приоритетов преподавателей по всем дисциплинам – **P**.

2. Методика определения матрицы приоритетов преподавателей по дисциплинам.

Для решения задачи определения матрицы приоритетов преподавателей по дисциплинам методами экспертной оценки представляется методика, которая состоит из следующих этапов:

1) Дисциплины делятся и собираются на родственные группы

2) Создается группа экспертов по каждой родственной группе дисциплин в количестве 3-4 человека. Эксперты должны иметь педагогический опыт или опыт академической или методической управленческой работы, знания предметной области по выделенной группе дисциплин

3) Формируется портфолио преподавателей и определенный шаблон для выставлениями экспертами своих мнений

4) Для работы экспертов, им выдается портфолио преподавателей с одинаковым набором информации о деятельности преподавателя.

5) Эксперты независимо друг от друга оценивают каждого преподавателя по дисциплине, согласно выданному шаблону. Эксперты заполняют три вида шаблона: Шаблон 1 – путем выставления баллов от «1» до «10», где

«10» является самой высокой оценкой приоритета преподавателя; Шаблон 2 – путем парного сравнения преподавателей друг с другом, с выставлением значений «0», если приоритет первого преподавателя ниже, чем второго, значение «0,5» - если оба преподавателя равны, значение «1» - если приоритет первого преподавателя выше, чем второго; Шаблон 3 – путем ранжирования преподавателей с выставлением рангов, где значение ранга «1» соответствует самому высокому приоритету преподавателя

6) Рассчитывается согласованность мнений экспертов

7) Рассчитывается приоритет каждого преподавателя по каждой дисциплине методом экспертных оценок и таким образов формируется матрица приоритетов преподавателей по всем дисциплинам. Применяются методы обработки экспертных мнений: метод средних значений, метод групповых оценок, метод парных сравнений и метод Кемени.

2. Оценка согласованности мнений экспертов.

Для принятия результатов, полученных методом экспертных оценок, важен этап проверки согласованности мнений экспертов.

Принято (Ivchenko et al., 2010), что оценка согласованности мнений экспертов, когда число экспертов больше двух, определяется на основе коэффициента конкордации.

Задаем через *n* – число экспертов, а через *m* – число преподавателей, тогда коэффициент конкордации будет вычисляться по формуле:

$$W = \frac{12\,S}{n^2(m^3 - m)},$$

$$S = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} x_{ij} - \frac{1}{2}\,n\,(m+1)\right)^2,$$

где x_{ij} – это результат оценивания *j*-ым экспертом *i*-го преподавателя (*j* = 1, *i* = 1).

Коэффициент конкордации равняется единице, если мнения всех экспертов согласованы, и в противном случае, равен нулю. Принято считать (Kendall et al., 1939), что согласованность будет достаточной, если коэффициент конкордации больше, чем *0.5*.

Для определения с какой доверительной вероятностью мнения экспертов являются согласованными, используется критерий Пирсона (Kendall et al., 1939):

$$\chi^2_{cal} = m(n-1)W. \qquad (1)$$

Необходимо найти табличное значение $\chi^2_{(tab,\alpha)}$, взятое для степени свободы *m-1* и которое больше, чем вычисленное значение χ^2_{cal}, по которому определяется уровень значимости α.

Мнения экспертов будем считать хорошо согласованными с доверительной вероятностью = 1 - . Расчет согласованности мнений экспертов на приведен в примере 1.

Пример 1. При заданном коэффициенте W = 0.55, при количестве экспертов n = 3 и количестве преподавателей m = 15 необходимо определить доверительную вероятность согласованности мнений экспертов по критерию Пирсона.

Решение: Вычисляем значение χ^2_{cal} по формуле (1): χ^2_{cal} = 15.4. Находим табличное значение $\chi^2_{(tab,\alpha)}$ для степени свободы равное *14* и больше, чем *15.4*, то есть $\chi^2_{cal} < \chi^2_{(tab,\alpha)}$ (Chi-Square Table, 2025). Имеем согласно данных табличной формы: $\chi^2_{(tab,\alpha)}$ = 23.685 для уровня значимости α = 0.05. Следовательно, считаем, что мнения экспертов хорош согласованы с доверительной вероятностью ρ = 0.95.

4. Методы экспертных оценок и алгоритмы решения для определения приоритете преподавателей по одной дисциплине.

Путем анализа методов экспертных оценок на трудоемкость метода решения, временные затраты на решение и круг требуемых ресурсов для решения, выбраны для рассмотрения четыре метода экспертных оценок: метод средних занчений, метод групповой оценки, метод парных сравнекний и метод Кемени.

Для каждого метода разработан алгоритм решения и реализована программа на Python.

В программе предусмотрен блок для оценки согласованности мнений экспертов, который также предлагает рассмотреть замену какого-либо эксперта, если его мнение не соответствует постановке задачи.

Для удобства обработки данных, в коде программы реализована задача загрузки данных из листов файла формата Excel.

1. Метод средних значений. Метод является наиболее широко используемым при обработке данных на практике в образовательной сфере. Для определения приоритета преподавателей по дисциплине, находим по каждому преподавателю по этой дисциплине среднее значение по всем мнениям экспертов. Чем выше согласованность экспертов, тем точнее будет результат, вычисленный методом средлних значений.

В примере 2, рассмотрена задача с применением метода средних значений.

Пример 2. На кафедре, имеется 7 преподавателей, каждый из которых имеет квалификацию для преподавания определенной дисциплины.

Необходимо определить приоритет преподавателей по этой дисциплине.

Для решения задачи собрана группа из трех экспертов, каждому из которых выдано портфолио преподавателей со следующей информацией: педагогический стаж, академическая и научная квалификация и результаты опросов студентов. Экспертам представлен шаблон, в котором они должны отразить свое мнение по этим 7 преподавателям. Эксперты проводят

оценивание независимо друг от друга, путем выставления баллов от «1» до «10», где «10» - самая высокая оценка данная преподавателю.

Решение задачи методом средних значений. Задаем в виде таблицы 1 мнения экспертов, полученные путем непосредственного оценивания преподавателей.

Таблица 1. – Мнения экспертов по оценке приоритета преподавателей по дисциплине

№	Преподаватели	Эксперт 1 (Э1)	Эксперт 2 (Э2)	Эксперт 3 (Э3)
1	П1	9	8	10
2	П2	7	9	7
3	П3	5	6	7
4	П4	10	4	8
5	П5	1	7	5
6	П6	3	5	4
7	П7	6	4	7

Согласно предложенной методики, далее определяется согласованность мнений экспертов. По результатам вычислений программой, были получены следующие значения параметров согласованности мнений экспертов: коэффициент конкордации W = 0.71; расчетное значение критерия Пирсона χ^2_{cal} = 12.78; табличное значение $\chi^2_{(tab,\alpha)}$ = 16.81 (степень свободы – 6). Табличное значение для заданной степени свободы, больше, чем расчетное значение критерия Пирсона.

Согласно таблице $\chi^2_{(tab,\alpha)}$ = 16.81 соответствует уровню значимости α = 16.81 соответствует уровню значимости α = 0.01 (Chi-Square Table, 2025).

Вывод – мнения экспертов хорошо согласованы с доверительной вероятностью = 0.99.

Следующий блок разработанной программы запускает расчет приоритете преподавателей по дисциплине по методу средних значений. Нормированный результат к шкале «10», полученный в результате вычислений программой представлен в таблице 2.

Таблица 2. – Результат вычислений приоритета преподавателей по дисциплине методом средних значений

№	Преподаватели	Итоговая экспертная оценка по методу средних значений
1	П1	10
2	П2	9
3	П3	7
4	П4	8
5	П5	5
6	П6	4
7	П7	6

По результатам вычислений методом средних значений имеем следующую приоритетность преподавателей по дисциплине: «П1>П2>П4>П3>П7>П5>П6». Самый высокий приоритет у преподавателя под номер 1 (значение приоритета равно – «10») и самый низкий – у преподавателя под номером 6 ((значение приоритета равно – «4»).

1) **Метод групповой оценки**. Согласно методу формируется матрица с результатами оценивания *j*-ым экспертом *i*-го преподавателя по одной дисциплине: **E** = $[x_{ij}]$ ($i = \overline{1,m}$, $j = \overline{1,n}$).

Для определения компетенности экспертов и приоритета преподавателей по дисциплине (Mirkin, 1974), находим матрицы:

$$\mathbf{A} = \mathbf{E}\mathbf{E}^{\mathrm{T}}, \ \mathbf{B} = \mathbf{E}^{\mathrm{T}}\mathbf{E} \qquad (2)$$

Нахождением максимального собственного значения матрицы **B** и соотвествующего собственного вектора $\bar{k} = \{k_j\}$, $j = \overline{1,n}$, определяется компетентность экспертов.

Нахождением максимального собственного значения матрицы **A** и соотвествующего собственного вектора $\bar{p} = \{p_i\}$, $i = \overline{1,m}$, определяется приоритете преподавателей.

Исследования (Kireyev et al., 2004) практических вычислений векторов $\bar{k}$ и $\bar{p}$ показывают, что при большой размерности целесообразно вычисления проводить итерационным способом.

Шаги рекуррентного алгоритма: на шаге $t = 0$, принимается, что $k_j^0 = \frac{1}{n}$, $j = \overline{1,n}$., далее для каждого t= 1,2,3, ….. проводятся вычисления $\bar{p}^t$, λ^t, $\bar{k}^t$ по формулам:

$$\bar{p}^t = \mathbf{E}\bar{k}^{t-1} \qquad (3)$$

$$\bar{k}^t = \frac{1}{\lambda^t}\mathbf{E}^T\bar{p}^t \qquad (4)$$

Имеем , отсюда – нормировочный коэффициент вычисляется как:
$\lambda^t = \sum_{j=1}^{n} K_j^t$.

При $t = 0$ имеем: $\lambda^0 = \sum_{j=1}^{n}\left(\frac{1}{n}\right) = 1$. Тогда согласно (4) получаем:

$$\bar{k}^{t-1} = \frac{1}{\lambda^{t-1}}\mathbf{E}^T\bar{p}^{t-1}. \qquad (5)$$

Подставляя (5) в (3)

$$\bar{p}^t = \frac{1}{\lambda^{t-1}}\mathbf{E}\mathbf{E}^T\bar{p}^{t-1}. \qquad (6)$$

Подставляя (3) в (4)

$$\bar{k}^t = \frac{1}{\lambda^t}\mathbf{E}^T\mathbf{E}\bar{k}^{t-1}. \qquad (7)$$

Учитывая (2), можем записать (6) и (7) в следующем виде:

$$\bar{p}^t = \frac{1}{\lambda^{t-1}}\mathbf{A}\,\bar{p}^{t-1}, \ \bar{k}^t = \frac{1}{\lambda^t}\mathbf{B}\bar{k}^{t-1}$$

Итерации по вычислению проводим до тех пор, пока.

В работе (Yevlanov et al., 1978) показана сходимость данного алгоритма к собственному вектору, соответствующему максимальному собственному числу матрицы **A**.

Блок схема алгоритма вычисления приоритета преподавателей представлена на рисунке 1.

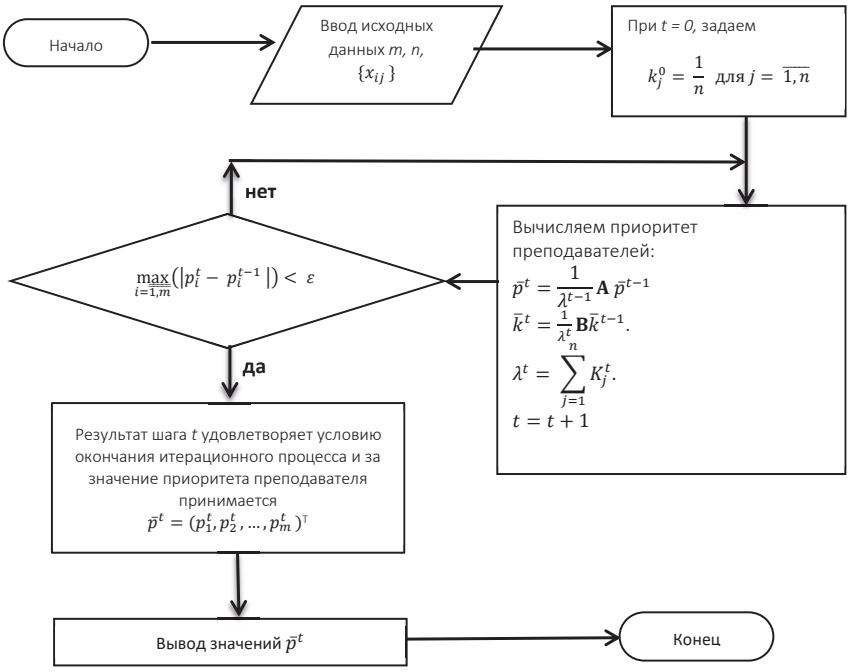


Рисунок 1. – Алгоритма вычисления приоритета преподавателей по дисциплине методом групповой экспертной оценки

На примере 2 рассмотрим решение задачи по определению приоритета преподавателей по дисциплинам с применением метода групповых оценок.

Решение задачи методом групповых оценок. Задачу решаем для данных по мнению экспертов, которые приведены в той же таблице 1.

Согласно рекуррентного алгоритма, представленного на рисунке 1,

получаем результаты вычисления по специально разработанному блоку программы на Python для метода групповых оценок.

Нормированный результат к шкале «10», полученный в результате вычислений программой представлен в таблице 3.

Таблица 3. – Результат вычислений приоритета преподавателей по дисциплине методом групповых оценок

№	Преподаватели	Итоговая экспертная оценка по методу групповой оценки
1	П1	10
2	П2	9
3	П3	7
4	П4	8
5	П5	5
6	П6	4
7	П7	6

По результатам вычислений методом групповых оценок имеем следующую приоритетность преподавателей по дисциплине: «П1>П2>П4>П3>П7>П5>П6». Самый высокий приоритет у преподавателя под номер 1 (значение приоритета равно – «10») и самый низкий – у преподавателя под номером 6 ((значение приоритета равно – «4»).

3) **Метод парных сравнений**. Согласно методу, для n экспертов, задается для сравнения между собой *m* преподаватлей. Через r_{ij}^{h} обозначается результат *h*-ым экспертом по сравнению *i*-го преподавателя с *j*-ым по одной дисциплине:

$$r_{ij}^{h} = \begin{cases} 1 \text{ , если } i - \text{ый преподаватель более значим, чем } j - \text{ый;} \\ 0{,}5 \text{ , если преподаватели } i \text{ и } j \text{ являются равноправными;} \\ 0 \text{ , если } i - \text{ый преподаватель менее значим, чем } j - \text{ый,} \end{cases}$$

где $h = \overline{1, n}$, $i,j = \overline{1, m}$.

Результаты сравнений преподавателей между собой, представляются в виде таблиц парных сравнений по каждому эксперту.

На основании всех сравнений по полученным r_{ij}^{h} стоится матрица математических ожиданий **O** = [o_{ij}]:

$$o_{ij} = M\left[r_{ij}^{h}\right] = 1 * \frac{n_{ij}^{-1}}{n} + 0{,}5 * \frac{n_{ij}^{0}}{n} + 0 * \frac{n_{ij}^{1}}{n} \qquad (8)$$

где $h = \overline{1, n}$, $i,j = \overline{1, m}$, а n_{ij}^{-1} – количество экспертов, которые предпочли *j*-го преподавателя с *i*-му, n_{ij}^{1} - количество экспертов, которые предпочли *i*-го преподавателя с *j*-му, n_{ij}^{0} - количество экспертов, которые приравняли приоритеты *i*-го и *j*-го преподавателя.

Так как общее количество экспертов равно сумме трех возможных мнений экспертов, то есть $n = n_{ij}^{-1} + n_{ij}^{1} + n_{ij}^{0}$, то отсюда можно определить занчение n_{ij}^{0} и подставляя его в (8) получаем значение o_{ij}:

где (9)

При этом.

Далее вычисляется вектор коэффициентов относительной важности (Orlov, 2002), то есть приоритет преподавателей по дисциплине для каждого шага итерации:

$$p_i^t = \frac{1}{\lambda^t} * \mathbf{O} * p_i^{t-1},\ i = \overline{1,m},\ t = 1,2,3, \ldots..$$

Для $t = 0$, задается одинаковое значение приоритета преподавателей: $p_i = 1,\ i = \overline{1,m}.$

Итерация прекращается при

В работе (Kireyev et al., 2004) показана сходимость данного алгоритма вычисления коэффициентов относительной важности преподавателей.

Блок схема алгоритма вычисления приоритета преподавателей методом парных сравнений представлена на рисунке 2.

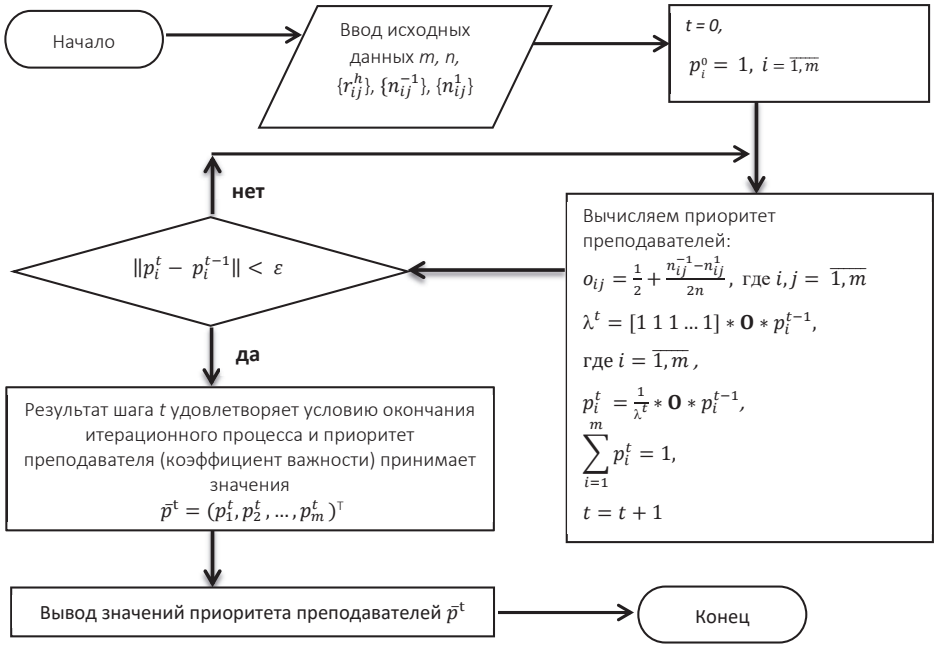


Рисунок 2. – Алгоритм вычисления приоритета преподавателей по дисциплине методом парных сравнений

На примере 2 рассмотрим решение задачи по определению приоритета преподавателей по дисциплинам с применением метода парных сравнений.

Решение задачи методом парных сравнений . Для решения задачи преобразуем данные по мнению экспертов, которые приведены в таблице 1. Согласно метода парных сравнений каждый эксперт попарно сравнивает преподавателей между собой и выставляет значение «0», если приоритет *i*-го преподавателя по дисциплине ниже, чем у *j*-го преподавателя, значение «0.5» - если оба преподавателя равны, значение «1» - если приоритет *i*-го преподавателя по дисциплине выше, чем у *j*-го преподавателя. После преобразования результаты парного сравнения получаем, согласно таблицы 4.

Таблица 4. – Результаты парного сравнения приоритета преподавателей методом парных сравнений

Э1	П1	П2	П3	П4	П5	П6	П7	**Э2**	П1	П2	П3	П4	П5	П6	П7	**Э3**	П1	П2	П3	П4	П5	П6	П7
П1	0.5	1	1	0	1	1	1	П1	0.5	1	1	1	1	1	1	П1	0.5	1	1	1	1	1	1
П2	0	0.5	1	0	1	1	1	П2	1	0.5	1	1	1	1	1	П2	0	0.5	0.5	0	1	1	0.5
П3	0	0	0.5	0	1	1	0	П3	0	0	0.5	1	0	1	1	П3	0	0.5	0.5	0	1	1	0.5
П4	1	1	1	0.5	1	1	1	П4	0	0	0	0.5	0	1	0.5	П4	0	1	1	0.5	1	1	1
П5	0	0	0	0	0.5	0	0	П5	0	0	1	1	0.5	1	1	П5	0	0	0	0	0.5	1	0
П6	0	0	0	0	1	0.5	0	П6	0	0	0	1	0	0.5	1	П6	0	0	0	0	0	0.5	0
П7	0	0	1	0	1	1	0.5	П7	0	0	0	0.5	0	0	0.5	П7	0	0.5	0.5	0	1	1	0.5

Согласно рекуррентного алгоритма, представленного на рисунке 2, получаем результаты вычисления по специально разработанному блоку программы на Python для метода парных сравнений. В коде программы реализована загрузка данных из трех листов файла формта Excel, где каждый лист это результат парного сравнения преподавателей между собой одним экспертом.

Нормированный результат к шкале «10», полученный в результате вычислений программой представлен в таблице 5.

Таблица 5. – Результат вычислений приоритета преподавателей по дисциплине методом парных сравнений

№	**Преподаватели**	**Итоговая экспертная оценка по методу парных сравнений**
1	П1	10
2	П2	9
3	П3	7
4	П4	9
5	П5	6
6	П6	6
7	П7	7

По результатам вычислений методом парных сравнений, имеем следующую приоритетность преподавателей по дисциплине: «П1>П2=П4>П3=П7>П5=П6». Самый высокий приоритет у преподавателя

под номер 1 (значение приоритета равно – «10») и самый низкий – у преподавателей под номером 5 и 6 (значение приоритета равно – «6»).

Итоговое значение приоритета преподавателей, вычисленное по методу парных сравнений, не совпадает с вычислениями по предыдущим двум методам.

4) **Метод Кемени**. Согласно методу, n экспертов проводят ранжирование *m* преподаватлей (Orlov, 2002). Так как мы имеем матрицу **E** = [x_{ij}] ($i = \overline{1,m}$, $j = \overline{1,n}$) по результатам оценивания *j*-ым экспертом *i*-го преподавателя, на основе этой матрицы проведем ранжирование преподавателей и результаты запишем в a_{ij} ($i = \overline{1,m}$, $j = \overline{1,n}$).

Согласно описания метода (Kemeni, 1972), вычисляется величина суммы разностей (по абсолютной величине) мнений двух экспертов по каждому преподаваателю, то есть рассотяние между мнениями *j*-го и *k*-го экспертов, которую обозначим r_{jk} ($j = \overline{1,n}$, $k = \overline{1,n}$).

Далее вычислется расстояние от *j*-го эксперта до остальных - R_j ($j = \overline{1,n}$).

Среди вычисленных суммарных расстояний ищется минимальное, которое и принимается за итоговое мнение экспертов.

Разработан алгоритм вычисления приоритета преподавателей по методу Кемени, представлен в блок-схеме на рисунке 3.

Имеется специально разработанный блок программы на Python для метода Кемени для получения результатов вычислений по итогам ранжирования преподавателей.

На примере 2 рассмотрим решение задачи по определению приоритета преподавателей по дисциплинам с применением метода Кемени.

Решение задачи методом Кемени. Для решения задачи преобразуем данные по мнению экспертов, которые приведены в таблице 1. Проводится ранжирование преподавателей каждым экспертом через проставления значений от «1» до «7», где «1» самый высокий ранг приоритета преподавателя, результаты раскрыты в таблице 6.

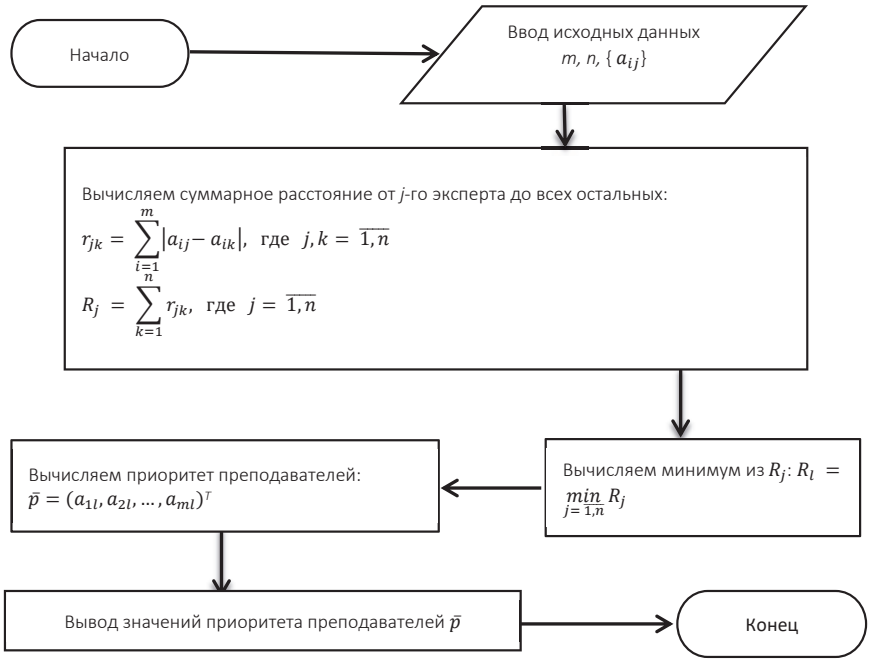


Рисунок 3. – Алгоритм вычисления приоритета преподавателя по дисциплине по методу Кемени

Таблица 6. – Результаты ранжирования преподавателей экспертами

№	Преподаватели	Эксперт 1 (Э1)	Эксперт 2 (Э2)	Эксперт 3 (Э3)
1	П1	2	2	1
2	П2	3	1	3
3	П3	5	4	4
4	П4	1	6	2
5	П5	7	3	6
6	П6	6	5	7
7	П7	4	7	5

Нормированный результат к шкале «10», полученный в результате вычислений программой представлен в таблице 7.

Таблица 7. – Результат вычислений приоритета преподавателей по дисциплине методом Кемени

№	Преподаватели	Итоговая экспертная оценка по методу Кемени
1	П1	10
2	П2	8
3	П3	7
4	П4	9
5	П5	5
6	П6	4
7	П7	6

По результатам вычислений методом Кемени, имеем следующую приоритетность преподавателей по дисциплине: «П1>П4>П2>П3>П7>П5>П6». Самый высокий приоритет у преподавателя под номер 1 (значение приоритета равно – «10») и самый низкий – у преподавателя под номером 6 (значение приоритета равно – «4»).

Итоговое значение приоритета преподавателей, вычисленное по методу Кемени, не совпадает с вычислениями по предыдущим трем методам.

5) Сравнительный анализ результатов исследования. На примере 2 были применены четыре метода экспертных оценок для нахождения приоритете преподавателей: метод средних значений, метод групповых оценок, метод парных сравнений и метод Кемени.

В таблице 8 приведены результаты по приоритетности преподавателей, которые получены в ходе решения задачи по вычислению приоритета преподавателей по дисциплине, указанными методами. А в таблице 9 – значения приоритетов преподавателей.

Таблица 8. – Приоритетность преподавателей выявленная в ходе вычисления методом средних значений, методом групповых оценок, методом парных сравнений и методом Кемени

Метод вычисления приоритета преподавателей по дисциплине	**Приоритетность преподавателя**
Метод средних значений	П1>П2>П4>П3>П7>П5>П6
Метод групповых оценок	П1>П2>П4>П3>П7>П5>П6
Метод парных сравнений	П1>П2=П4>П3=П7>П5=П6
Метод Кемени	П1>П4>П2>П3>П7>П5>П6

Таблица 9. – Значения приоритета преподавателей, вычисленные методом средних значений, методом групповых оценок, методом парных сравнений и методом Кемени

№	**Преподаватели**	**Итоговая экспертная оценка приоритетеа преподавателей**			
		Метод средних значений	**Метод групповых оценок**	**Метод парных сравнений**	**Метод Кемени**
1	П1	10	10	10	10
2	П2	9	9	9	8
3	П3	7	7	7	7
4	П4	8	8	9	9
5	П5	5	5	6	5
6	П6	4	4	6	4
7	П7	6	6	7	6

Результаты вычисленных приоритетов по одной дисциплине рассмотренными методами для задачи примера 2, согласно данных отображенных в таблице 9 раскрыты в виде графического изображения на рисунке 4.

Графики наглядно показывают, что имеется единый тренд – приоритеты преподавателей постепенно снижаются от первого преподавателя (П1) к

шестому преподавателю (П6) с последующим незначительным ростом значений у седьмого преподавателя (П7). Наблюдается минимальный разброс значений между третьим преподавателем (П3) и седьмым (П7). Различия между результатами методов не превышают значения от 0.2 до 0.3, что говорит об общей согласованности мнений экспертов.

Устойчивость экспертных мнений и надежность результатов показывает наблюдаемый тренд для большинства преподавателей – это высокая степень совпадения результатов по методам.

Локальные отклонения получаются при использовании метода парных сравнений, линия показывающая приоритет четвертого (П4) и пятого (П5) преподавателя демонстрирует более высокие значения приоритете по сравнению с другими методами. То есть экспертное мнение очень чувствительно именно при проведении сравнения преподавателей между собой.

Метод Кемени, в отличии от других методов показывает резкий спад на втором преподавателе, что вызвано спецификой медианного агрегирования, которое минимизирует противоречия в исходном ранжировании.

Линии тренда метода средних значений и метода групповых оценок совпадают, что доказывает схожесть принципа средней агрегации экспертных мнений.

Обсуждение. Таким образом, несмотря на имеющиеся локальные различия в линиях тренда, большинство имеющихся точек тренда, показывает наличие стабильного распределения приоритетов преподавателей по дисциплине полученных указанными выше методами.

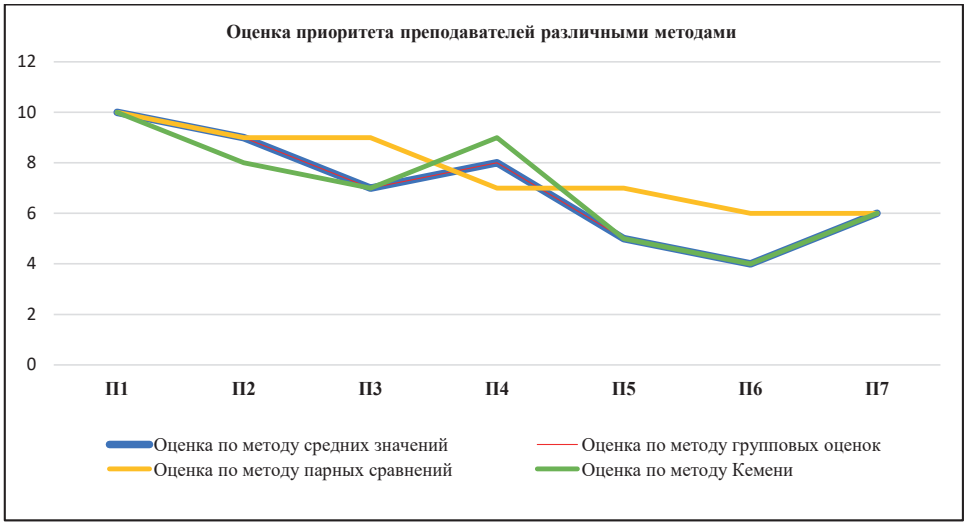


Рисунок 4. – График результатов вычислений приоритетов преподавателей по дисциплине методом средних значений, методом групповых оценок, методом парных сравнений и методом Кемени

Исходя из трудоемкости и ресурсозатратности (по времени, по количеству задействованных лиц и объему подготавливаемой информации), в практическом применении самыми удобными являются метод средних значений и метод групповых оценок. Но необходимо учитывать, что эти методы дают хороший результат только при наличии хорошей согласованности мнений экспертов. То есть при применении этих методов необходимо правильно подбирать экспертную группу.

В итоге решения задачи, определяя пошагово приоритет преподавателей по каждой дисциплине можно составить матрицу приоритетов по преподавателя всей кафедры.

Заключение.

В целях решения поставленной задачи, исследованы существующие методики определения приоритете преподавателей в организациях образования разных стран. Изучен и применен опыт принятия решений в организациях образования Республики Казахстан, что позволило разработать методику определения матрицы приоритетов преподавателей всей кафедры по каждой дисциплине. Полученная, согласно методики и рекомендуемых методов экспертных оценок, матрица приоритетов преподавателей всей кафедры по каждой дисциплине дает в дальнейшем возможность решить актуальную задачу распределения учебной нагрузки между преподавателями.

По результатам проведенных исследований, можно сделать заключения, что методы экспертных оценок жизнеспособны и дают хороший результат для применения в организациях образования. Из множества имеющихся методов экспертной оценки наиболее быстрый результат по затраченному времени на решение, дает метод средних значений, наиболее долгий – метод парных сравнений. Рекомендуется для использования метод групповых оценок и метод Кемени.

Результаты исследований также показали необходимость проверки согласованности мнений экспертов на начальном этапе, что предусмотрено разработанной методикой и продемонстрировано на решении заданного примера 2. Оценка согласованности мнений экспертов основана на расчете коэффициента конкордации и коэффициента Пирсона. Например для заданного примера вывод о хорошей согласованности мнений экспертов сделан с доверительной вероятностью r = 0.99 на результатах расчета, согласно которому: коэффициент конкордации W = 0.71; расчетное значение критерия Пирсона χ^2_{cal} = 12.78; табличное значение $\chi^2_{(tab,\alpha)}$ = 16.81.

Проведен сравнительный анализ результатов вычислений приоритета преподавателей по одной дисциплине для заданного примера 2 методом средних значений, методом групповых оценок, методом парных сравнений и методом Кемени. Сравнительный анализ на графике, представленном на рисунке 4, показывает, что имеется общая тенденция согласованности распределения приоритета при использовании указанных методов. Но в практическом применении в оперативных управленческих решениях наиболее

удобны метод групповых оценок и метод Кемени. Каждый из рассмотренных методов экспертных оценок можкет имспользоваться и для решения других управленческих задач, например, для определения лучшего преподавателя или рейтинега преподователя кафедры и т.д.

При написании статьи использован опыт применении ИИ для проверки переводов аннотаций на казахсий и английский языки, для проверки академического стиля в тексте обзора литературы, для формирования графика сравнительного результата с целью наглядного представления данных, а также в подготовке перечня литературы в требуемом формате (в том числе и транлитерация).

Литература

Tashev A., Takenova Z., Arshidinova M. (2023) Algorithms for Solving Problems of Resources Allocation in the Management of Business Processes in Educational Organizations. International Journal of Modern Education and Computer Science, no 15(5). — P. 14-27.

Berk R.A. (2005) Survey of 12 Strategies to Measure Teaching Effectiveness. International Journal of Teaching and Learning in Higher Education, no 17(1). — P. 48-62.

Юревич М.А. (2013) Методики оценки педагогических кадров в высшей школе в Европе, США и Австралии. Образовательные технологии, №2. —С.104-115

Исаева Т., Чуриков М., Котляренко Ю. (2015) Эффективность оценивания деятельности преподавателей вузов: сравнение отечественных и зарубежных методик. Интернет-журнал «Науковедение», №7(3)

Кравченко Т.К. (2010) Экспертная система поддержки принятия решений. Вестник Томского государственного университета, № 6. —С.147-156

Данелян Т.Я. (2015) Формальные методы экспертных оценок. Государственное управление. Электронный вестник, №1. —С.183-187

Tsukida K., Gupta M.R. (2011) How to analyze paired comparison data. UWEE Technical report: UEEETR-2011-0004. — P. 24.

Thurstone L.L. (2017) A law of comparative judgment. Scaling: A Sourcebook for Behavioral Scientists. — P. 81-92.

Saaty T.L. (2008) Decision making with the analytic hierarchy process. Int. J. Services Sciences, no 1(1). — P. 83-98.

Kou G., Lu Y., Peng Y., Shi Y. (2012) Evaluation of classification algorithms using MCDM and rank correlation. Int. J. Information Technology & Decision Making, no 11(1). — P. 197-225.

Rowe G., Wright G. (1999) The Delphi Technique as a Forecasting Tool: Issues and Analysis. International Journal of Forecasting, no 15(4). —P. 353–375.

Krivulin N.K., Sergeev S.N. (2019) Tropical implementation of the Analytical Hierarchy Process decision method. Fuzzy Sets and Systems, no 377. — P. 31–51.

Meyer, M.A., Booker, J.M. (1991) Eliciting and Analyzing Expert Judgment: A Practical Guide. Academic Press: ASA-SIAM Series on Statistics and Applied Mathematics. — P. 441.

Евланов Л.Г., Кутузов В.А. (1978) Экспертные оценки в управлении. Москва: Экономика. —С. 88.

Миркин Б.Г. (1974) Проблема группового выбора. Москва: Наука. — С. 256

Никайдо Х. (1972) Выпуклые структуры и математическая экономика. Мир

Arici M., Такенова Ж. (2022) О некоторых вопросах распределения ресурсов при управлении сложными процессами. Advanced technologies and computer science, № 1(3). — С..29-38

Такенова Ж. (2024) Новые подходы в решении управленческих задач в организациях образования. Известия НАН РК. Серия физико-математическая, № 1. — С. 368-284

Такенова Ж. (2022) Вопросы формирования педагогической нагрузки в высшем учебном заведении. Сборник материалов международной научно-практической конференции «Современные тренды в архитектуре и строительстве: энергоэффективность, энергосбережение, BIM технологии, проблемы городской среды» по направлению «Инновационные тренды в современном высшем образовании». — С. 376-385

Ивченко Г.И., Медведев Ю.И. (2010) Введение в математическую статистику. Москва: Издательство ЛКИ. — С. 600

Kendall M.G., Babington Smith B. (1939) The Problem of m Rankings. The Annals of Mathematical Statistics. London and University of St. Andrews. Scotland, no 10(3). — P. 275-287.

Chi-Square Table. (n.d.). In Social Science Statistics. Retrieved June 9, 2025, from https://www.socscistatistics.com/tests/chisquare2/

Киреев В. И., Пантелеев А.Б. (2004) Численные методы в примерах и задачах. Учебное пособие. — С. 480.

Орлов А.И. (2002) Экспертные оценки. Учебное пособие. — С. 60

Кемени Дж., Снелл Дж. (1972) Кибернетическое моделирование: Некоторые приложения. Москва: Советское Радио. — С. 192.

References

Tashev A., Takenova Z., Arshidinova M. (2023) Algorithms for Solving Problems of Resources Allocation in the Management of Business Processes in Educational Organizations. International Journal of Modern Education and Computer Science, no 15(5). — P. 14-27 (in English)

Berk R.A. (2005) Survey of 12 Strategies to Measure Teaching Effectiveness. International Journal of Teaching and Learning in Higher Education, no 17(1). — P. 48-62 (in English)

Yurevich M.A. (2013) Metodiki otsenki pedagogicheskikh kadrov v vysshey shkole v Yevrope, SSHA i Avstralii [Methods of assessing teaching staff in higher education in Europe, the USA and Australia]. Obrazovatel'nyye tekhnologii, no 2. — P. 104-115 (in Russian)

Isayeva T., Churikov M., Kotlyarenko YU. (2015) Effektivnost' otsenivaniya deyatel'nosti prepodavateley vuzov: sravneniye otechestvennykh i zarubezhnykh metodik [Efficiency of assessing the activities of university teachers: comparison of domestic and foreign methods]. Internet-zhurnal «Naukovedeniye», no 7(3) (in Russian)

Kravchenko T.K. (2010) Ekspertnaya sistema podderzhki prinyatiya resheniy [Expert decision support system]. Vestnik Tomskogo gosudarstvennogo universiteta, no 6. — P. 147-156 (in Russian)

Danelyan T.YA. (2015) Formal'nyye metody ekspertnykh otsenok [Formal methods of expert assessments]. Gosudarstvennoye upravleniye. Elektronnyy vestnik, no 1. — P. 183-187 (in Russian)

Tsukida K., Gupta M.R. (2011) How to analyze paired comparison data. UWEE Technical report: UEEETR-2011-0004. — P. 24 (in English)

Thurstone L.L. (2017) A law of comparative judgment. Scaling: A Sourcebook for Behavioral Scientists. — P. 81-92 (in English)

Saaty T.L. (2008) Decision making with the analytic hierarchy process. Int. J. Services Sciences, no 1(1). — P. 83-98 (in English)

Kou G., Lu Y., Peng Y., Shi Y. (2012) Evaluation of classification algorithms using MCDM and rank correlation. Int. J. Information Technology & Decision Making, no 11(1). — P. 197-225 (in English)

Rowe G., Wright G. (1999) The Delphi Technique as a Forecasting Tool: Issues and Analysis. International Journal of Forecasting, no 15(4). — P. 353–375 p. (in English)

Krivulin N. K., Sergeev S. N. (2019) Tropical implementation of the Analytical Hierarchy Process decision method. Fuzzy Sets and Systems, no 377. — P. 31–51 (in English)

Meyer, M.A., Booker, J.M. (1991) Eliciting and Analyzing Expert Judgment: A Practical Guide. Academic Press: ASA-SIAM Series on Statistics and Applied Mathematics. — 441 (in English)

Yevlanov L.G., Kutuzov V.A. (1978) Ekspertnyye otsenki v upravlenii [Expert assessments in management]. Moscow: Economics. — P. 88 (in Russian)

Mirkin B.G. (1974) Problema gruppovogo vybora [The problem of group choice]. Moscow: Science. — P. 256 (in Russian)

Nikaydo KH. (1972) Vypuklyye struktury i matematicheskaya ekonomika [Convex structures and mathematical economics]. World (in Russian)

Arici M., Takenova ZH. (2022) O nekotorykh voprosakh raspredeleniya resursov pri upravlenii slozhnymi protsessami [On some issues of resource allocation in managing complex processes]. Advanced technologies and computer science, no 1(3). — P. 29-38 (in Russian)

Takenova ZH. (2024) Novyye podkhody v reshenii upravlencheskikh zadach v organizatsiyakh obrazovaniya [New approaches to solving management problems in educational organizations]. Izvestiya NAN RK. Seriya fiziko-matematicheskaya, no 1. — P. 368-284 (in Russian)

Takenova ZH. (2022) Voprosy formirovaniya pedagogicheskoy nagruzki v vysshem uchebnom

zavedenii [Issues of formation of pedagogical load in higher educational institution]. Sbornik materialov mezhdunarodnoy nauchno-prakticheskoy konferentsii «Sovremennyye trendy v arkhitekture i stroitel'stve: energoeffektivnost', energosberezheniye, BIM tekhnologii, problemy gorodskoy sredy» po napravleniyu «Innovatsionnyye trendy v sovremennom vysshem obrazovanii». — P. 376-385 (in Russian)

Ivchenko G.I., Medvedev YU.I. (2010) Vvedeniye v matematicheskuyu statistiku [Introduction to Mathematical Statistics.]. Moscow: LKI Publishing House. — P. 600 (in Russian)

Kendall M.G., Babington Smith B. (1939) The Problem of m Rankings. The Annals of Mathematical Statistics. London and University of St. Andrews. Scotland, no 10(3). — P. 275-287 (in English)

Chi-Square Table. (n.d.). In Social Science Statistics. Retrieved June 9, 2025, from https://www.socscistatistics.com/tests/chisquare2/ (in English)

Kireyev V.I., Panteleyev A.B. (2004) Chislennyye metody v primerakh i zadachakh [Numerical methods in examples and problems]. Study manual. — P.480 (in Russian)

Orlov A.I. (2002) Ekspertnyye otsenki [Expert assessments]. Study manual. — P. 60 (in Russian)

Kemeni Dzh., Snell Dzh. (1972) Kiberneticheskoye modelirovaniye: Nekotoryye prilozheniya [Cybernetic Modeling: Some Applications]. Moscow: Soviet Radio. — P. 192 (in Russian)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 228–242

https://doi.org/10.32014/2025.2518-1726.374

МРНТИ 27.47.19
УДК 512.647

Zh. Tashenova, A.R. Gabdullin, Zh. Abdugulova, Sh. Amanzholova, E. Nurlybaeva, 2025.
Department of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
E-mail: zhuldyz_tm@mail.ru

ANALYSIS OF MODERN WIRELESS NETWORK SECURITY PROTOCOLS AND PROSPECTS FOR THEIR DEVELOPMENT

Tashenova Zh. — PhD, Department of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: zhuldyz_tm@mail.ru, https://orcid.org/0000-0003-3051-1605;
Gabdullin A. — Master of Information Security Systems, Department of Information Security System, Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: anchorite.exe@gmail.com, https://orcid.org/0000-0003-3051-1605;
Abdugulova Zh. — Associated Professor, Department of Information Technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
E-mail: janat_6767@mail.ru, https://orcid.org/0000-0001-7462-4623;
Amanzholova Sh. — PhD, Kurmangazy Kazakh National Conservatory, Almaty, Kazakhstan,
E-mail: schirin75@mail.ru;
Nurlybaeva E. — PhD, Department of Information Technologies, The Kazakh National Academy of Arts named after T. Zhurgenova, Almaty, Kazakhstan,
E-mail: nuremek@mail.ru, https://orcid.org/0000-0003-3051-1605.

Abstract. Modern wireless networks rely on robust security protocols for data protection and offering secure connectivity. In this paper, we address the weaknesses and strengths of WPA2 (Wi-Fi Protected Access II) and its successor WPA3 (Wi-Fi Protected Access III), and examine prospects for their future development. We summarize authentication mechanisms of the protocols (including the SAE handshake of WPA3) and examine their resistance to popular attack vectors such as handshake capture, deauthentication, and the KRACK (Key Reinstallation Attack) vulnerability. Our results demonstrate that WPA3 eliminates a number of WPA2 weaknesses by neutralizing these common attacks: the improved handshake and mandatory protections of WPA3 entirely thwart the use of captured handshakes for offline cracking and significantly reduce exposure to deauthentication and key reinstallation attacks. There are, however, some open issues requiring further improvement in the protocols to counter emerging threats. These findings underscore

the imperative of universal WPA3 adoption and ongoing protocol improvements to deliver strong, future-resistant wireless network security. This article presents a comprehensive analysis of modern wireless network security protocols, focusing on their architecture, functionality, and resistance to contemporary cyber threats. The study examines widely used standards, including WPA3, TLS-based mechanisms, and emerging encryption approaches, highlighting their strengths and existing vulnerabilities.

Keywords: Wireless Networks, WPA3, WPA2, Security Protocols, Cyber Threats

Ж.М. Ташенова, А.Р. Габдуллин, Ж.К. Абдугулова, Ш.А. Аманжолова, Э.Н. Нурлыбаева, 2025.
Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан.
E-mail: zhuldyz_tm@mail.ru

ЗАМАНАУИ СЫМСЫЗ ЖЕЛІНІҢ ҚАУІПСІЗДІК ХАТТАМАЛАРЫН ТАЛДАУ ЖӘНЕ ОЛАРДЫҢ ДАМУ ПЕРСПЕКТИВАЛАРЫ

Ташенова Ж.М. — PhD, Ақпараттық технологиялар факультеті, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: zhuldyz_tm@mail.ru, https://orcid.org/0000-0003-3051-1605;
Габдуллин А.Р. — ақпараттық қауіпсіздік жүйелері магистрі, Ақпараттық қауіпсіздік жүйелері кафедрасы, Ақпараттық технологиялар факультеті, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: anchorite.exe@gmail.com, https://orcid.org/0000-0003-3051-1605;
Абдугулова Ж.К. — экономика ғылымдарының кандидаты, қауымдастырылған профессор, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, ақпараттық технологиялар факультеті, Астана, Қазақстан,
E-mail: janat_6767@mail.ru, https://orcid.org/0000-0001-7462-4623;
Аманжолова Ш.А. — PhD, Құрманғазы атындағы Қазақ ұлттық консерваториясы, Алматы, Қазақстан,
E-mail: schirin75@mail.ru, https:// orcid.org/0000-0002-6674-2766;
Нұрлыбаева Э.Н. — PhD, Т.Жүргенова атындағы Қазақ ұлттық өнер академиясы, ақпараттық технологиялар кафедрасы, Алматы, Қазақстан,
E-mail: nuremek@mail.ru, https://orcid.org/0000-0003-3051-1605.

Аннотация. Қазіргі заманғы сымсыз желілер деректерді қорғау мен қауіпсіз байланысты қамтамасыз ету үшін сенімді қауіпсіздік хаттамаларына сүйенеді. Бұл мақалада WPA2 (Wi-Fi Protected Access II) және оның мұрагері WPA3 (Wi-Fi Protected Access III) хаттамаларының әлсіз және күшті жақтары қарастырылып, олардың болашақтағы даму перспективалары талданады. Протоколдардың аутентификация механизмдері (соның ішінде WPA3-тегі SAE қол алысуы) сипатталып, оларды жиі кездесетін шабуыл түрлеріне – қол алысуды ұстап қалу, деаутентификация және KRACK (Key Reinstallation Attack) осалдығына қарсы төзімділігі зерттеледі. Зерттеу нәтижелері WPA3 нұсқасының WPA2-дегі бірқатар осал тұстарды жойып, қол алысуды жақсарту және міндетті қорғаныс тетіктері арқылы оффлайн-құпиясөзді бұзу әрекеттерін

толықтай болдырмайтынын, сондай-ақ деаутентификация мен кілтті қайта орнату шабуылдарына қарсы әлдеқайда тиімді қорғаныс беретінін көрсетті. Дегенмен, жаңа қауіптерге қарсы тұру үшін әлі де жетілдіруді қажет ететін мәселелер бар. Бұл тұжырымдар WPA3 хаттамасын әмбебап енгізу мен оның үздіксіз жетілдірілуінің заманауи сымсыз желілердің қауіпсіздігін қамтамасыз етуде шешуші маңызға ие екенін айқындайды. Мақалада заманауи сымсыз желілердің қауіпсіздік хаттамалары олардың архитектурасы, функционалдығы және киберқауіптерге төзімділігі тұрғысынан жан-жақты талданады. Сондай-ақ WPA3, TLS негізіндегі тетіктер мен жаңа шифрлау әдістері секілді кеңінен қолданылатын стандарттардың артықшылықтары мен осал тұстары көрсетіледі. Сонымен қатар, мақалада қауіпсіздік технологияларының даму үрдістері қарастырылып, постквантты криптографияның, нөлдік сенім үлгілерінің және жасанды интеллектке негізделген шешімдердің рөлі атап көрсетіледі. Талдау нәтижелері сымсыз желілердің қауіпсіздігін болашақта дамыту үшін технологиялық инновацияларды, нормативтік шараларды және бейімделгіш қауіптерге қарсы әрекет ету стратегияларын біріктіретін кешенді тәсілді қажет ететінін көрсетеді.

Түйін сөздер: сымсыз желілер, WPA3, WPA2, Қауіпсіздік хаттамалары, Киберқауіптер

Ж.М. Ташенова, А.Р. Габдуллин, Ж.К. Абдугулова, Ш.А.Аманжолова, Э.Н. Нурлыбаева, 2025.

Евразийский национальный университет им. Л.Н. Гумилёва,
Астана, Казахстан.
E-mail: zhuldyz_tm@mail.ru

АНАЛИЗ СОВРЕМЕННЫХ ПРОТОКОЛОВ БЕЗОПАСНОСТИ БЕСПРОВОДНЫХ СЕТЕЙ И ПЕРСПЕКТИВЫ ИХ РАЗВИТИЯ

Ташенова Ж.М. — PhD, факультет информационных технологий, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Республика Казахстан,
E-mail: zhuldyz_tm@mail.ru, https://orcid.org/0000-0003-3051-1605;
Габдуллин А.Р. — магистр по системам информационной безопасности, кафедра систем информационной безопасности, факультет информационных технологий, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан,
E-mail: anchorite.exe@gmail.com, https://orcid.org/0000-0003-3051-1605;
Абдугулова Ж.К. — доцент факультета информационных технологий Евразийского национального университета им. Л.Н. Гумилева, Астана, Казахстан,
E-mail: janat_6767@mail.ru, https://orcid.org/0000-0001-7462-4623;
Аманжолова Ш.А. — PhD, Казахская национальная консерватория им. Курмангазы, Алматы, Казахстан,
E-mail: schirin75@mail.ru, https:// orcid.org/0000-0002-6674-2766;
Нурлыбаева Э.Н. — PhD, Казахская национальная академия искусств им. Т. Жургеновой, кафедра информационных технологий, Алматы, Казахстан,
E-mail: nuremek@mail.ru, https://orcid.org/0000-0003-3051-1605.

Аннотация. Современные беспроводные сети полагаются на надежные протоколы безопасности для защиты данных и обеспечения безопасного подключения. В этой статье мы рассмотрим слабые и сильные стороны WPA2 (Wi-Fi Protected Access II) и его преемника WPA3 (Wi-Fi Protected Access III), а также рассмотрим перспективы их будущего развития. Мы суммируем механизмы аутентификации протоколов (включая рукопожатие SAE WPA3) и исследуем их устойчивость к популярным векторам атак, таким как захват рукопожатия, деаутентификация и уязвимость KRACK (атака переустановки ключа). Наши результаты показывают, что WPA3 устраняет ряд слабых сторон WPA2, нейтрализуя эти распространенные атаки: улучшенное рукопожатие и обязательная защита WPA3 полностью пресекают использование захваченных рукопожатий для офлайн-взлома и значительно снижают подверженность атакам деаутентификации и переустановки ключа. Однако есть некоторые открытые вопросы, требующие дальнейшего улучшения протоколов для противодействия новым угрозам. Эти результаты подчеркивают необходимость всеобщего принятия WPA3 и постоянного совершенствования протокола для обеспечения надежной и устойчивой к будущим изменениям безопасности беспроводных сетей. Исследуются широко используемые стандарты, включая WPA3, механизмы на основе TLS и новые методы шифрования, с выделением их преимуществ и существующих уязвимостей. Особое внимание уделяется проблемам обеспечения конфиденциальности, целостности и аутентификации в динамичных беспроводных средах. Кроме того, в статье рассматриваются актуальные тенденции развития технологий безопасности, подчеркивается роль постквантовой криптографии, моделей «нулевого доверия» и решений на основе искусственного интеллекта.

Ключевые слова: беспроводные сети, WPA3, WPA2, протоколы безопасности, киберугрозы

Introduction. Wi-Fi is a ubiquitous part of our lives nowadays. Wi-Fi covers nearly all locations, whether homes, offices, or public hotspots. Wi-Fi's security aspects have changed and have evolved enormously over 20 years. All this development came based on various pivotal standards that were introduced to make Wi-Fi secure and reliable. Initial standards like WEP were proven to have design flaws, and thus Wi-Fi Protected Access (WPA) and most popularly used WPA2 (IEEE 802.11i) were introduced during 2004. The newest, Wi-Fi Alliance, debuted WPA3 as a successor to WPA2 back in 2018 to make WLANs' encryption and validation more secure. All these modern 802.11 security protocols (WPA2, WPA3) were designed to offer confidentiality, integrity, and access control for wireless networks, and these are currently available as the default settings for personal routers and business Wi-Fi deployments (Halbouni, Ong, & Leow, 2023; Lounis & Zulkernine, 2020).

Literature Review

Wi-Fi Protected Access II (WPA2) is the dominant WLAN security protocol for well over a decade, providing strong encryption through AES-CCMP. Several researches have, nonetheless, unveiled serious flaws in WPA2's design. Among these, there is the Key Reinstallation Attack (KRACK), which utilizes a flaw in the four-way handshake to cause a nonce reusal and decrypt traffic without knowing Wi-Fi's password (Vanhoef & Ronen, 2020). Similarly, an attacker can capture a WPA2 handshake-derived value (the PMKID) to perform offline dictionary attacks, bypassing the need to intercept the full 4-way exchange (De Almeida Braga, Fouque, & Sabt, 2020). These revelations highlight that even strong ciphers can be undermined by protocol logic errors.

Moreover, WPA2 networks are vulnerable to denial-of-service (DoS) tactics due to unprotected management frames: malicious deauthentication and disassociation packets can be injected to knock clients off a network at will (Chatzoglou, Kambourakis, & Kolias, 2022; Gebresilassie et al., 2023). In practice, weak pre-shared keys also remain an Achilles' heel of WPA2, as attackers can readily crack poorly chosen passwords through offline guessing (Banakh et al., 2024; De Almeida Braga et al., 2020). The accumulation of such vulnerabilities ultimately motivated the development of WPA3 to fortify Wi-Fi security against these exploits.

The introduction of WPA3 brought important enhancements intended to address WPA2's shortcomings. Notably, WPA3-Personal replaces the pre-shared key exchange with the Simultaneous Authentication of Equals (SAE) handshake, a variant of the Dragonfly key exchange, to provide forward secrecy and better resistance to offline password guessing (Lounis & Zulkernine, 2019). WPA3 also mandates Protected Management Frames (PMFs) to defend against deauthentication spoofing and introduces individualized data encryption even on open networks through Opportunistic Wireless Encryption (OWE) (Halbouni, Ong, & Leow, 2023; Lounis & Zulkernine, 2020). These improvements raised expectations that WPA3 would resolve the prevalent issues in WPA2.

Yet early analyses of WPA3 have shown that it is not immune to vulnerabilities. Security researchers discovered design and implementation flaws in WPA3 shortly after its release. For instance, the Dragonblood study uncovered a suite of attacks that included handshake downgrades, side-channel leaks, and denial-of-service exploits (Vanhoef & Ronen, 2020). These findings revealed that an attacker could undermine WPA3's SAE handshake—gaining the ability to run offline dictionary attacks or even overload access points with excessive processing requests—despite the protocol's new protections. Additionally, prior cryptanalysis of the Dragonfly handshake underlying SAE had identified potential weaknesses in certain parameter choices, hinting at the challenges in balancing usability with cryptographic rigor (De Almeida Braga et al., 2020; Lounis & Zulkernine, 2019).

Attackers often leverage the above protocol weaknesses through well-known Wi-Fi attack techniques. One such threat is the Evil Twin attack, wherein a rogue

access point impersonates a legitimate Wi-Fi network to lure victims into connecting (Shrivastava, Kumar, & Kataoka, 2020). By duplicating a trusted network's SSID and settings, an adversary can perform man-in-the-middle interception once clients unknowingly join the fake hotspot (Banakh et al., 2024; Chatzoglou et al., 2022). Evil Twin attacks are frequently coupled with deauthentication floods: the attacker forcefully disconnects users from the genuine AP, prompting them to reconnect—often to the stronger malicious signal (Gebresilassie et al., 2023; Shrivastava et al., 2020).

Deauthentication and related denial-of-service (DoS) attacks exploit the fact that, under WPA2, management frames are not authenticated, allowing any device to broadcast spoofed disconnection commands (Gebresilassie et al., 2023; Schepers, Ranganathan, & Vanhoef, 2022). The result is a simple but effective DoS that can disrupt service or facilitate further exploits like Evil Twin man-in-the-middle hijacking. Beyond spoofed frames, adversaries can also launch DoS attacks at the physical layer (jamming the Wi-Fi spectrum) or via resource exhaustion (flooding the network), rendering the channel unusable (Marais, Coetzee, & Blauw, 2021). These attack techniques demonstrate how weaknesses in Wi-Fi's protocol layers are actively exploited in practice, emphasizing that improvements in standards (e.g., WPA3's PMF to counter deauth) must be complemented by vigilance against a range of attack vectors.

Research Objectives and Tasks

Despite continuous improvements, current wireless security protocols still suffer from serious vulnerabilities that expose users and organizations to attacks. Given these persistent weaknesses and the fast-evolving tactics of attackers, there is a clear need for ongoing evaluation of wireless security protocols under real-world conditions.

In this work, we take a practical approach to assess the resilience of modern Wi-Fi security standards. We have built a dedicated wireless security testbed that simulates a typical network environment and allows controlled execution of various attacks (including Evil Twin setups, deauthentication floods, handshake interception, etc.) against WPA2- and WPA3-protected networks. By performing our own independent experiments, we can verify the severity of known weaknesses and observe how effectively the protocols' defenses hold up outside of theoretical analysis or vendor claims.

Through this hands-on evaluation, we present a synthesized analysis of the strengths and weaknesses of WPA2 and WPA3, and we offer insights into their suitability for different use cases. In particular, this study:

1. Identifies which known vulnerabilities remain applicable (or have been mitigated) in real deployments of WPA2 vs. WPA3 (Chatzoglou et al., 2022; Gao et al., 2021).

2. Pinpoints security gaps where further improvements or best practices are needed (for instance, in handling rogue AP threats or ensuring robust user authentication) (Gebresilassie et al., 2023; Schepers et al., 2022).

3. Provides guidance on selecting and configuring Wi-Fi security protocols for distinct contexts—from end-user home networks to large enterprise infrastructures—in light of their current security posture (Halbouni et al., 2023; Lounis & Zulkernine, 2019).

By emphasizing the practical significance of new protocol advances, this research seeks to make both researchers and practitioners of networks aware of state-of-the-art wireless security and how it is likely to evolve in the future. Ultimately, the purpose of this study lies in explaining how far advanced Wi-Fi security is and what there is still to be done to point toward and enable wiser, more robust future wireless security standards.

Materials and Methods

The testbed used for this research is illustrated in Figure 1. The setup consists of an off-the-shelf Wi-Fi router supporting both WPA2-PSK and WPA3-SAE (Personal mode), a victim client device, and an attacker's laptop running Kali Linux with an Alfa AWUS036ACH USB Wi-Fi adapter in monitor mode.

A wired control server was connected to the router's LAN to generate traffic and log attack impacts. The testbed simulates a small-office/home network, ensuring realistic conditions while allowing controlled execution of attacks.

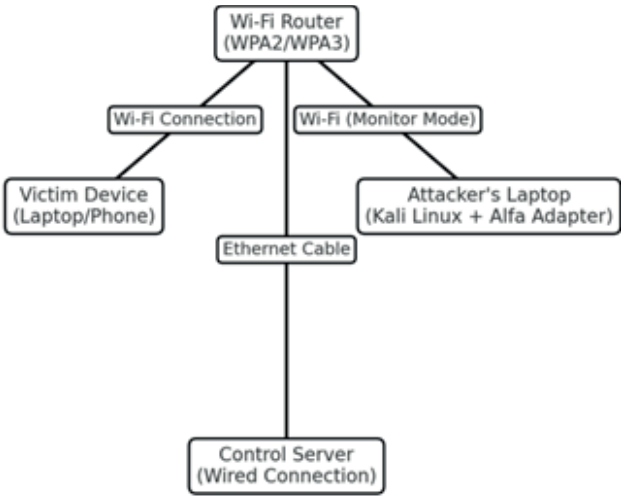


Figure 1. Diagram of the testbed setup used for hands-on experiments

Research Procedures & Attack Scenarios

We tested the resilience of WPA2 and WPA3 security protocols under practical conditions by applying a systematic penetration testing approach with Kali Linux as the basis for the test bed. We tested for four principal attack vectors for Wi-Fi exploitation: (1) Deauthentication Attack (Deauth), (2) Continuous Deauthentication as a Denial-of-Service (DoS), (3) Handshake Capture for Offline Attacks, and (4) Key Reinstallation Attack (KRACK) solely for WPA2.

Each of these types of attacks was performed within a controlled laboratory setting against both WPA2- and WPA3-secured networks sequentially to have a consistent basis for comparison. A five-step process was employed for all phases of testing to preserve methodological rigor and reproducibility. In a first step, baseline measurements of usual network operation were made to define standard performance benchmarks. Then, the specific attack script was carried out against the test network. All wireless traffic and protocol-specific packets were collected during attack instances with Wireshark and tcpdump for later examination. Then, the impact of each attack was determined based upon principal indicators such as packet loss, disconnection time, and recovery behavior of the client. Lastly, WPA2 and WPA3 results were comparatively examined to find differences between security resilience and robustness of protocol.

We repeated each attack scenario five times under the same conditions to guarantee statistical accuracy of results. Through this repetition, we could average the results obtained and look out for consistent trends or deviations between test runs to validate the experimental results.

Attack Implementation & Algorithms. Deauthentication Attack

A set of hands-on attack implementations were performed to test WPA2 and WPA3 protections with standardized procedures and tools under the testbed setup. The first attack involved a deauthentication attack, whereby an attack is performed by sending spoofed IEEE 802.11 deauthentication frames to disconnect a client forcefully from the network. As there is no protection for management frames provided by WPA2, these frames can be injected by any device with range. The attack process was conducted by putting it into monitor mode for the attack's wireless adapter, capturing the MAC address of the victim through Wireshark or airodump-ng, and sending forged deauthentication frames through aireplay-ng. As predicted, WPA2 clients got automatically disconnected as soon as spoofed packets were received, while WPA3 clients, which have Protected Management Frames (PMF) with them, remained unaffected because of the requirement for authentication for such frames.

The second attack model was a Denial-of-Service (DoS) attack through constant deauthentication, which extends the simple deauthentication attack by continually sending deauth frames with high frequency to disallow reassociation of the client. It was carried out with a loop script sending packets around 0.1 seconds apart. On WPA2 networks, this attack was found to cause heavy service disruption, with as high as 95% observed packet loss and almost total disconnection of the client. On WPA3 networks with activated PMF, there was no perceivable impact, as rogue management frames were well-filtered and discarded.

The third attack targeted handshake capture and offline cracking of passwords, with an attack on the WPA2-Personal process of authentication. The reconnection was forced with a deauth frame, and the resultant EAPOL 4-way handshake was captured with the help of packet analyzers. The handshake was pulled out with

aircrack-ng, and an offline dictionary attack was performed to try and retrieve the pre-shared key (PSK). For WPA2, this attack was successful with all attempts that had the password as part of the dictionary. For WPA3, based on the Simultaneous Authentication of Equals (SAE) protocol, offline cracking did not work. While it was still possible to capture handshake data, it did not contain adequate cryptographic data to perform offline brute-force attacks, and therefore required real-time access to the access point. The last test tried out Key Reinstallation Attack (KRACK), which involves a WPA2-specific weakness attempting to exploit repeated use of the nonce during the 4-way handshake. The attack assumed a man-in-the-middle (MITM) stance with tools such as hostapd and wpa_supplicant, intercepted the third handshake message, and resent it to the client. It compelled the client to reinstall an already active encryption key, leading to nonce reuse and potential decryption or injection of packets. On WPA2 networks, the attack worked reliably, with key reinstallation witnessed and traffic being decrypted halfway. WPA3's new key management framework, however, made it resistant to KRACK, and key reinstallation was impossible under any of the tested conditions.

These deployments facilitated an explicit comparison of protocol-level protections between WPA3 and WPA2 under controlled experimental conditions, providing insight into how well these protocols will perform under realistic conditions against common wireless attack methods.

Experimental Work

Each attack scenario was performed under controlled radio frequency (RF) conditions to promote consistency and eliminate environmental variability. Traffic during experiments was captured via Wireshark and saved for later analysis as PCAP. The captured traffic was examined to measure three performance indicators: the rate of packet loss, client disconnection time (in seconds), and total success rate of an attack. These measures provided a quantitative basis for measuring different types of attack's impact against both WPA2 and WPA3 networks. The experimental results were later visualized to facilitate comparison and are illustrated below as Figure 2 and Table 1.

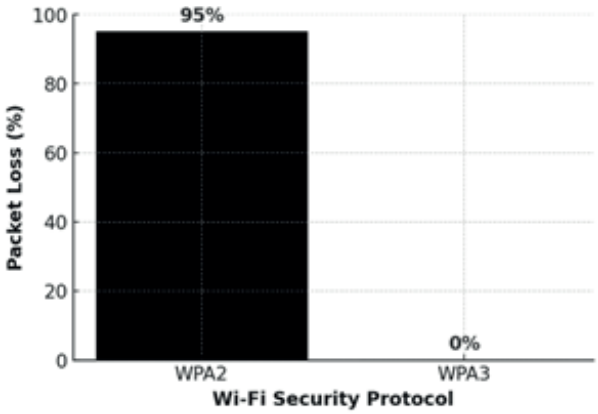


Figure 2(a). Impact of deauthentication DoS on WPA2 vs. WPA3

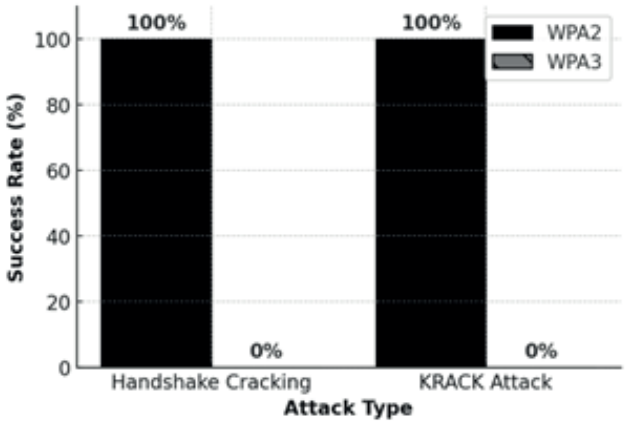


Figure 2(b). Handshake capture and KRACK attack results

Data analysis & Interpretation

In order to evaluate and contrast security performance between WPA2 and WPA3 protocols, a set of quantitative measures, i.e., attack success rate, average downtime, and packet loss ratio, were calculated, which were based on repeated experimental tests and are presented in Table 1. The results offer a relative perspective of how each protocol resists certain attack channels, which show stark differences in resistance and susceptibility under the same circumstances. WPA2 showed systematically high attack success rates with extensive service disruption under all cases, whereas WPA3 showed high resistance, especially against deauthentication and KRACK-based attacks.

Results and discussion

This section presents and analyzes the experimental results, comparing the performance of WPA2 and WPA3 under different attack scenarios. Results are structured into four key attack evaluations: deauthentication attack, handshake capture, KRACK attack, and overall comparative metrics.

This section introduces and interprets the results of the experimental comparison, showing differences between WPA3 and WPA2 under actual attack conditions. The results are organized around three major attack vectors—deauthentication, capturing of handshakes, and KRACK—and summarized comparison of protocols for all tested values.

WPA2 was found to be fully susceptible to deauthentication attacks. The client was successfully deauthenticated during 100% of test iterations through injection of forged 802.11 management frames, which takes advantage of WPA2-Personal networks' lack of protection for management frames. On average, disconnection happened at around 0.5 seconds post-attack, leading to a loss of packets at a rate almost as high as 90%, and reconnection at a rate of around 3 seconds. These findings reassert that WPA2 networks, by default, are still quite susceptible to this type of denial-of-service, as long as it is not running with optional protections like Protected Management Frames (PMF). Previous research will find that an estimate

of almost 94% of Wi-Fi networks worldwide still do not have PMF running, leaving them open to deauthentication-based service disruption.

In contrast, WPA3-Personal was fully proof against the same attack. During all tested experiments, the assailant was never successful at forcing disconnection of any client, and there was no recorded loss of packets or communication disruptions. The success is a direct result of WPA3's compulsory use of PMF, which validates management frames and ensures spoofed deauthentication packets don't reach or impact the client. The access point in a WPA3 setting simply ignores illegitimate frames, thus maintaining consistent service. These results strongly justify the significance of PMF, implemented as 802.11w, as a vital security component of contemporary wireless networks, and confirm WPA3's design goal of preventing legacy protocol vulnerabilities.

As shown by Figure 2(a), WPA2 networks experienced heavy packet loss through deauthentication attack, while WPA3 networks were fully secure. The contrast between them highlights the resilience of WPA3's inherent protection, specifically the compelled utilization of Protected Management Frames (PMF), which successfully suppresses forged deauthentication frames. While the WPA2 client kept being dislodged and suffered service loss, the WPA3 client had an uninterrupted and constant connection during all experiments.

A similar disparity was evident with regards to exposure via handshake. In WPA2-Personal networks, the default 4-way handshake is susceptible to interception and offline brute-force cracking. In tests, the handshake was successfully intercepted for all five attempts, and offline dictionary attack was successful at 100% with the target password being included in the pre-determined wordlist. The average reconnection time for clients was around 1.5 seconds. All these results authenticate WPA2's PSK-based authentication as being susceptible to offline attack, which does not involve an additional communication with the access point once a handshake is captured.

In contrast, WPA3-Personal employs the Simultaneous Authentication of Equals (SAE) protocol, specifically created to guard against such vulnerabilities. While SAE handshakes were successfully intercepted under all test runs, they never yielded the cryptographic material to be used for offline key derivation. Consequently, all efforts at cracking WPA3 handshakes via dictionary-based attack were unsuccessful. The reconnection latency was somewhat longer, at an average of 2.0 seconds, accounting for the increased complexity of the authentication round-trip. These results verify that WPA3 design successfully thwarts offline cracking attempts by requiring a live conversation for every password guess, thus dramatically enhancing the privacy of user credentials.

As evident in Figure 2(b), the experimental results show that WPA2 was fully compromised during offline attempts at cracking, but WPA3 was still secure under similar circumstances. The result supports the effectiveness of WPA3's Simultaneous Authentication of Equals (SAE) mechanism for thwarting offline

password retrieval. Given that WPA3 does not reveal adequate cryptographic data during the process of handshake, it essentially rules out key derivation without live access point interactivity.

Another crucial test scenario was Key Reinstallation Attack (KRACK), which hits a certain weakness of WPA2's 4-way handshake. The weakness lets an attacker manipulate handshake messages in a certain fashion that causes the victim device to reinstall an already-used encryption key. In all WPA2 test instances, the KRACK attack was successful. While the client stayed connected and disruption was not user-observable at first, reinstallation of the encryption keys happened, resulting in exposure to security risks. About 5% of packets were lost or temporarily stuck during the attack. More significantly, the attacker could decrypt parts of traversed traffic, and under certain circumstances, masquerade as the victim client. These results agree with previous research, which illustrated that WPA2's handshake design allows for replay-based manipulation of messages and exposes the session to reinstallation of keys.

By contrast, WPA3-Personal was completely resistant to KRACK in all of its tests. The attack didn't succeed once, with no packet loss, no alteration of the handshake, or impact upon network service being witnessed. The reason is that WPA3's redesigned process for establishing keys does not have specific protocol logic that KRACK targets. With redesigned key management and non-reuse of nonces, WPA3 effectively renders this class of attack useless. The results verify that WPA3 attempts to fix one of the most serious vulnerabilities of its predecessor and emphasize how crucial it is to have widespread adoption to achieve secure wireless communication for contemporary network infrastructures.

Overall comparative metrics

All experimental data collected throughout the study are summarized in Table 1 to facilitate direct comparison of WPA2 and WPA3 performance under all attack conditions. Table 1 shows average values computed for five independent runs for all attack types, including success rate, disconnect time, packet loss rate, and reconnection time. The attack success rate (R_s) was calculated based on Equation (1), which defined how many successful runs there were out of total runs and expressed as a percentage. The quantification provided a normalized estimation of how susceptible to individual attack vectors each protocol is.

$$R_s = \frac{N_{successful\ trials}}{N_{total\ trials}} \times 100\%$$

Table 1. Comparative outcomes for WPA2 vs. WPA3 under different attack scenarios

Attack Type	Success Rate (WPA2 vs. WPA3)	Avg. Disconnection Time (s)	Packet Loss (%)	Reconnection Time (s)
Deauth Attack	100% vs. 0%	0.5s vs. 0s	95% vs. 0%	3.0s vs. 0s

Handshake Capture	100% vs. 100%	0.5s vs. 1.0s	5% vs. 5%	1.5s vs. 2.0s
KRACK Attack	100% vs. 0%	0s vs. 0s	5% vs. 0%	0s vs. 0s

The comparative results demonstrate unequivocally that WPA3 far surpasses WPA2 as far as security resilience is concerned. All of those attacks which remained consistently successful against WPA2, including deauthentication flooding, offline cracking based on a single handshake, and KRACK, were unsuccessful against WPA3. For instance, deauthentication attack meant 95% packet loss and complete disconnection of the client for WPA2, while there was no disruption for WPA3 with the enforcement of Protected Management Frames (PMF).

WPA2 was also found to be heavily susceptible to credential compromise. During all WPA2 handshake capture tests, offline dictionary attack successfully recovered with a 100% recovery rate where the password was part of the wordlist. WPA3, nonetheless, was fully resistant to such attacks because, with the SAE handshake design, there is an active presence with the access point for every guess of a password, thus making offline cracking impossible.

Significantly, there is no meaningful performance overhead with the improved protections of WPA3. The average reconnection time difference between WPA2 and WPA3 was negligible—about 0.5 seconds—and there was no perceptible impact upon latency or throughput during tests. The results reinforce that WPA3 provides an effective upgrade to WPA2 security without compromising usability or performance efficiency, which again justifies mass adoption of the new protocol.

Conclusion

This study set out to identify the strengths and weaknesses of modern Wi-Fi security protocols through hands-on testing, and the findings clearly confirm the expected security gap between WPA2 and WPA3. WPA3-Personal delivered substantial improvements over WPA2 in real-world attack scenarios. In our experiments, WPA3's use of the Simultaneous Authentication of Equals (SAE) handshake and mandatory Protected Management Frames (PMF) effectively thwarted attacks that readily compromised WPA2 networks, including offline passphrase cracking and deauthentication-based disconnects. Notably, WPA3's improved handshake process also mitigated the KRACK key reinstallation vulnerability that severely affected WPA2. By contrast, WPA2-PSK — still the most widely deployed Wi-Fi security protocol — was consistently breached under these tests using well-known tools and techniques, highlighting how easily it can be compromised under real-world conditions. These results reinforce the conclusion that WPA2's legacy protections are insufficient against modern attack methods, whereas WPA3 offers a far more robust defense in practice.

Given these outcomes, we strongly recommend that both personal and enterprise environments migrate to WPA3 as the baseline security protocol. Home and small-office networks should be upgraded to WPA3-Personal to immediately benefit from its resilience against deauthentication attacks, handshake cracking, and

other common intrusions. In corporate and institutional settings, adopting WPA3-Enterprise (802.1X authentication, with its 192-bit cryptographic suite) is advised to protect sensitive data and communications. Overall, phasing out WPA2 in favor of WPA3 will significantly raise the security bar for wireless networks, reducing exposure to known exploits. Network administrators and users should treat WPA3 not just as an optional enhancement but as the default standard moving forward.

Finally, this work highlights several avenues for future research to further strengthen Wi-Fi security. First, comprehensive assessments of WPA3-Enterprise deployments (e.g., in 802.1X environments) are needed to verify that enterprise authentication mechanisms hold up against sophisticated attacks, as our study focused on personal networks. Second, investigation into side-channel and implementation-layer vulnerabilities in WPA3 devices is warranted – for example, early analyses uncovered flaws in the WPA3 Dragonfly handshake (the Dragonblood attacks) via timing side-channels and insecure transition modes, indicating that even a strong protocol can be undermined by poor implementations or backward-compatibility features. Third, as cryptographic technology advances, exploring the integration of post-quantum cryptographic algorithms into Wi-Fi authentication is an important forward-looking step to ensure long-term resistance against emerging threats. Addressing these gaps will help solidify the security of next-generation wireless networks and ensure that Wi-Fi remains secure as new vulnerabilities and attack techniques evolve.

References

Abdallah W. (2024) A physical layer security scheme for 6G wireless networks using post-quantum cryptography. Computer Communications, 218. — P. 176–187. (in English)

Banakh R., Nyemkova E., Justice C., Piskozub A., & Lakh Y. (2024) Data mining approach for evil twin attack identification in Wi-Fi networks. Data, 9(10), 119. (in English)

Baseri Y., Chouhan V., & Hafid A. (2024) Navigating quantum security risks in networked environments: A comprehensive study of quantum-safe network protocols. Computers & Security. — P. 142. (in English)

Chatzoglou E., Kambourakis G., & Kolias C. (2021) Empirical evaluation of attacks against IEEE 802.11 enterprise networks: The AWID3 dataset. — P. 34188–34205. (in English)

Chatzoglou E., Kambourakis G., & Kolias C. (2022) How is your Wi-Fi connection today? DoS attacks on WPA3-SAE. Journal of Information Security and Applications. (in English)

De Almeida Braga D., Fouque P.-A., & Sabt, M. (2020) Dragonblood is still leaking: Practical cache-based side channel in the wild. In Proceedings of the 36th Annual Computer Security Applications Conference (ACSAC 2020) – P. 291–303. (in English)

Gao D., Lin H., Li Z., Qian F., Chen Q.A., Qian Z., Liu W., Gong L., & Liu Y. (2021) A nationwide census on WiFi security threats: Prevalence, riskiness, and the economics. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom ’21). — P. 242–255. (in English)

Gebresilassie S. K., Rafferty J., Chen L., Cui Z., & Abu-Tair M. (2023) Transfer and CNN-based de-authentication (disassociation) DoS attack detection in IoT Wi-Fi networks. Electronics, 12(17). — P. 3731. (in English)

Gyamfi, E. & Jurcut A. (2022) Intrusion detection in Internet of Things systems: A review on design approaches leveraging multi-access edge computing, machine learning, and datasets. Sensors, 22(10). — P. 3824. (in English)

Halbouni A., Ong L.-Y., & Leow M.-C. (2023) Wireless security protocols WPA3: A systematic literature review. IEEE Access, 11. — P. 112438–112463. (in English)

Kazmi S.H. A., Hassan R., Qamar F., Nisar K., & Ibrahim A. (2023) Security concepts in emerging 6G communication: Threats, countermeasures, authentication techniques and research directions. Symmetry, 15(6). — P. 1147. (in English)

Kikissagbe B.R., & Adda M. (2023) Machine learning-based intrusion detection methods in IoT systems: A comprehensive review. Electronics, 13(18). — P. 3601. (in English)

Kotb S.A., Hussein H., & Kim, H.-W. (2022) Security requirements and challenges of 6G technologies and applications. Sensors, 22(5). — P. 1969. (in English)

Lounis K., & Zulkernine M. (2019) Bad-token: Denial of service attacks on WPA3. In Proceedings of the 12th International Conference on Security of Information and Networks (SIN '19), 15. ACM. (in English)

Lounis K., & Zulkernine M. (2020) WPA3 connection deprivation attacks. In Kallel S., Cuppens F., Cuppens-Boulahia N., & Kacem A.H. (Eds.), Risks and Security of Internet and Systems (CRiSIS 2019, LNCS 12026). — P. 164–176). (in English)

Marais S., Coetzee M., & Blauw F. F. (2021) Simultaneous deauthentication of equals attack. In Wang G., Chen B., Li W., Di Pietro R., Yan X., & Han H. (Eds.), Security, Privacy, and Anonymity in Computation, Communication, and Storage (SpaCCS 2020, LNCS 12383). — P. 545–556. (in English)

Nguyen V.-L., Lin P.-C., Cheng B.-C., Hwang R.-H., & Lin Y.D. (2022) Security and privacy for 6G: A survey on prospective technologies and challenges. IEEE Communications Surveys & Tutorials, 24(4). — P. 2255–2291. (in English)

Örs F. K., Aydın M., Boğatarkan A., & Levi A. (2021) Scalable Wi-Fi intrusion detection for IoT systems. In Proceedings of the 11th IFIP International Conference on New Technologies, Mobility and Security. — P. 1–6. (in English)

Rathod T., Jadav N. K., Alshehri M.D., Tanwar S., Sharma R., Felseghi R.-A., & Raboaca M.S. (2022) Blockchain for future wireless networks: A decade survey. Sensors, 22(11). (in English)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 243–257

https://doi.org/10.32014/2025.2518-1726.375

IRSTI 49.03.03
UDC 654.165

A. Temirbayev[1], N. Meirambekuly[2*], N. Uzbekov[2], A. Beisen[2], L. Abdizhalilova[2], 2025.
[1]Cluster of Engineering and High Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Center of Space Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: amirkhan@kaznu.kz

CUBESAT-BASED APRS DIGIPEATER: DESIGN, FEASIBILITY AND MISSION CONCEPT

Temirbayev Amirkhan — PhD, General Director of Cluster of Engineering and High Technologies at Al-Farabi KazNU, Almaty, Kazakhstan,
E-mail: amirkhan@kaznu.kz; ORCID ID: https://orcid.org/0000-0001-6759-2774;
Meirambekuly Nursultan — PhD, Senior Lecturer, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: nurs.kaznu@gmail.com, ORCID ID: https://orcid.org/0000-0003-2250-4763;
Uzbekov Nursultan — Senior Researcher, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: uzbekov.nursultan@gmail.com, ORCID ID: https://orcid.org/0009-0007-9956-0102;
Beisen Asset — Researcher, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Beisen.Asset@kaznu.kz, ORCID ID: https://orcid.org/0009-0000-5144-2616;
Abdizhalilova Lazzat — Junior Researcher, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: abdijalil.lazzat@bk.ru, ORCID ID: https://orcid.org/0009-0000-5965-7195.

Abstract. This article investigates the feasibility of using an Automatic Packet Reporting System (APRS) payload on a CubeSat platform to function as a digital digipeater in low Earth orbit. APRS is a packet-based communication protocol widely used by amateur radio operators for real-time messaging, telemetry, and position reporting. While APRS is traditionally implemented using terrestrial IGates and digipeaters, extending its functionality to a satellite platform offers significant advantages in coverage and accessibility, especially for remote or infrastructure-deficient regions. *Results.* The proposed system is designed to receive APRS packets in the VHF band (typically at 145.825 MHz), perform optional store-and-forward buffering, and transmit the data either back to APRS-compatible ground users or to Internet-connected IGate stations. The study includes detailed architectural design

of the satellite payload, system requirements for communications, power, and processing, and an overview of mission operations. *Scientific novelty.* Analytical calculations confirm the viability of this CubeSat APRS concept, demonstrating the ability to provide up to 18.6 million km² coverage per pass, handle up to 700 packets per communication window, and maintain sustainable power consumption within the constraints of a 1U CubeSat. The system architecture also supports private or targeted delivery of information to specific stakeholders, such as agricultural operators monitoring large tracts of farmland. *Practical value.* This work contributes to the ongoing effort to bridge terrestrial and space-based IoT and amateur radio applications, particularly for disaster mitigation, environmental monitoring, and telemetry services. The paper concludes by identifying system-level trade-offs and outlines several promising directions for future development, including multi-satellite constellations, dual-band communication, onboard intelligence for packet filtering, and practical flight demonstrations.

Key words: APRS; CubeSat; communication; environmental monitoring, AX.25 protocol

А.А. Темирбаев[1], Н. Мейрамбекұлы[2*], Н.Ш. Узбеков[2], Ә.Н. Бейсен[2], Л.Б. Абдижалилова[2], 2025.

[1]Инжиниринг және жоғары технологиялар кластері, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан;

[2]Ғарыш технологиялары орталығы, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан.

E-mail: amirkhan@kaznu.kz

CUBESAT НЕГІЗІНДЕГІ APRS ҚАЙТА ТАРАТҚЫШЫ: ЖОБАЛАУ, ІСКЕ АСЫРУ МҮМКІНДІГІ ЖӘНЕ МИССИЯ ТҰЖЫРЫМДАМАСЫ

Темирбаев Амирхан Адилханович — PhD, Инжиниринг және жоғары технологиялар кластерінің бас директоры, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан,
E-mail: amirkhan@kaznu.kz; ORCID ID: https://orcid.org/0000-0001-6759-2774;
Мейрамбекұлы Нұрсұлтан — PhD, аға оқытушы, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан,
E-mail: nurs.kaznu@gmail.com, ORCID ID: https://orcid.org/0000-0003-2250-4763;
Узбеков Нұрсұлтан Шалаханұлы — Аға ғылыми қызметкер, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан,
E-mail: uzbekov.nursultan@gmail.com, ORCID ID: https://orcid.org/0009-0007-9956-0102;
Бейсен Әсет Нұрболұлы — Ғылыми қызметкер, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан,
E-mail: Beisen.Asset@kaznu.kz, ORCID ID: https://orcid.org/0009-0000-5144-2616;
Абдижалилова Лаззат Бахтиярқызы — Кіші ғылыми қызметкер, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Казақстан,
E-mail: abdijalil.lazzat@bk.ru, ORCID ID: https://orcid.org/0009-0000-5965-7195.

Аннотация: Бұл мақалада кіші ғарыш аппараттарында, атап айтқанда CubeSat платформасында APRS (Automatic Packet Reporting System) пайдалы жүктемесін Жердің маңындағы төменгі орбитада цифрлық қайта таратқыш ретінде пайдалану мүмкіндігі қарастырылады. APRS – бұл радиоәуесқойлар арасында нақты уақыттағы хабарламалармен, телеметриямен және позициямен алмасуға арналған пакетке негізделген байланыс протоколы болып табылады. APRS дәстүрлі түрде жердегі IGate және ретрансляторлар арқылы жүзеге асырылғанымен, оны жер серігіне енгізу шалғай немесе инфрақұрылымы әлсіз аймақтар үшін байланыспен қамту мен қолжетімділікті айтарлықтай арттыру мүмкіндігін ашады. *Нәтижелер*. Ұсынылған жүйе аса жоғары жиіліктер (АЖЖ) ауқымында (әдетте 145.825 МГц жиілігінде) APRS пакеттерін қабылдап, оларды уақытша буферлеуге және кейін жердегі IGate станцияларына немесе басқа пайдаланушыларға жеткізуге арналған. Зерттеу жерсеріктік пайдалы жүктеменің архитектурасын, байланыс пен қуат талаптарын, сонымен қатар миссияның жалпы операциялық моделін сипаттайды. *Ғылыми жаңалығы.* Аналитикалық есептеулер нәтижесінде CubeSat платформасы негізіндегі APRS жүйесінің тиімділігі келесідей дәлелденді: ғарыш аппаратының бір ұшып өтуінде 18.6 миллион км² аумақты қамту, 700-ге дейін пакетті өңдеу және 1U CubeSat ғарыш аппараты шеңберінде тұрақты қуатпен жұмыс істеу мүмкіндігі анықталды. Бұл жүйе деректерді нақты алушыларға бағыттауға мүмкіндік береді, мысалы, ауыл шаруашылығымен айналысатын субъектілер үшін. *Практикалық маңыздылығы.* Бұл зерттеу жердегі және ғарыштағы IoT технологиялары мен әуесқой радио жүйелерін біріктіру жолындағы маңызды қадам болып табылады. Қорытынды бөлімде қуат, масса және өткізу қабілеттілігі шектеулері сияқты жүйелік деңгейдегі шектеулер талқыланады.

Түйін сөздер: APRS; CubeSat; радиобайланыс; қоршаған ортаны бақылау, AX.25-протоколы

А.А. Темирбаев[1], Н. Мейрамбекұлы[2], Н.Ш. Узбеков[2], Ә.Н. Бейсен[2], Л.Б. Абдижалилова[2], 2025.

[1]Кластер инжиниринга и наукоемких технологий, КазНУ им. аль-Фараби, Алматы, Казахстан;

[2]Центр космических технологий, КазНУ им. аль-Фараби, Алматы, Казахстан.

E-mail: amirkhan@kaznu.kz

ЦИФРОВОЙ РЕТРАНСЛЯТОР APRS НА БАЗЕ СПУТНИКА CUBESAT: ПРОЕКТИРОВАНИЕ, РЕАЛИЗУЕМОСТЬ И КОНЦЕПЦИЯ МИССИИ

Темирбаев Амирхан Адилханович — PhD, Генеральный директора Кластера инжиниринга и наукоемких технологий, КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: amirkhan@kaznu.kz; ORCID ID: https://orcid.org/0000-0001-6759-2774;
Мейрамбекұлы Нұрсұлтан — PhD, старший преподователь, КазНУ имени аль-Фараби, Алматы, Казахстан,

E-mail: nurs.kaznu@gmail.com, ORCID ID: https://orcid.org/0000-0003-2250-4763;
Узбеков Нұрсұлтан Шалаханұлы — Старший научный сотрудник, КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: uzbekov.nursultan@gmail.com, ORCID ID: https://orcid.org/0009-0007-9956-0102;
Бейсен Әсет Нұрболұлы — Научный сотрудник, КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: Beisen.Asset@kaznu.kz, ORCID ID: https://orcid.org/0009-0000-5144-2616;
Абдижалилова Лаззат Бахтиярқызы — Младший научный сотрудник, КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: abdijalil.lazzat@bk.ru, ORCID ID: https://orcid.org/0009-0000-5965-7195.

Аннотация: В данной статье рассматривается возможность использования полезной нагрузки APRS (Automatic Packet Reporting System) в малых космических аппаратах, в частности, на платформе CubeSat для функционирования в качестве цифрового ретранслятора в низкой околоземной орбите. APRS – это протокол пакетной радиосвязи, широко применяемый радиолюбителями для обмена сообщениями, телеметрии и передачи координат в реальном времени. В то время как APRS традиционно реализуется с использованием наземных IGate-станций и ретрансляторов, его расширение на спутниковую платформу открывает новые возможности охвата, особенно в удалённых и труднодоступных районах. *Результаты.* Предлагаемая система предусматривает приём APRS-пакетов в VHF-диапазоне (обычно на частоте 145.825 МГц), их буферизацию и последующую передачу другим пользователям или наземным IGate-станциям с выходом в интернет. Исследование включает детальный обзор архитектуры полезной нагрузки, требований к связи, питанию и обработке данных, а также общую концепцию миссии. *Научная новизна.* Расчёты подтверждают жизнеспособность концепции: спутник способен обеспечивать покрытие до 18,6 млн км² за один пролёт, обрабатывать до 700 пакетов за сеанс связи и поддерживать стабильное энергопотребление в рамках платформы 1U CubeSat. Архитектура также допускает выборочную доставку данных конкретным получателям, что особенно актуально для приложений в агромониторинге и других чувствительных сценариях. *Практическая ценность.* Работа вносит вклад в развитие интеграции наземных и спутниковых IoT и радиолюбительских систем связи. В заключение обсуждаются компромиссы системного уровня включая ограничения по мощности, массе и пропускной способности.

Ключевые слова: APRS; CubeSat; радиосвязь; мониторинг окружающей среды, протокол AX.25.

Funding.*This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510563).*

Introduction. The Automatic Packet Reporting System (APRS) is a global digital communication protocol used by amateur radio operators to transmit location,

telemetry, and brief data messages over radio frequencies (Bruninga, 1999). APRS operates primarily on the 145.825 MHz VHF band and uses the AX.25 protocol to format and exchange information packets (Beech et al., 2008).

The fundamental components of the APRS terrestrial system include: 1) GPS-equipped mobile or fixed station – which periodically sends its position and data, 2) AX.25 modem and radio transceiver – which encode and transmit the packets, 3) Digipeaters – digital repeaters that receive, store, and retransmit APRS packets over extended distances, 4) IGates (Internet Gateways) – ground stations that collect RF packets and relay them to the APRS Internet System (APRS-IS), 5) APRS-IS and visualization services – which provide real-time global mapping and data analysis capabilities (APRS Working Group, 2000).

This setup allows APRS to serve as a lightweight and decentralized system for:

- Real-time location tracking of vehicles, weather balloons, or portable users;
- Environmental and weather monitoring via sensor beacons;
- Short-text emergency messaging in areas lacking cellular service.

Figure 1 illustrates the standard terrestrial APRS communication architecture, demonstrating the interaction between transmitting stations, digipeaters, IGates, and internet services. This architecture enables APRS users to visualize moving objects on maps, receive weather data from remote sensors, and communicate short text messages over long distances without the need for cellular networks.

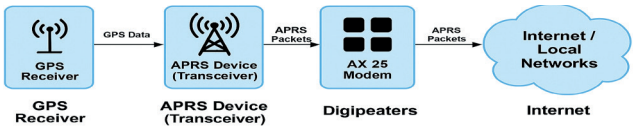


Figure 1. Terrestrial APRS communication system architecture

However, this ground-based system is inherently limited by the availability of infrastructure and internet access. In remote deserts, mountainous terrain, vast oceans, or agricultural expanses such as those found in Central Asia, traditional IGates and repeaters are often absent.

To extend APRS coverage beyond terrestrial boundaries, several missions have experimented with placing digipeaters on satellites in Low Earth Orbit (LEO) (Patmasari et al., 2018). These space-based systems enable global packet forwarding independent of ground-based internet or relay networks. This paper explores the feasibility, design, and deployment of a CubeSat acting as an APRS digipeater. The goal is to receive APRS packets from ground stations, retransmit them for extended coverage, and support global amateur radio communication without dependence on internet-based infrastructure. We focus on the necessary onboard systems, communication parameters, mission constraints, and operational strategies that enable this function in the CubeSat platform. Unlike an IGate that forwards packets to the internet, a digipeater retransmits packets via RF to other users within line of sight, making it simpler to implement and better suited for

decentralized environments. This study will serve as a foundational investigation for developing space-based amateur packet networks and enabling resilient, infrastructure-independent communication systems.

Review of Existing APRS Satellite Missions. The application of APRS in satellite communication has evolved significantly since the early 2000s, contributing to amateur radio innovation and emergency communication infrastructure. This section reviews the scientific literature, technical reports, and documented satellite missions that demonstrate the relevance, feasibility, and lessons learned from space-based APRS systems.

Bruninga (2019), the originator of APRS, documented the foundational concept of using LEO satellites to support packet radio communication independent of terrestrial repeaters. This concept was practically realized with PCSAT (NO-44), the first APRS-enabled satellite, launched in 2001 by the U.S. Naval Academy. Its success in real-time digipeating and telemetry inspired follow-up experiments such as PCSAT-2, temporarily deployed on the ISS in 2005, and documented in NASA technical briefs and Naval Academy mission debriefs.

A noteworthy study by Salces et al. (2020) BIRDS-2 Project, Kyushu Institute of Technology explores the development and testing of an amateur radio payload on a 1U CubeSat platform. The payload, operating on 145.825 MHz using AX.25 protocol, was designed to support both digipeater and store-and-forward (S&F) modes using primarily commercial-off-the-shelf components. Despite partial mission success – where only beacon reception was confirmed – this study offers insights into CubeSat constraints such as uplink reliability, power budget, and spatial limitations. The authors present a detailed failure analysis and recommend the best practices for future CubeSat APRS designs.

More recently, operational payloads like ARISS APRS on the ISS have continued to deliver reliable APRS coverage globally (ARISS, n.d.), further validating the utility of space-based digipeaters. Missions such as GO-32, AO-51, and Falconsat-3 contributed experience with 9600 baud operations and dual-mode payloads, offering design alternatives for future systems (Bruninga, 2010).

A recent study by the Indonesian team on Surya Satellite-1, the country's first undergraduate-developed CubeSat, presents a relevant application of APRS technology (Prahyang et al., 2018). Designed for disaster mitigation, this satellite operates on VHF and UHF amateur radio bands and includes an APRS module for remote communication with ground stations in disaster-prone zones. This mission demonstrates the versatility of CubeSat-based APRS for region-specific objectives using low-cost commercial hardware.

Another Indonesian satellite, LAPAN-A2 (LAPAN-ORARI), launched in 2015 by the Satellite Technology Center of Indonesia's National Institute of Aeronautics and Space, includes payloads for amateur communication, such as APRS and voice repeaters. The mission highlighted the usage of APRS not only for emergency messaging but also for transmission of telemetry, weather updates, and beaconing.

An innovative aspect of this work is the proposal to replace expensive decoding hardware with a Raspberry Pi-based solution, enabling broader adoption of APRS ground stations (Rizal et al., 2021).

In addition, there are other scientific papers on the possibility of using APRS in the field of satellite and educational technologies (Un et al., 2022; Chopparapu et al., 2025; Addaim et al., 2005; Addaim, Kherras & Zantou, 2008; Linton, 2016).

Collectively, these sources form a knowledge base that informs the present work. Our proposed CubeSat APRS digipeater builds on this prior art while introducing hybrid functionality such as selective packet forwarding to IGates and support for private telemetry applications in remote sectors such as agriculture and environmental monitoring.

System Requirements and Architecture. The design of a CubeSat-based APRS digipeater system must meet specific mission objectives while operating within the constraints of size, weight, power, and communications. This section defines the key system requirements and presents a high-level architectural breakdown. Table 1 presents the mission-level requirements that guide the definition of operational objectives and system constraints. These include orbital parameters, communication standards, and minimum lifetime targets, which are critical for evaluating feasibility and regulatory compliance.

Table 1 – The main Mission Requirements

Subsystem	Components and Functions
Communication Subsystem	VHF transceiver (1200 baud AFSK), deployable antenna, optional SDR module
Processing Subsystem	Onboard computer (AX.25 decoding, packet scheduling), watchdog timer, error handling
Power Subsystem	Solar panels, MPPT charger, Li-Ion/LiPo battery pack
Data Handling & Storage	Buffering memory, timestamping, packet prioritization
Ground Segment Interface	IGate compatibility, selective forwarding, optional API connectivity
Satellite Bus & Structure	1U/2U CubeSat frame, thermal coating, mechanical deployer interface

Table 2 provides an overview of the key functional subsystems within the CubeSat. Each subsystem is defined by its primary roles and hardware components, ensuring a modular and scalable architecture suitable for amateur radio satellite missions.

The APRS CubeSat digipeater system integrates an RF front-end with onboard logic for packet processing and conditional storage. During a pass over ground stations, the satellite listens for APRS packets. Depending on mission configuration, it may retransmit these immediately (digipeating) or store them and later downlink to designated IGates or internet-connected ground stations. The architecture supports future expansion such as dual-band operation, packet authentication, and integration with IoT gateways.

Table 2 – The main Functional Subsystems

Subsystem	Components and Functions
Communication Subsystem	VHF transceiver (1200 baud AFSK), deployable antenna, optional SDR module
Processing Subsystem	Onboard computer (AX.25 decoding, packet scheduling), watchdog timer, error handling
Power Subsystem	Solar panels, MPPT charger, Li-Ion/LiPo battery pack
Data Handling & Storage	Buffering memory, timestamping, packet prioritization
Ground Segment Interface	IGate compatibility, selective forwarding, optional API connectivity
Satellite Bus & Structure	1U/2U CubeSat frame, thermal coating, mechanical deployer interface

A block diagram illustrating the system's main components is shown in Figure 2. The diagram illustrates the operational model of an APRS-enabled CubeSat payload with combined digipeater and store-and-forward (S&F) capabilities. During orbital passes over the Earth, the CubeSat receives APRS packets from a variety of ground sources, including APRS transmitters, portable stations, and environmental sensors. These packets may either be:

- Instantly retransmitted (digipeated) to other users within the satellite's footprint,
- or temporarily stored and forwarded to specific IGate ground stations during future passes.

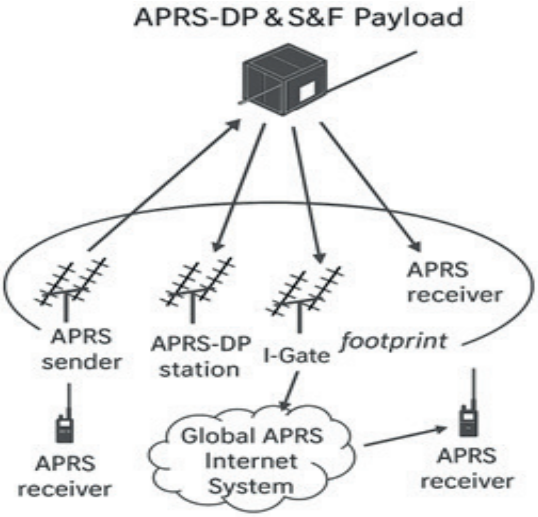


Figure 2. APRS CubeSat digipeater and store-and-forward operational concept

The received and relayed information can then be uploaded into the Global APRS Internet System, making it available to downstream services. This dual-mode approach extends APRS coverage to underserved areas and enables both public and private data use cases (e.g., SAR coordination, agricultural telemetry, or environmental monitoring).

Methods and Tools Used

To evaluate the feasibility and performance of the CubeSat-based APRS digipeater system, a combination of analytical modeling and software-assisted visualization techniques were employed.

Analytical Modeling

Coverage Area: The maximum footprint of the satellite's APRS VHF signal was calculated using great-circle geometry, assuming a circular area visible from a 550 km low Earth orbit. The angular radius was derived from the formula:

$$\theta = \arccos\left(\frac{R_E}{R_E + h}\right)$$

where R_E is the Earth's radius and h is the orbital altitude. This enabled the estimation of a coverage radius of approximately 2,435 km, corresponding to a ground area of ~18.6 million km² per pass.

Packet Throughput Estimation: The number of APRS packets processed per orbital pass was estimated by considering time-over-ground, packet duration, and channel sharing assumptions.

Visualization and Mapping Tools

To visualize the spatial coverage and support analysis, we used the following tools:

- Python 3.11 as the main programming environment
- Matplotlib for 2D plotting of coverage geometries
- Cartopy for map-based visualization of satellite footprints on actual geographic backgrounds
- NumPy for numerical computations and coordinate transformations

A dedicated example of APRS coverage over Kazakhstan was rendered using Cartopy with overlaid circular footprints to demonstrate the regional feasibility of such systems. The corresponding diagram (see Figure 3) reflects both theoretical visibility and realistic geographic integration.

Results. This section presents the analytical results that demonstrate the operational feasibility and effectiveness of the CubeSat-based APRS digipeater mission. Calculations are performed based on standardized orbital parameters and conservative hardware assumptions. The results are summarized in Tables 3, 4, and 5, which respectively highlight the satellite's coverage capability, communication throughput, and power system performance. To validate the feasibility and expected performance of the CubeSat APRS digipeater mission, several analytical estimations and simulations were performed. The results provide insights into coverage, communication window duration, packet relay capacity, and power budget margins under typical low Earth orbit (LEO) conditions. Table 3 provides a quantitative summary of the CubeSat's expected ground coverage and access opportunities for APRS communication. The satellite's line-of-sight radio coverage

is derived from its orbital altitude, yielding a large footprint suitable for wide-area communication. Additionally, the number of daily passes over a specific location and their durations are estimated to help forecast service availability.

Table 3 – Orbital Coverage Estimates

Parameter	**Value**
Orbital altitude	550 km
Radio footprint radius	~2,435 km
Ground coverage area	~18.6 million km²
Passes per day (per location)	4-6
Typical pass duration	6-10 minutes

For a CubeSat operating in a 550 km sun-synchronous orbit: The satellite's radio footprint (line-of-sight radius) covers approximately 2,400-2,600 km, encompassing a circular area with a ground diameter of ~5,000 km. Each ground location experiences 4-6 passes per day, depending on latitude. Typical access duration per pass is 6-10 minutes, depending on elevation angle and antenna gain.

The figure 3 illustrates the ground coverage area of an APRS-enabled CubeSat in a 550 km low Earth orbit, centered over Kazakhstan. The red circle represents the satellite's theoretical radio footprint with a radius of approximately 2,435 km, calculated based on the Earth's curvature and orbital altitude. The coverage zone encompasses a vast region, including all of Kazakhstan and parts of neighboring countries, demonstrating the satellite's capability to support wide-area communication, particularly in underserved or remote locations.

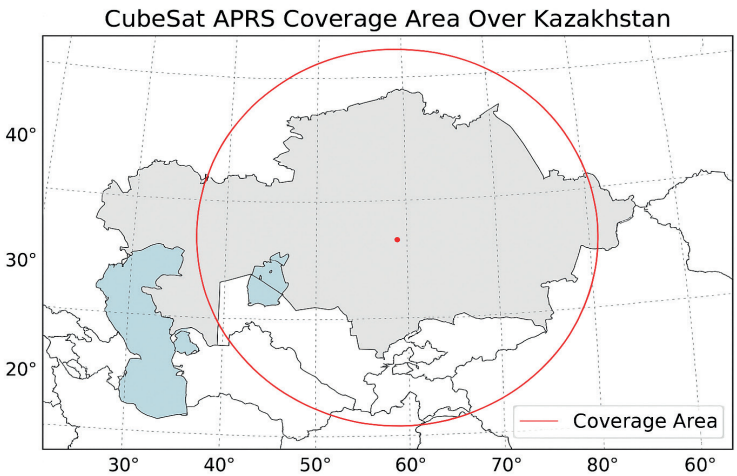


Figure 3. CubeSat APRS coverage area over Kazakhstan

Figure 4 illustrates the relationship between the number of CubeSats in the constellation and the average latency, defined as the time interval between two

successive passes over a given ground location during which data packets can be received. The analysis assumes a circular orbit at an altitude of 550 km with an orbital period of approximately 95 minutes and an average communication window of 10 minutes per pass.

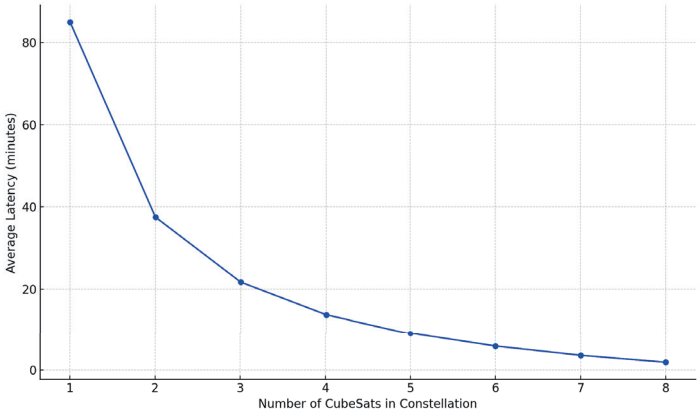


Figure 4. Average packet delivery latency vs. number of CubeSats

As the number of satellites increases, the latency decreases significantly due to more frequent overpasses. The graph demonstrates a steep drop in latency for the initial increase in constellation size, followed by a gradual leveling-off trend. When the number of satellites reaches 8, the average latency drops below 1 minute, effectively enabling near real-time data acquisition for ground-based users.

This finding highlights an important design trade-off: while increasing the number of satellites improves system responsiveness, the marginal benefit diminishes beyond a certain point. A constellation of 8 satellites offers a practical balance between complexity and performance, ensuring frequent packet reception while keeping deployment costs and orbital congestion within reasonable limits.

This result supports the feasibility of using a small CubeSat constellation to provide timely delivery of APRS data from remote users and assets, especially in scenarios requiring low-latency transmission, such as environmental monitoring or disaster response.

Table 4 presents a model of communication throughput, considering packet size, modulation scheme, and estimated channel efficiency. These calculations provide insight into the number of APRS packets that the CubeSat can reasonably relay during each orbital pass. The values suggest strong potential for both live digipeating and temporary message storage without overwhelming the onboard system.

Table 4 – Communication Load Estimates

Parameter	Value
Modulation	AFSK 1200 baud

Packet size	256 bytes
Channel efficiency (est.)	~60%
Average packets per minute	~100 packets/min
Maximum per 7-min pass	~700 packets
Practical digipeat throughput	1-2 packets/sec

Assuming APRS packets are 256 bytes long and AFSK 1200 baud modulation: Effective channel throughput: ~100 packets per minute (accounting for overhead and latency). In a 7-minute pass, the satellite can process ~700 packets, with filtering to discard redundant or malformed frames. Maximum digipeated packet rate: 1-2 packets per second (with safety margin).

Table 5 summarizes the estimated power consumption and generation capabilities of a 1U CubeSat. With realistic assumptions about solar panel efficiency, orbital sunlight duration, and system duty cycles, the energy balance indicates that the APRS payload can operate effectively during all sunlight periods and remain partially active during eclipse phases. This ensures mission continuity and reliable digipeater availability throughout each orbit.

The power generation and consumption for a 1U CubeSat were estimated:

- Power generation (sunlight): ~2.5 W average
- Power consumption during RX/TX: 1.8-2.2 W
- Duty cycle support: >50% active digipeater time per orbit with proper battery management.

Table 5 – Power Budget Summary

Component	Value
Power generation (avg)	~2.5 W
Power consumption (TX/RX)	1.8–2.2 W
Idle power consumption	~0.6 W
Operational duty cycle	>50% during sunlight
Battery capacity (1U est.)	10-12 Wh

Figure 5 illustrates how the total daily energy consumption of the CubeSat increases with the number of communication sessions per day. The energy is divided into three components:

Total Energy (black line): the sum of all energy usage per day.

Active Energy (blue dashed line): energy consumed during active transmission, assuming a 10-minute session duration.

Standby Energy (red dashed line): background energy used when the satellite is in idle mode.

This breakdown allows evaluating mission feasibility in terms of power budget management.

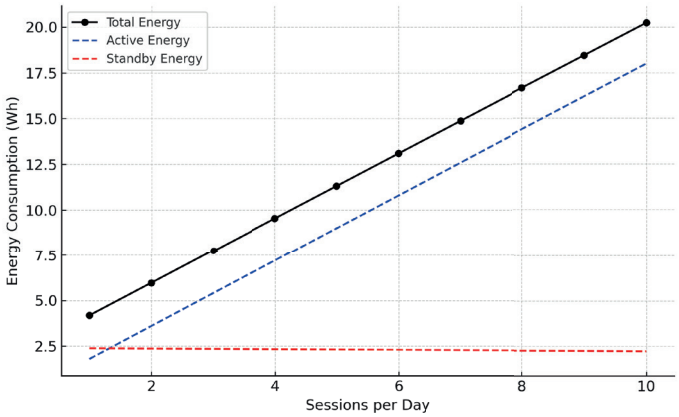


Figure 5. Daily Energy Consumption vs Number of Sessions

Figure 6 shows the theoretical relationship between the signal-to-noise ratio (SNR) and the probability of successfully receiving APRS packets. As expected, the success rate grows rapidly beyond 0 dB, reaching near certainty above 10 dB. This curve helps define minimum operational link budget requirements for reliable uplinks.

These results support the conclusion that the CubeSat can operate reliably in real-world APRS use cases, including support for multiple users per pass and storage of data for scheduled downlink to IGates in hybrid missions.

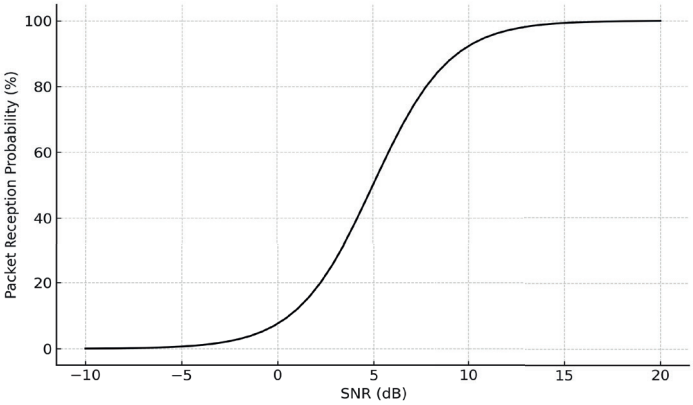


Figure 6. Packet reception probability vs signal-to-noise ratio (SNR)

Discussion. The results obtained from this study demonstrate the practicality of deploying a CubeSat-based APRS digipeater system for wide-area amateur radio communication. The mission design provides meaningful coverage for rural and remote regions and maintains compatibility with existing APRS protocols and equipment.

One of the key takeaways from the analysis is that even a 1U CubeSat, when placed in a 550 km sun-synchronous orbit, can provide reliable access to APRS ground transmitters several times per day. The use of store-and-forward functionality offers increased flexibility, enabling data relay even when no IGate is directly in range during a pass. This functionality makes it feasible to support use cases such as environmental monitoring, agricultural telemetry, and emergency communication in areas with limited terrestrial infrastructure.

The results also highlight important trade-offs, particularly between power availability and communication duty cycle. While the energy budget suffices for basic APRS operation, the system would benefit significantly from energy-efficient components and optimized transmission scheduling.

Furthermore, the selective forwarding of data to designated IGates, as discussed in the system design, enables a semi-private mode of operation, which is crucial for use cases involving sensitive or proprietary data (e.g., land monitoring by private agricultural enterprises).

Conclusion. This paper explored the design and feasibility of deploying an APRS digipeater payload on a CubeSat platform. By integrating low-power communication hardware, a modular onboard computer, and standard VHF APRS protocols, a compact and scalable architecture was proposed.

Through analytical modeling and subsystem-level evaluation, we demonstrated that such a satellite can:

- Support APRS packet reception and retransmission at usable rates,
- Maintain sufficient power levels within a 1U form factor,
- Cover wide geographic areas with multiple daily access windows.

These findings confirm that CubeSat-based APRS relays represent a valuable extension to the terrestrial APRS infrastructure, especially in regions lacking IGate coverage. Future work may focus on flight testing of prototype payloads, security extensions, dual-band operation (VHF/UHF), and long-term constellation planning to enable continuous APRS coverage.

References

Addaim, A., Kherras, A., El Bachir, Z., Abdelhafid Zantou, El Zantou, & Er-Radi, A. (2005) Design of APRS network using low-cost nanosatellite. (in. Eng.)

Addaim, A., Kherras, A., & Zantou, E.B. (2008) DSP implementation of integrated store-and-forward APRS payload and OBDH subsystems for low-cost small satellite. Aerospace Science and Technology, 12(4). — P. 308–317. https://doi.org/10.1016/j.ast.2007.08.002 (in. Eng.)

APRS Working Group (2000) APRS protocol reference version 1.0. Retrieved July 3, 2025. from http://www.aprs.org/doc/APRS101.PDF (in. Eng.)

ARISS. (n.d.). About ARISS. Retrieved July 3, 2025. from https://www.ariss.org/about-ariss.htm l (in. Eng.)

Beech W.A., Nielsen D.E., & Taylor J. (2008) AX.25 link access protocol for amateur packet radio. Retrieved July 3, 2025. from http://www.tapr.org/pdf/AX25.2.2.pdf (in. Eng.)

Bruninga, B. (1999). APRS articles. Retrieved July 3, 2025. from http://www.aprs.org/APRS-docs/ARTICLES.TXT (in. Eng.)

Bruninga B. (2010) APRS articles. Retrieved July 3, 2025, from https://www.aprs.org/GO32-ops.html (in. Eng.)

Bruninga B. (2019) PSAT2 – Amateur radio communications transponders. Retrieved July 3, 2025. from http://aprs.org/psat2.html (in. Eng.)

Chopparapu H.N., Surya Teja D., Sushma N., Vijaya Santhi P., Latha R., Kotamraju S.K., & Chinnari Sri Kavya K. (2025. March 20). Real-time telemetry transmission for CubeSat mission using KL AP – APRS. Journal of Information Systems Engineering and Management, 10(3). — P. 479–488. https://doi.org/10.52783/jisem.v10i3.5340 (in. Eng.)

Linton, G. (2016). Virtualization of CubeSat downlink ground stations using the APRS I-Gate network (Master's thesis). University of Manitoba. https://doi.org/10.13140/RG.2.2.33526.19520 (in. Eng.)

Patmasari R., Wijayanto I., Deanto R.S., Gautama Y.P., & Vidyaningtyas H. (2018) Design and realization of automatic packet reporting system (APRS) for sending telemetry data in Nano satellite communication system. JMECS (Journal of Measurements, Electronics, Communications, and Systems), 4(1). — P. 1–7 (in. Eng.)

Prahyang S.Y., et al. (2018) APRS communication experiment in nanosatellite. IOP Conference Series: Earth and Environmental Science, 149, 012072. https://doi.org/10.1088/1755-1315/149/1/012072 (in. Eng.)

Rizal S., et al. (2021) APRS data receiver using Raspberry Pi in LAPAN-A2 satellite. Spektral, 2(2). — P. 76–82. https://doi.org/10.32722/spektral.v2i2.4250 (in. Eng.)

Salces A.C., Sejera M.P., Kim S., Maui H., & Cho M. (2020) Development and investigation of communication issues on a CubeSat-onboard amateur radio payload with APRS digipeater and store-and-forward capabilities. UNISEC Space Takumi Journal, 9(2). — P. 17–46. (in. Eng.)

Un S., Po K., Thourn K., Pec R., Srun C., & Seven S. (2022) Design of emergency position reporting system for disasters using amateur radio and automatic packet reporting system (APRS) as a mobile station operator for educational purposes. Indonesian Journal of Educational Research and Technology, 3(3). — P. 257–264. https://doi.org/10.17509/ijert.v3i3.58888. (in. Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 258–270

https://doi.org/10.32014/2025.2518-1726.376

UDC 004.4, 004.8, 004.6

N. Temirbekov[1,2], D. Tamabay[1,2,*], S. Kasenov[1,2], A. Temirbekov[1,2], A. Baimankulov[3], 2025.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]National Engineering Academy of the RK, Almaty, Kazakhstan;
[3]Kostanay Regional University named after A. Baitursynuly, Kostanay, Kazakhstan.
E-mail: dtamabay@gmail.com

A WEB-BASED SYSTEM FOR AIR POLLUTION MONITORING WITH API-INTEGRATED DATA SOURCES

Temirbekov Nurlan Mukhanovich — doctor of physical and mathematical sciences, professor of the department of Mathematical and Computer Modeling, Al-Farabi Kazakh National University, 050040, Almaty, Kazakhstan,
E-mail: temirbekov@rambler.ru, ORCID ID: https://orcid.org/0000-0001-7542-3778;
Tamabay Dinara Orazbekkyzy — lecturer of the department of Computational Sciences and Statistics, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: dtamabay@gmail.com, ORCID ID: https://orcid.org/0000-0001-8315-5849;
Kasenov Syrym Erkinovich — head of the department of Mathematics, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: syrym.kasenov@gmail.com, ORCID ID: https://orcid.org/0000-0002-0097-1873;
Temirbekov Almas Nurlanovich — head of the department of Computational Sciences and Statistics, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: almas_tem@mail.ru, ORCID ID: https://orcid.org/0000-0002-4157-2799;
Baimankulov Abdykarim Tungushbayevich — doctor of physical and mathematical sciences, professor of Kostanay Regional University named after A. Baitursynuly, Kostanay, Kazakhstan,
E-mail: bat_56@mail.ru, ORCID ID: https://orcid.org/0000-0002-6435-9560

Abstract. The developed web-based platform for atmospheric air monitoring is a multicomponent system that sequentially collects, processes, and visualizes pollution data. Its information sources are automated monitoring stations and accumulated historical records, which provide real-time access to current concentrations of $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, as well as the Air Quality Index (AQI). A key feature of the platform is the use of cartographic tools to clearly display the spread of pollutants and to analyze their spatiotemporal dynamics. Users can compare current observations with past periods, identify trends, peaks, and episodes of elevated risk without resorting to cumbersome tables. The system generates recommendations for vulnerable population groups based on World Health Organization standards, thereby increasing the practical value of the

environmental data presented. Combining up-to-date measurements with historical time series supports forecasting of changes in the environmental situation and helps to assess potential air-quality deterioration in advance. The resulting assessments, distribution maps, and time-series plots form a basis for developing effective air-quality management strategies and for making environmentally sound decisions at various levels—from the day-to-day actions of individual users to long-term planning aimed at improving urban and natural environments. In this way, the platform links observation data with applied analytics, making information delivery, short-term response, and planning processes more transparent and well-grounded. The platform focuses on air quality and its dynamics in both time and space.

Keywords: information and analytical platform, geographic information system, air pollution, health risk assessment, air quality index

Acknowledgements. *This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant number BR27100483 «Development of predictive exploration technologies for identifying ore-prospective areas based on data analysis from the unified subsurface user platform "Minerals.gov.kz" using artificial intelligence and remote sensing methods»).*

Н. Темирбеков[1,2], Д. Тамабай [1,2,*], С. Касенов [1,2],
А. Темирбеков [1,2], А. Байманкулов [3], 2025.
[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Қазақстан Республикасы Ұлттық инженерлік академиясы,
Алматы, Қазақстан;
[3]А. Байтұрсынұлы атындағы Қостанай өңірлік университеті,
Қостанай, Қазақстан.
E-mail: dtamabay@gmail.com

API-ИНТЕГРАЦИЯЛАНҒАН ДЕРЕККӨЗДЕРІ БАР АТМОСФЕРАЛЫҚ АУАНЫҢ ЛАСТАНУЫН БАҚЫЛАУҒА АРНАЛҒАН ВЕБ-НЕГІЗДЕЛГЕН ЖҮЙЕ

Темирбеков Нурлан Муханович — физика-математика ғылымдарының докторы, әл-Фараби атындағы Қазақ ұлттық университетінің Математикалық және компьютерлік модельдеу кафедрасының профессоры, Алматы, Қазақстан,
E-mail: temirbekov@rambler.ru, ORCID ID: https://orcid.org/0000-0001-7542-3778;
Тамабай Динара Оразбекқызы — әл-Фараби атындағы Қазақ ұлттық университетінің Есептеу ғылымдары және статистика кафедрасының оқытушысы, Алматы, Қазақстан,
E-mail: dtamabay@gmail.com, ORCID ID: https://orcid.org/0000-0001-8315-5849;
Касенов Сырым Еркинович — әл-Фараби атындағы Қазақ ұлттық университетінің Математика кафедрасының меңгерушісі, Алматы, Қазақстан,
E-mail: syrym.kasenov@gmail.com, ORCID ID: https://orcid.org/0000-0002-0097-1873;
Темирбеков Алмас Нурланович — әл-Фараби атындағы Қазақ ұлттық университетінің Есептеу ғылымдары және статистика кафедрасының меңгерушісі, Алматы, Қазақстан,

E-mail: almas_tem@mail.ru, ORCID ID: https://orcid.org/0000-0002-4157-2799;
Байманкулов Абдыкарим Тунгушбаевич — физика-математика ғылымдарының докторы, А. Байтұрсынұлы атындағы Қостанай өңірлік университетінің профессоры, Қостанай, Қазақстан, E-mail: bat_56@mail.ru, ORCID ID: https://orcid.org/0000-0002-6435-9560

Аннотация. Атмосфералық ауаның ластануын бақылауға арналған әзірленген веб-негізделген платформа — бұл ауаның ластануын мониторингтеуге арналған көпқұрамды жүйе болып табылады және ластану жөніндегі деректерді рет-ретімен жинайды, өңдейді әрі көрнекілейді. Экологиялық дереккөздер ретінде автоматтандырылған мониторинг станциялары мен жинақталған тарихи жазбалар қолданылады, бұл пайдаланушыға нақты уақыт режимінде $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2 концентрацияларының өзекті мәндеріне және Ауа сапасының индексіне (AQI) қолжеткізуді қамтамасыз етеді. Платформаның негізгі ерекшелігі — ластаушы заттардың таралуын көрнекі көрсету және олардың кеңістіктік-уақыттық динамикасын талдау үшін картографиялық құралдарды қолдану. Пайдаланушы ағымдағы бақылауларды өткен кезеңдермен салыстырып, үрдістерді, шыңдарды және жоғары тәуекел эпизодтарын көлемді кестелерге жүгінбей-ақ айқындай алады. Жүйе Дүниежүзілік денсаулық сақтау ұйымының стандарттарына сүйене отырып, халықтың осал топтарына арналған ұсынымдарды қалыптастырады, бұл ұсынылатын экологиялық деректердің практикалық маңызын арттырады. Өзекті өлшемдерді тарихи қатарлармен біріктіру экологиялық жағдайдың өзгерістерін болжауды қолдайды және ауа сапасының ықтимал нашарлауын алдын ала бағалауға көмектеседі. Алынатын бағалар, таралу карталары және уақыттық графиктер ауа сапасын басқарудың тиімді стратегияларын әзірлеуге және әртүрлі деңгейлерде — жекелеген пайдаланушылардың күнделікті әрекеттерінен бастап қалалық және қоршаған ортаны жақсартуға бағытталған ұзақ мерзімді жоспарлауға дейін — экологиялық негізделген шешімдер қабылдауға негіз болады. Осылайша, платформа бақылау деректерін қолданбалы аналитикамен байланыстырып, ақпараттандыру, қысқа мерзімді әрекет ету және іс-шараларды жоспарлау үдерістерін неғұрлым ашық әрі негізді етеді; ол ауа сапасына және оның уақыт пен кеңістіктегі динамикасына шоғырланады.

Түйін сөздер: ақпараттық-талдамалық платформа, географиялық ақпараттық жүйе, ауа ластануы, денсаулық тәуекелдерін бағалау, ауа сапасының индексі

Н. Темирбеков[1,2], Д. Тамабай[1,2]*, С. Касенов[1,2], А. Темирбеков[1,2], А. Байманкулов[3], 2025.
[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;
[2]Национальная инженерная академия Республики Казахстан, Алматы, Казахстан;
[3]Костанайский региональный университет имени А. Байтурсынулы, Костанай, Казахстан.
E-mail: dtamabay@gmail.com

ВЕБ-СИСТЕМА МОНИТОРИНГА ЗАГРЯЗНЕНИЯ ВОЗДУХА С API-ИНТЕГРИРОВАННЫМИ ИСТОЧНИКАМИ ДАННЫХ

Темирбеков Нурлан Муханович — доктор физико-математических наук, профессор кафедры математического и компьютерного моделирования Казахского национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: temirbekov@rambler.ru, ORCID ID: https://orcid.org/0000-0001-7542-3778;
Тамабай Динара Оразбекқызы — преподаватель кафедры вычислительных наук и статистики Казахского национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: dtamabay@gmail.com, ORCID ID: https://orcid.org/0000-0001-8315-5849;
Касенов Сырым Еркинович — заведующий кафедры Математики Казахского национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: syrym.kasenov@gmail.com, ORCID ID: https://orcid.org/0000-0002-0097-1873;
Темирбеков Алмас Нурланович — заведующий кафедрой вычислительных наук и статистики Казахского национального университета имени аль-Фараби, Алматы, Казахстан,
E-mail: almas_tem@mail.ru, ORCID ID: https://orcid.org/0000-0002-4157-2799;
Байманкулов Абдыкарим Тунгушбаевич — доктор физико-математических наук, профессор Костанайского регионального университета имени А. Байтурсынулы, Костанай, Казахстан,
E-mail: bat_56@mail.ru, ORCID ID: https://orcid.org/0000-0002-6435-9560.

Аннотация. Разработанная веб-платформа для мониторинга атмосферного воздуха представляет собой многокомпонентную систему, которая последовательно собирает, обрабатывает и визуализирует данные о загрязнении. Источниками информации служат автоматизированные станции мониторинга и накопленные исторические записи, что обеспечивает доступ к актуальным значениям концентраций $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, а также к индексу качества воздуха (AQI) в режиме реального времени. Ключевая особенность платформы — применение картографических средств для наглядного отображения распространения загрязняющих веществ и анализа их пространственно-временной динамики. Пользователь может сопоставлять текущие наблюдения с прошлыми периодами, выявлять тенденции, пики и эпизоды повышенного риска без обращения к громоздким таблицам. Система формирует рекомендации для уязвимых групп населения, ориентируясь на стандарты Всемирной организации здравоохранения, что повышает практическую значимость представляемых экологических данных. Объединение оперативных измерений с историческими рядами

поддерживает прогнозирование изменения экологической обстановки и помогает заранее оценивать возможные ухудшения качества воздуха. Получаемые оценки, карты распределения и временные графики служат основой для выработки эффективных стратегий управления качеством воздуха и принятия экологически обоснованных решений на различных уровнях — от повседневных действий отдельных пользователей до долгосрочного планирования мер по улучшению городской и окружающей среды. Тем самым платформа связывает данные наблюдений с прикладной аналитикой, делая процессы информирования, краткосрочного реагирования и планирования мероприятий более прозрачными и обоснованными. Разработанная платформа фокусируется на качестве воздуха и его динамике во времени и пространстве.

Ключевые слова: информационно-аналитическая платформа, геоинформационная система, загрязнение воздуха, оценка риска для здоровья, индекс качества воздуха

Introduction. In the early 1960s, with the development of technology and increased attention to environmental issues, large-scale introduction of sensors for monitoring atmospheric air quality began. The environmental monitoring system had to quickly detect sudden significant releases of radioactive materials for operational activities, as well as regularly measure the levels of hazardous substances. It should detect radiation and radioactivity levels before the surrounding population is exposed to exposure exceeding established limits.; This means that the higher the dose, the earlier the detection should be performed. It also implies the need for prompt notification and appropriate measures to protect the public and the environment. The research results (Ishihara, 1967) showed that this monitoring system functioned in the following three ways: (1) centralized continuous automatic monitoring; (2) regular surveys of radiation levels; and (3) environmental sample analysis, especially for use as biological indicators. The selection of suitable biological indicators for monitoring environmental pollution and the development of methods for measuring changes in low dose levels, for example, 10^{-6} rad × hr^{-1}, were important before designing an effective environmental monitoring system. Three different centralized continuous monitoring systems for nuclear installations were presented in the work.

The paper (Mukaro et al., 1999) describes a battery-powered, microcontroller-based DAS for remote solar radiation monitoring. It uses an ST62E20 microcontroller and a SolData pyranometer, operating in low-power mode with 10-minute data sampling. Data is stored in EEPROM and transferred via RS232 for offline analysis. The system prioritizes efficiency and cost-effectiveness, showing ±13 W/m^2 accuracy in field tests.

A study (Lodwick et al., 1981) at the University of New South Wales has shown that changes in the environment can be effectively measured using LANDSAT images. A set of programs has been developed for the CYBER 72/171 system that process data from MSS tapes provided by NASA. The results were used to monitor

droughts, predict forest fires, and evaluate crops. Programs written in Fortran used minimal memory requirements and provided image processing in five hours of processor time.

The work (Morita et al., 1983) is devoted to monitoring the environment of the Japanese Tokai and Oarai regions, where there were seventeen enterprises related to nuclear energy, varying in size and type of activity. Environmental monitoring was carried out with the cooperation of the Government, local authorities and enterprises in accordance with plans developed by the prefectural monitoring Committee. The main monitoring objectives included: 1) assessment of the dose to the public based on data on environmental radioactivity and emissions of radioactive substances; 2) determination of the accumulation of radioactive materials over long time intervals. 3) Early detection of abnormal emissions from enterprises.

As can be seen from papers above, the importance of environmental monitoring and the development of tracking systems was recognized back in the early 1960s. This time became the starting point for the implementation of systems that made it possible to monitor the state of the environment and respond promptly to changes in it. Today, thanks to advances in high technology, environmental monitoring has become more efficient and accurate. Modern sensors allow for round-the-clock monitoring without having to worry about data storage. The received data can be processed using cloud computing and transmitted over long distances without obstacles.

The study (Lee et al., 2024) uses geoinformation system and isochronous mapping to assess air quality risks for vulnerable populations in Gangseo, Seoul, South Korea. It identifies underserved areas with limited monitoring, proposing strategic sites for air quality stations. Findings offer policymakers recommendations for better environmental protection and monitoring infrastructure.

The article (Chen et al., 2008) highlights the importance of information systems as a key technological factor capable of contributing to environmental sustainability. The presented conceptual model and the proposed provisions reveal the potential roles of IP in sustainable development strategies, providing a scientific basis for future research in this area.

The article (Ghaemi et al., 2009) presents a web platform for interactive environmental planning in Southern California's Green Visions Plan. It helps municipalities identify project sites for parks, biodiversity, and watershed improvements. Featuring a site-level park analysis tool, it estimates user demographics. A client-server model ensures efficient geospatial data processing.

The article (Paynter et al., 1998) examines the process of developing web-based information systems, including the stages of design, implementation and security. The study highlights the importance of using modern web technologies to effectively manage data and improve information exchange processes in various fields.

The article (Giglione et al., 2022) is devoted to the development of an integrated Geographic Information System-based web platform (GIS) for monitoring the environmental status associated with industrial emissions. The paper presents

an approach to the use of modern technologies for the collection, analysis and visualization of air pollution data related to the activities of industrial enterprises. The main focus is on the functionality of the platform, which provides users with access to real-time data and allows them to analyze pollution levels and their dynamics.

The article (Culshaw et al., 2006) describes a web platform designed to support decision-making in planning in the UK. It provides access to environmental information by integrating data from URGENT and ODPM research. The system covers key planning functions and includes 11 thematic areas such as air quality, groundwater protection and flood risk. The use of geoinformation technologies and modular architecture make it possible to take into account new legal requirements and local data. The study demonstrates the effectiveness of e-governance principles, offering economic and operational benefits to planners and developers.

In summary, it can be stated that information ecology is a newly emerging field that has been actively developing in recent years, and research in this area represents a multidisciplinary subject (Wang et al., 2017). Studies on information ecology primarily focus on information ecosystems, information ecology in e-commerce, and information ecology in networks.

The purpose of this research is to develop an information and analytical platform for monitoring the atmospheric air of industrial cities in Kazakhstan, which receives and analyzes air condition data from automated monitoring systems (AMS). This system provides the collection, transmission and processing of information in real time, which allows you to quickly monitor changes in the level of pollution in the atmosphere and provide recommendations to the public. Developed website *https://aipol.kz/* is an integrated platform that combines data on meteorological conditions and pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2 and CO with the calculation of the Air Quality Index (AQI).

Materials and Methods. Information and analytical systems are an important tool for monitoring the environment and supporting decision-making. They provide environmental data analysis to identify trends, anomalies, and forecast changes, which contributes to a deeper understanding of ecosystem processes and reduces risks to public health.

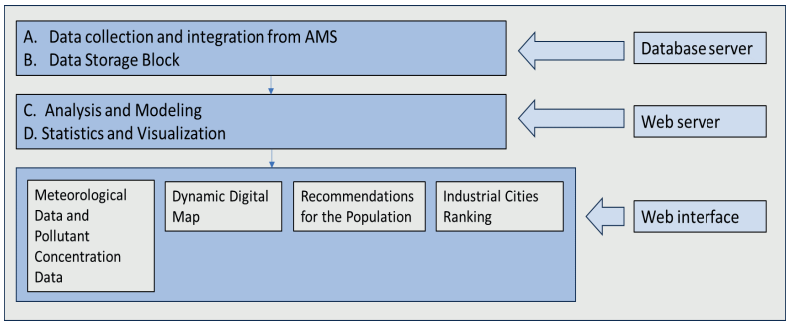


Figure 1. The structure of the information and analytical platform

Platforms increase the availability of environmental information by contributing to information and environmental education for citizens, as well as providing government agencies with informed data for environmental policy development.

The integration of GIS technologies enables comprehensive visualization and spatial representation of environmental data, significantly enhancing the ability to analyze patterns, trends, and anomalies in air pollution distribution. By leveraging GIS-based spatial analysis, researchers and decision-makers can identify high-risk areas, assess pollution sources, and evaluate the effectiveness of mitigation strategies with greater accuracy.

In the context of scientific research, web-based platforms that incorporate GIS technologies facilitate access to extensive datasets collected from various automated monitoring systems, remote sensing technologies, and open-source databases (Figure 1). This integration supports interdisciplinary collaboration by enabling researchers from different fields - such as environmental science, meteorology, public health, and urban planning - to utilize shared geospatial data for developing predictive models, conducting impact assessments, and formulating evidence-based policies. Furthermore, the ability to integrate real-time and historical data enhances the capacity for temporal analysis, enabling a deeper understanding of pollution dynamics and long-term environmental changes.

Data collection and integration. The air pollution data integration module, located in the database server, includes several key components, each of which performs important functions to ensure efficient information collection and processing. Its structure includes an Application Programming Interface (API) client, a data processing unit, and a storage interface.

The API client is responsible for executing HTTP requests to the pollution data API from AMS of Ecoservice-S LLP, managing authentication, and configuring request parameters. It provides reliable and secure data transfer between the platform and an external information source. The data processing unit accepts the received responses in JSON format, cleans them of unnecessary information and converts them into a format suitable for future use. In addition, this component performs error checking and data validation, which guarantees the accuracy and integrity of the information.

From a functional point of view, the module periodically sends API requests to get up-to-date pollution data. It is configured to collect information on key indicators such as $PM_{2.5}$, NO_2 concentrations, and AQI values. To ensure the timely receipt of data, a query planning mechanism is used via cron or special tasks, which allows you to regularly update information about the state of atmospheric air.

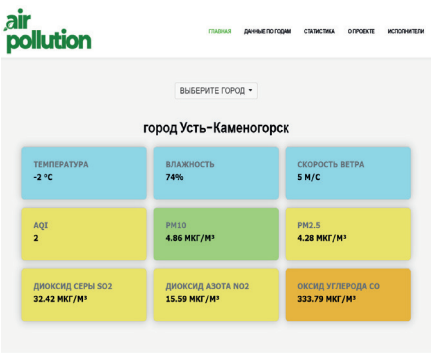
After processing, all data is stored in a database for further analysis and visualization. The module also includes notification and reporting functions: in case of missing data from the remote server or other problems, the system sends notifications. This integration automates the data collection and analysis process, ensuring real-time monitoring of air quality with minimal delays and maximum accuracy.

Data storage block. The created atmospheric pollution database for the industrial city of Ust-Kamenogorsk is designed to generate long-term historical records of the state of the atmosphere, identify the most dangerous pollutants and accumulate metadata from various sources. The system allows you to securely store and process information on server hosted in the LLP Akademset data center. The developed geoinformation system provides access to up-to-date and archived data from mobile and stationary automated sensors installed in the most polluted areas of cities.

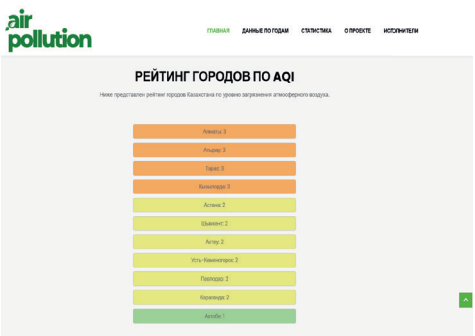
The data repository of the information-analytical platform stores spatiotemporal analysis of pollution, providing regulatory authorities with tools for informed decision-making in environmental safety and sustainable development. The data are available for scientific research on environmental conditions and facilitate ecological monitoring and natural resource management.

Real-time displayed data. The site is integrated with external APIs that provide real-time data on weather conditions and air pollution levels. One of these sources is the Pollution Data API of Ecoservice-S LLP, which aggregates data from automated air quality monitoring stations operated by the company. Additionally, data from other organizations' monitoring stations are collected and made available via API of the OpenWeatherMap platform. This integration enables comprehensive environmental monitoring by ensuring access to accurate and up-to-date atmospheric data.

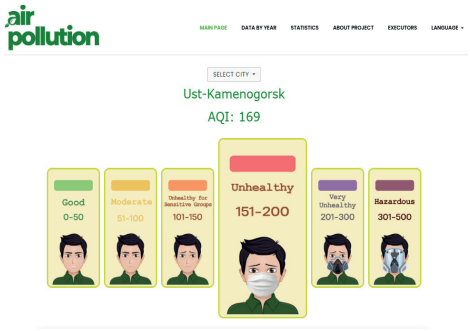
The website interface provides the following information: on the main page of the platform, meteorological parameters, concentrations of pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2) and AQI are presented, which makes it possible to assess the level of pollution and its impact on human health (Figure 2 (a)-(c)). It also displays recommendations for the population (Temirbekov et al., 2023, (a)), in accordance with World Health Organization (WHO) global air quality standards, an AQI rating of cities, and a frequently asked questions section to improve navigation and accessibility of information.



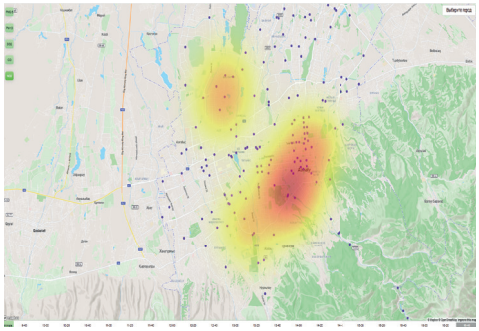
(a)



(b)



(c)



(d)

Figure 2. The website interface of the information and analytical system

Figure 2(c) presents recommendations aimed at informing citizens about possible health risks at different levels of atmospheric air pollution and providing adapted precautions for different population groups.

The recommendation system for the population, implemented within the developed platform, is designed in accordance with internationally recognized air quality standards. It is based on a comprehensive analysis of environmental parameters and aligns with the global guidelines established by the WHO. This system ensures the provision of scientifically grounded recommendations aimed at mitigating the adverse effects of air pollution on public health, thereby enhancing environmental awareness and promoting informed decision-making among users.

Visualization of pollutant dispersion. The cartographic section is a key tool for visualizing environmental data in real time, including the location of automated monitoring stations, their current indicators and dynamic maps of the dispersion of pollutants in industrial zones of Kazakhstan, which allows for spatial analysis and assessment of local sources of pollution.

For visualization of spread of harmful substances in city, a model of the transport of pollutants in the atmospheric air of industrial cities is used with an accurate determination of the concentration of emissions from manufacturing enterprises, as in (Temirbekov et al., 2023 (b), Temirbekov et al., 2024 (a), Temirbekov et al., 2023 (c), Temirbekov et al., 2024 (b)).

The transport equation is examined, where the initial urban area is transformed into a dimensionless region. Within this dimensionless computational domain $0 \le x \le 1,\ 0 \le y \le 1,\ 0 \le t \le T$:

$$\frac{\partial \varphi_q}{\partial t} + u\frac{\partial \varphi_q}{\partial x} + \upsilon\frac{\partial \varphi_q}{\partial y} = \Delta\varphi_q + \alpha_q\varphi_q + \beta_q + f_q \tag{1}$$

$$\varphi_q(x,y,0) = \varphi_0(x,y),\ \varphi_q(0,y,t) = 0, \varphi_q(1,y,t) = 0, \tag{2}$$

$$\varphi_q(x,0,t) = 0,\ \ \varphi_q(x,1,t) = 0, \tag{3}$$

where φ_q is the concentration of the impurity, u, v are components of wind speed, f_q is the power of the pollution sources, $\Delta\varphi_q = \frac{\partial}{\partial x}\left(\frac{\partial \varphi_q}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial \varphi_q}{\partial y}\right)$, in terms of the concentration of the harmful substance in the impurity, the coefficient β_q is the constant of the rate of formation of the substance, α_q is the constant of the rate of decrease.

Additional information for solving this problem is p_q – data on the values of pollutants of AMS. For this problem, the inverse problem is considered, in which it is required to determine the source based on the data received from the monitoring system. The essence of the inverse problem is to minimize the following Lagrange functional

$$L(f_q) = \int_0^T dt \int_\Omega \left[\frac{\partial \varphi_q}{\partial t} + u\frac{\partial \varphi_q}{\partial x} + v\frac{\partial \varphi_q}{\partial y} - \Delta\varphi_q - \alpha_q \varphi_q - \beta_q - f_q\right]\varphi^* d\Omega + \sum_{i=1}^{n} \lambda_i \int_0^T dt \int_\Omega (p_q - \varphi_q)^2 \delta(\vec{r} - \vec{r}_i)\, d\Omega, \quad (4)$$

where r_j is the radius vector of the location of pollution sources and AMS, $\delta(x)$ is the Dirac delta function, λ_i – the coefficient of preference.

This model takes into account the inaccuracy of data on the concentration and volume of emissions, as well as their changes during photochemical reactions. The proposed approach solves this problem as an inverse problem using the theory of conjugate equations. This provides a more accurate understanding of the impact of industrial facilities on air quality and provides a basis for decision-making in the field of regional environmental policy.

For data visualization, Mapbox and OpenStreetMap maps are used, which allows to track environmental information in real time (Figure 2(d)).

Results and discussion. The study of atmospheric pollution monitoring has gained significant attention in recent years due to the increasing anthropogenic impact on air quality and public health (Okabayashi Miyaji et al., 2021). Numerous specialized platforms provide real-time data on pollutant concentrations, leveraging automated monitoring systems to facilitate continuous observation and early detection of environmental risks. These platforms, including AirKaz.org, IQAir, AQI India, AirNow, SmartEco, etc. serve as valuable tools for public access to air quality indices and pollutant concentrations ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2). However, they primarily focus on data aggregation and visualization without advanced spatial analysis or predictive capabilities.

The developed platform (Air pollution, 2024) introduces a more advanced analytical approach by integrating automated monitoring data with meteorological parameters, enabling spatiotemporal modeling of pollutant dispersion. This methodological framework enhances the ability to identify emission hotspots, track

pollution sources, and assess the impact of atmospheric dynamics on pollutant transport. The visualization module employs dynamic geospatial mapping techniques to represent real-time pollution levels and historical trends, offering an in-depth perspective on environmental fluctuations.

The integration of computational algorithms allows for the real-time assessment of pollution dynamics and the generation of adaptive recommendations for vulnerable population groups, including individuals with respiratory conditions, children, and the elderly. By providing health-oriented advisories and exposure minimization strategies, the platform bridges the gap between environmental science and public health management.

Conclusion. The development of a platform for monitoring atmospheric air quality and informing the public represents a significant step in the field of environmental management and environmental protection. The proposed solutions are based on modern information technologies that ensure the collection, processing and analysis of large amounts of data on air pollution. The integration of automated monitoring stations and the use of historical data allows not only to assess the current state of the environment, but also to identify long-term trends and predict environmental changes.

One of the key features of the platform is the accessibility and visibility of the information provided. Users can monitor the concentrations of pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2) and the Air quality Index in real time, which helps to raise environmental awareness. An important component of the platform is a recommendation system based on WHO international standards, which offers tailored precautions for various population groups, including children, the elderly, and people with chronic diseases.

Visualization of data using cartographic tools provides a deeper understanding of the spatial distribution of pollutants and allows you to quickly identify problem areas. This is especially important for industrial regions, where the concentration of emissions from production facilities significantly affects air quality. The ability to track the spread of pollution and localize sources of emissions supports informed decision-making to develop measures to reduce anthropogenic impact.

Thus, the developed platform becomes an important tool for air quality management and making environmentally significant decisions. Its implementation contributes to improving environmental monitoring, increasing the level of environmental safety and public health. Further development of the platform is expected in the future, including expansion of functionality, integration of additional data sources and improvement of forecasting algorithms, which will increase its efficiency and accuracy of environmental analysis.

References

Ishihara T. (1967) Environmental radiological monitoring system at nuclear installations. Health Physics, Volume 13, No 6. — P. 549–558. https://doi.org/10.1097/00004032-196706000-00002. (in Eng.)

Mukaro R., Carelse X.F. (1999) A microcontroller-based data acquisition system for solar radiation

and environmental monitoring. IEEE Transactions on Instrumentation and Measurement, Volume 48, No 6. — P. 1232–1238. DOI: 10.1109/19.816142 (in Eng.)

Lodwick G. D. (1981) A computer system for monitoring environmental changes in multitemporal Landsat data. Canadian Journal of Remote Sensing, Volume 7, No 1. — P. 24–33. https://doi.org/10.1080/07038992.1981.10855006 (in Eng.)

Morita S. (1983) Environmental Radiation Monitoring System in Tokai and Oarai Areas. Journal of the Atomic Energy Society of Japan, Volume 25, No 10. — P. 801–807. https://doi.org/10.3327/jaesj.25.801 (in Eng.)

Lee J.; Jang J.; Im J.; Lee J.H. (2024) GIS-Based Spatial Analysis and Strategic Placement of Fine Dust Alert Systems for Vulnerable Populations in Gangseo District. Applied Sciences, Volume 14, 10610. https://doi.org/10.3390/app142210610 (in Eng.)

Chen A.J.W., Boudreau M.C., Watson R.T. (2008) Information systems and ecological sustainability. Journal of Systems and Information Technology, Volume 10, No 3. — P. 186–201. https://doi.org/10.1108/13287260810916907 (in Eng.)

Ghaemi P. et al. (2009) Design and implementation of a web-based platform to support interactive environmental planning. Computers, Environment and Urban Systems, Volume 33, No 6. — P. 482–491. https://doi.org/10.1016/j.compenvurbsys.2009.05.002 (in Eng.)

Paynter J., Pearson M. (1998) A case study of the Web-based information systems development. Department of Management Science and Information Systems, University of Auckland, New Zealand. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1795c6dda389e6a5930f9bc8f58aaf7a027be8ee (in Eng.)

Giglione G. et al. (2022) An integrated web-based GIS platform for the environmental monitoring of industrial emissions: preliminary results of the project. Applied Sciences, Volume 12, No 7. — P. 3369. https://doi.org/10.3390/app12073369 (in Eng.)

Culshaw M.G. et al. (2006) The role of web-based environmental information in urban planning—the environmental information system for planners. Science of the Total Environment, Volume 360, No 1–3. — P. 233–245. https://doi.org/10.1016/j.scitotenv.2005.08.037 (in Eng.)

Wang X. et al. (2017) Information ecology research: past, present, and future. Information Technology and Management, Volume 18. — P. 27–39. https://doi.org/10.1007/s10799-015-0219-3 (in Eng.)

Temirbekov N.; Temirbekova M.; Tamabay D.; Kasenov S.; Askarov S.; Tukenova Z. (2023) Assessment of the Negative Impact of Urban Air Pollution on Population Health Using Machine Learning Method. International Journal of Environmental Research and Public Health, Volume 20. — P. 6770. https://doi.org/10.3390/ijerph20186770 (in Eng.)

Temirbekov N.; Temirbekov A.; Kasenov S.; Tamabay D. (2024) Numerical modeling for enhanced pollutant transport prediction in industrial atmospheric air. International Journal of Design & Nature and Ecodynamics, Volume 19, No 3. — P. 917–926. https://doi.org/10.18280/ijdne.190321 (in Eng.)

Temirbekov N.; Malgazhdarov, Y.; Tamabay, D.; Temirbekov, A. (2023). Atmospheric modelling of photochemical transformations of pollutants: Impact of weather conditions and diurnal cycle (Case study: Ust-Kamenogorsk, Kazakhstan). Mathematical Modelling of Engineering Problems, Volume 10, No 5. — P. 1699–1705. https://doi.org/10.18280/mmep.100520 (in Eng.)

Temirbekov, N.; Tamabay, D.; Tanashova M. (2024. Spread of harmful substances in the atmosphere of industrial cities of Kazakhstan: modeling and data refinement. Indonesian Journal of Electrical Engineering and Computer Science, Volume 37, No 1. — P. 636. DOI:10.11591/ijeecs.v37.i1. —P. 636-647 (in Eng.)

Okabayashi Miyaji, R.; Valencia de Almeida, F.; de Oliveira Bauer, L.; Madureira Ferrari, V.; Pizzigatti Corrêa, P. L.; Varanda Rizzo, L.; Prakash, G. (2021). Spatial Interpolation of Air Pollutant and Meteorological Variables in Central Amazonia. Data, Volume 6. — P. 126 (in Eng.)https://doi.org/10.3390/data6120126

Air pollution, Available online: https://aipol.kz/ (accessed on 9 January 2024). (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 271–285

https://doi.org/10.32014/2025.2518-1726.377

IRSTI 70.85.31
UDC 556.3.01:630.2(574)(282.255.5)

A.A. Tlepiyev[1*], A. Mukhamedgali[1], Y.T. Kaipbayev[2], A.N. Kalmashova[2], Y.G. Mukhanbet[2], 2025.
[1]Kazakh-British Technical University, Almaty, Kazakhstan;
[2]Kazakh National Agrarian Research University, Almaty, Kazakhstan.
E-mail: armantlepiev123@gmail.com

SURFACE WATER MONITORING IN KAZAKHSTAN USING NDWI AND RANDOM FOREST: A CASE STUDY OF LAKE AKKOL

Tlepiyev Arman Arystanovich — master student, Kazakh-British Technical University, Almaty, Kazakhstan,
E-mail: armantlepiev123@gmail.com;
Mukhamedgali Adil — PhD, Kazakh-British Technical University, Almaty, Kazakhstan,
E-mail: a.mukhamedgali@kbtu.kz, ORCID iD: https://orcid.org/0009-0003-4437-3757;
Kaipbayev Yerbolat Tolganbayevich — PhD, Associate Professor of the Water Resources and Melioration Department, Kazakh National Agrarian Research University, Almaty, Kazakhstan,
E-mail: yerbolat.kaipbayev@kaznaru.edu.kz, ORCID iD: https://orcid.org/0000-0002-7931-7881;
Kalmashova Ainur Nurlepesovna — PhD, Senior Lecturer of the Water Resources and Melioration Department, Kazakh National Agrarian Research University, Almaty, Kazakhstan,
e-mail: Ainur.Kalmashova@kaznaru.edu.kz, ORCID iD: https://orcid.org/0009-0007-7552-8271
Mukhanbet Yerlan Gabitovich — PhD student, Kazakh National Agrarian Research University, Almaty, Kazakhstan,
E-mail: yerlan.mukhanbet@kaznaru.edu.kz, ORCID iD: https://orcid.org/0009-0006-1365-2042.

Abstract. For nations like Kazakhstan, where dry and semi-arid climates together with human activity put increasing strain on lakes and rivers, monitoring water resources has become ever more crucial in recent years. Accurate and timely information on surface water dynamics is essential for effective water management, environmental protection, and adaptation to climate change. Advances in remote sensing technologies, particularly the use of indices like NDWI and machine learning algorithms such as Random Forest, have significantly enhanced the ability to detect and analyze surface water changes over time. These tools offer scalable, cost-effective solutions for continuous monitoring, especially in remote and vast landscapes typical of Central Asia. This work offers a useful method based on the Normalized Difference Water Index (NDWI) to detect water bodies. Each of the tools we used – QGIS, Python and Google Earth Engine (GEE) – had unique benefits for the work. We applied a supervised Random Forest technique

using several spectral bands and indices to separate water covered from dry areas. Examining seasonal and long term fluctuations in water levels, our main case study was on Lake Akkol in the Zhambyl Region. To grasp their influence on local water dynamics, we also examined information from the Assy and Talas rivers. The consistent and dependable results across platforms underlined the great spatial and temporal heterogeneity of water distribution in the area and supported the need for continuous satellite based monitoring.

Key words: Normalized Difference Water Index, Google Earth Engine, QGIS, Python, water resources, monitoring

Acknowledgments

The author would like to thank the developers of QGIS, Python, and GEE for providing open source tools, and the Copernicus Data Space Ecosystem for access to Sentinel 2 data. Special thanks to colleagues and academic advisers for their support throughout the project.

А.А. Тлепиев[1], А. Мухамедгали [1], Е.Т. Кайпбаев [2], А.Н. Калмашова[2], Е.Ғ. Муханбет[2], 2025.

[1]Қазақстан-Британ техникалық университеті, Алматы, Қазақстан;
[2]Қазақ ұлттық аграрлық зерттеу университеті, Алматы, Қазақстан.
E-mail: armantlepiev123@gmail.com

ҚАЗАҚСТАНДАҒЫ БЕТКІ СУЛАРДЫ NDWI ЖӘНЕ RANDOM FOREST ӘДІСІ АРҚЫЛЫ МОНИТОРИНГІЛЕУ: АҚКӨЛ КӨЛІНІҢ МЫСАЛЫНДА

Тлепиев Арман Арыстанович — Қазақстан-Британ техникалық университетінің магистранты, Алматы, Қазақстан,
E-mail: armantlepiev123@gmail.com,
Мухамедгали Адиль — PhD, Қазақстан-Британ техникалық университеті, Алматы, Қазақстан,
E-mail: a.mukhamedgali@kbtu.kz; ORCID iD: https://orcid.org/0009-0003-4437-3757;
Кайпбаев Ерболат Толганбаевич — PhD, «Су ресурстары және мелиорация» кафедрасының қауымдастырылған профессоры, Қазақ ұлттық аграрлық зерттеу университеті, Алматы, Қазақстан,
E-mail: yerbolat.kaipbayev@kaznaru.edu.kz, ORCID iD: https://orcid.org/0000-0002-7931-7881;
Калмашова Айнур Нурлепесовна — PhD, Су ресурстары және мелиорация» кафедрасының аға оқытушысы, Қазақ ұлттық аграрлық зерттеу университеті, Алматы, Қазақстан,
E-mail: Ainur.Kalmashova@kaznaru.edu.kz, ORCID iD: https://orcid.org/0009-0007-7552-8271;
Муханбет Ерлан Габитович — Қазақ ұлттық аграрлық зерттеу университетінің докторанты, Алматы, Қазақстан,
E-mail: yerlan.mukhanbet@kaznaru.edu.kz, ORCID iD: https://orcid.org/0009-0006-1365-2042.

Аннотация. Қазақстан сияқты құрғақ және жартылай құрғақ климат белдеулерінде орналасқан елдер үшін, адамның шаруашылық әрекетімен қатар, көлдер мен өзендерге түсетін жүктеме жылдан-жылға артып келеді. Сондықтан су ресурстарын бақылау соңғы жылдары айрықша

маңызды бола түсті. Беткі сулардың динамикасы туралы дәл әрі уақытылы ақпарат – су ресурстарын тиімді басқару, қоршаған ортаны қорғау және климаттың өзгеруіне бейімделу үшін аса қажет. Қашықтықтан зондтау технологияларының дамуымен, әсіресе NDWI сияқты индекстер мен Random Forest сияқты машиналық оқыту алгоритмдерінің қолданылуы, уақыт ішінде беткі судағы өзгерістерді анықтау мен талдау мүмкіндіктерін едәуір арттырды. Бұл құралдар, әсіресе Орталық Азияға тән кең және қашық аймақтарда, тұрақты бақылау жүргізу үшін тиімді және үнемді шешімдер ұсынады. Бұл жұмыста су айдындарын анықтау үшін нормаланған су индексіне (NDWI) негізделген тиімді әдіс ұсынылады. Қолданылған әрбір құралдың – QGIS, Python және Google Earth Engine (GEE) – өзіндік артықшылықтары болды. Біз су мен құрғақ жерлерді ажырату үшін бірнеше спектралды арналар мен индекстерді пайдалана отырып, Random Forest бақылаулы әдісін қолдандық. Маңызды зерттеу нысаны ретінде Жамбыл облысындағы Аққөл көлі алынып, су деңгейінің маусымдық және ұзақ мерзімді өзгерістері зерттелді. Сондай-ақ, жергілікті су динамикасына әсерін түсіну мақсатында Асы және Талас өзендері жөніндегі ақпарат та қарастырылды. Әртүрлі платформаларда алынған тұрақты әрі сенімді нәтижелер өңірдегі су таралуының кеңістіктік және уақыттық әркелкілігін көрсетті және спутниктік бақылаудың үздіксіз жүргізілу қажеттілігін дәлелдеді.

Түйін сөздер: нормаланған су индексі (NDWI), Google Earth Engine, QGIS, Python, су ресурстары, мониторинг

А.А. Тлепиев[1], А. Мухамедгали[1], Е.Т. Кайпбаев[2], А.Н. Калмашова[2], Е.Г. Муханбет[2], 2025.

[1]Казахстанско-Британский технический университет, Алматы, Казахстан;
[2]Казахский национальный аграрный исследовательский университет, Алматы, Казахстан.
E-mail: yerbolat.kaipbayev@kaznaru.edu.kz

МОНИТОРИНГ ПОВЕРХНОСТНЫХ ВОД В КАЗАХСТАНЕ С ИСПОЛЬЗОВАНИЕМ NDWI И МЕТОДА СЛУЧАЙНОГО ЛЕСА: НА ПРИМЕРЕ ОЗЕРА АККОЛЬ

Тлепиев Арман Арыстанович — магистрант Казахстанско-Британского технического университета, Алматы, Казахстан,
E-mail: armantlepiev123@gmail.com,
Мухамедгали Адиль — PhD, Казахстанско-Британский технический университет, Алматы, Казахстан,
E-mail: a.mukhamedgali@kbtu.kz, ORCID iD: https://orcid.org/0009-0003-4437-3757;
Кайпбаев Ерболат Толганбаевич — PhD, ассоциированный профессор кафедры «Водные ресурсы и мелиорация», Казахский национальный аграрный исследовательский университет, Алматы, Казахстан,
E-mail: yerbolat.kaipbayev@kaznaru.edu.kz, ORCID iD: https://orcid.org/0000-0002-7931-7881;

Калмашова Айнур Нурлепесовна — PhD, старший преподаватель кафедры «Водные ресурсы и мелиорация», Казахский национальный аграрный исследовательский университет, Алматы, Казахстан,
E-mail: ainur.kalmashova@kaznaru.edu.kz, ORCID iD: https://orcid.org/0009-0007-7552-8271;
Муханбет Ерлан Габитович — докторант Казахского национального аграрного исследовательского университета, Алматы, Казахстан,
E-mail: yerlan.mukhanbet@kaznaru.edu.kz, ORCID iD: https://orcid.org/0009-0006-1365-2042.

Аннотация. Для таких стран, как Казахстан, где засушливый и полузасушливый климат в сочетании с человеческой деятельностью усиливают нагрузку на озера и реки, мониторинг водных ресурсов в последние годы становится все более важным. Точная и своевременная информация о динамике поверхностных вод необходима для эффективного управления водными ресурсами, охраны окружающей среды и адаптации к изменениям климата. Развитие технологий дистанционного зондирования, особенно использование индексов, таких как NDWI, и алгоритмов машинного обучения, например, случайного леса (Random Forest), значительно повысило возможности выявления и анализа изменений поверхностных вод во времени. Эти инструменты обеспечивают масштабируемые и экономически эффективные решения для постоянного мониторинга, особенно в удалённых и обширных ландшафтах, характерных для Центральной Азии. В данной работе предлагается эффективный метод обнаружения водоёмов на основе нормализованного водного индекса (NDWI). Каждый из использованных нами инструментов – QGIS, Python и Google Earth Engine (GEE) – имел свои уникальные преимущества. Мы применили контролируемый метод случайного леса с использованием нескольких спектральных каналов и индексов для разделения водоёмов и сухих участков. Основным объектом исследования стало озеро Акколь в Жамбылской области, на котором мы изучали сезонные и долгосрочные колебания уровня воды. Также были проанализированы данные по рекам Асы и Талас для оценки их влияния на местную водную динамику. Последовательные и надёжные результаты, полученные на разных платформах, подчеркнули высокую пространственную и временную неоднородность распределения водных ресурсов в регионе и подтвердили необходимость постоянного спутникового мониторинга.

Ключевые слова: нормализованный водный индекс (NDWI), Google Earth Engine, QGIS, Python, водные ресурсы, мониторинг

Introduction. Particularly in dry and semi-arid areas like southern Kazakhstan, water is among the most vital natural resources for maintaining life, agriculture, and economic development. But climate change, upstream water diversion, poor irrigation techniques, and growing population pressure have made water availability in the region quite erratic. This fluctuation jeopardizes not only ecosystems but also the livelihoods of those depending on a constant water supply for home consumption and agriculture (Kozykeyeva, 2020).

Monitoring surface water dynamics across vast distances and over several times has proved to depend critically on remote sensing technologies. Among these, the Normalized Difference Water Index (NDWI) is among the most useful spectral indices for satellite image based open water body detection. Originally suggested by McFeeters in 1996 (Straube, 2013), NDWI highlights water features by enhancing the reflectance difference between the green and near infrared (NIR) bands, hence suppressing non water features including soil and plant. Sentinel 2 MSI Level 2A imagery was used in this work to calculate NDWI by means of atmospherically corrected surface reflectance data. NDWI values were computed espe- cially using Band 03 (Green, 560 nm) and Band 8A (Narrow NIR, 865 nm). Downloaded with a spatial resolution of 20 meters from the Copernicus Data Space Ecosystem platform (Gao, 1996), the imagery was focusing on the monitoring of Lake Akkol, a representative inland water body prone to seasonal and interannual fluctuations in the Zhambyl Region of Kazakhstan, this study aims. Important tributaries like the Assy and Talas rivers, which help to sustain water levels in surrounding lakes and marshes, have hydrologic impact across the area.

Three platforms – QGIS (Messager, 2016 and Breiman, 2001) and Google Earth Engine (GEE) – were used to compute NDWI and improve water classification and change detection. Moreover, a Random Forest machine learning model (U.S. Geological Survey,2024) was trained to use spectral indices and bands to distinguish non water from water locations. This combined approach offers a strong, flexible framework for evaluating the strengths and limits of any computational technique as well as for studying water dynamics.

The main goal of this work is to show how combining remote sensing indices with machine learning can improve the accuracy and efficiency of surface water monitoring in areas sensitive to water shortage.

Study Area and Data. The study focuses on Lake Akkol, in the semi arid continental environment of southern Kazakhstan's Zhambyl Region, with hot summers and cold winters. Lake Akkol is a small inland water feature whose water cover varies greatly both seasonally and yearly. Maintaining rural livelihoods, biodiversity, and local agriculture as well as helping to support.

The hydrology of the lake is affected by nearby tributary rivers, particularly the Talas River and Assy River, which are essential seasonal sources of inflow. Apart from surface runoff, these rivers affect the temporal dynamics of the water extent of the lake in seasons of snowmelt and spring rains. As its precipitation and evaporation rates vary as well, the region is a relevant test site for evaluating water resource changes under climatic stress.

For image processing and analysis, the area of interest (AOI) was selected to be a buffer zone around the lake and inflow channels.

This work takes use of freely accessible Sentinel 2 MSI Level 2A satellite photos including surface reflectance data modified for atmospheric impacts, obtained from the Ecosystem site of Copernicus Data Space (Gao, 1996).

Sentinel 2 offers high resolution multispectral data at 10–20 meter spatial

resolution, which is well suited for detecting and monitoring surface water features. Two specific bands were used for water index analysis:

- Band 03 (Green, 560 nm)
- Band 8A (Narrow Near Infrared, 865 nm)

The work focused on just low cloud cover or cloud free models. Long term water dynamics and interannual comparisons were investigated with photos spanning many years: 2016–2024.

This work manually acquired Sentinel 2 photos from the Copernicus Data Space Ecosystem web portal (Gao, 1996). Visual and metadata based filtering was part of the chosen strategy to guarantee low cloud coverage or cloud free conditions over the Lake Akkol area. Scenes were selected depending on:

- **Cloud Coverage:** less than 10% total cloud cover.
- **Spatial Resolution:** 20 meter products (B03 and B8A bands).
- **Temporal Range:** July–August for representative seasonal water extent (2016 and 2024).
- **Level:** Sentinel 2 MSI Level 2A (surface reflectance, atmospherically corrected).

Images were previewed directly in the platform's browser interface using the quick look rendering option and downloaded as SAFE products. Bands were later extracted and preprocessed using Python and QGIS tools (Messager, 2016 and Breiman, 2001) for consistency across platforms.

№1 table – Summary statistics and metadata for Sentinel-2 Bands B8A and B03

Property	B8A	B03
Minimum Value	975	1038
Maximum Value	10077	11420
Mean Value	3536.03	2456.12
Standard Deviation	649.71	712.08
Projection Method	Universal Transverse Mercator (UTM)	
CRS	EPSG:32642 – WGS 84 / UTM zone 42N	

In addition to satellite imagery, the study incorporated vector data from the HydroLAKES database (European Space Agency, 2024), which provides global lake boundary polygons. These prevali-dated lake polygons were used to extract training samples for the Random Forest model, enabling more accurate and consistent labeling of water bodies across different image sources. Pixels located within HydroLAKES polygons were labeled as "water," while pixels outside were used to represent "nonwater" classes.

№2 table – Classification metrics with definitions and formulas

Metric	Purpose	Formula
Accuracy	Overall correctness of predictions. May be misleading with imbalanced classes.	$\frac{TP+TN}{TP+TN+FP+FN}$

Precision	How many predicted positives (e.g. wa- ter) are actually correct. Important when false positives are costly.	$\frac{TP}{TP+FP}$
Recall	Measures how many actual positives are correctly detected. Crucial when missing water areas is critical.	$\frac{TP}{TP+FN}$
F1 score	Balances precision and recall. Reliable metric for imbalanced datasets.	$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$

Where: TP – true positives, FP – false positives, TN – true negatives, FN – false negatives

This approach allowed for the creation of a balanced and geographically diverse training dataset, improving the robustness of the machine learning classification and supporting generalization across different types of water bodies.

The interpretation scale presented in Table 3 is grounded in widely accepted practices for evaluating classification models across domains such as remote sensing, medical diagnostics, and information retrieval. Although there is no universally standardized threshold, the defined value ranges (e.g., Excellent: 90 to 100%, Good: 75 to 84%) are commonly used in applied machine learning literature and supported by evidence from prior studies.

Maxwell et al. (2018) highlight that overall accuracy values above 90% are typically considered high performing in remote sensing classification tasks, especially when the dataset is well prepared and balanced. Similarly, Sokolova and Lapalme (2009) emphasize that metrics such as precision and F1 score provide more nuanced insights in imbalanced classification problems, which are common in environmental data analysis.

№3 table – Qualitative interpretation scale for classification metrics

Interpretation	Value Range (%)
Excellent	90 to 100
Very Good	85 to 89
Good	75 to 84
Moderate	60 to 74
Poor	Below 60

This qualitative scale helps provide intuitive, interpretable guidance when assessing model effectiveness, especially in interdisciplinary contexts where stakeholders may not be familiar with the technical significance of raw metric values. It also supports communication of results to non specialist audiences and aligns with performance reporting practices recommended in tools such as scikit learn (Gillies et al., 2022).

Methods and materials. This study combines remote sensing techniques with

machine learning to detect and monitor surface water dynamics in the Lake Akkol region. The workflow consists of three main components: NDWI calculation using multiple platforms, training data preparation using HydroLAKES polygons (European Space Agency, 2024), and supervised classification using the Random Forest algorithm (Breiman, 2001).

The Normalized Difference Water Index (NDWI) (Straube et.al., 2013) was computed using the following formula:

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR}$$

where:

Green is Band 03 (560 nm),

NIR is Band 8A (865 nm) from Sentinel 2 MSI Level 2A imagery.

To evaluate and compare processing flexibility, speed, and visual output quality, NDWI was calculated using three platforms:

QGIS: NDWI was computed using the Raster Calculator (Messager et.al., 2016). This approach was useful for localized analysis and manual quality control.

Python: A Python script utilizing rasterio, numpy, and matplotlib was developed to automate NDWI calculation and integrate with machine learning pipelines (U.S. Geological Survey, 2024).

Google Earth Engine (GEE): The NDWI was calculated and visualized using GEE's JavaScript API. This method allowed efficient processing of large datasets across multiple years.

To create labeled training data for the classification model, vector polygons from the HydroLAKES global database were used (European Space Agency, 2024). These polygons represent the outlines of known lakes and reservoirs.

Pixels within the HydroLAKES polygons were labeled as water.

Pixels from nearby nonwater areas (e.g., bare soil, vegetation) were labeled as non-water.

The dataset was balanced to reduce bias and ensure model generalization across vary- ing conditions and landscapes.

This semiautomated labeling method significantly improved the consistency and scal-ability of training data generation.

A Random Forest model was trained using the scikit learn Python library (Gillies et al., 2022). The classifier used a combination of spectral features and remote sensing indices to improve classification accuracy.

The following input features were included:

- **NDWI**: Highlights open water by contrasting green and NIR reflectance.
- **NDVI (Normalized Difference Vegetation Index)**, computed using:

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR}$$

Where: NIR is Band 8A, and Red is Band 04 (665 nm) (McFeeters, 1996). NDVI was included to help differentiate water from vegetated areas.

- **Spectral Bands**:
  - B03 (Green)
  - B04 (Red)
  - B08A (Narrow NIR)
  - B11 (SWIR1, 1610 nm)
  - B12 (SWIR2, 2190 nm)

These bands are particularly effective for distinguishing water from built up or dry areas, as SWIR reflectance is typically low over water but high over soil and manmade surfaces.

The model was trained on labeled pixels from the HydroLAKES dataset (European Space Agency, 2024) and validated using independent regions near Lake Akkol. Evaluation was based on:

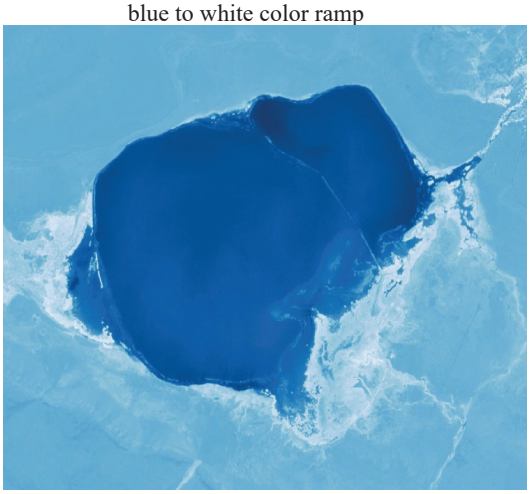
- Confusion matrix
- Overall classification accuracy
- Precision and recall for the water class

The Random Forest algorithm (Breiman, 2001) was chosen due to its ability to handle nonlinear relationships, resistance to overfitting, and strong performance in remote sensing applications.

Results and Discussion

NDWI maps generated using QGIS (Messager et al., 2016) provided a detailed view of water extent in the region. The NDWI raster was visualized using a **single band pseudocolor** rendering style, which enhances contrast between water and nonwater areas. A **blue to white color ramp** was applied, where darker blue tones represent higher NDWI values (closer to water), and white indicates drier or nonwater surfaces.

1- Figure. NDWI map of Lake Akkol generated in QGIS using single band pseudocolor with a blue to white color ramp



The computed NDWI values from the Sentinel 2 image in QGIS ranged from:

- **Minimum:**-0.63236
- **Maximum:** 0.330626

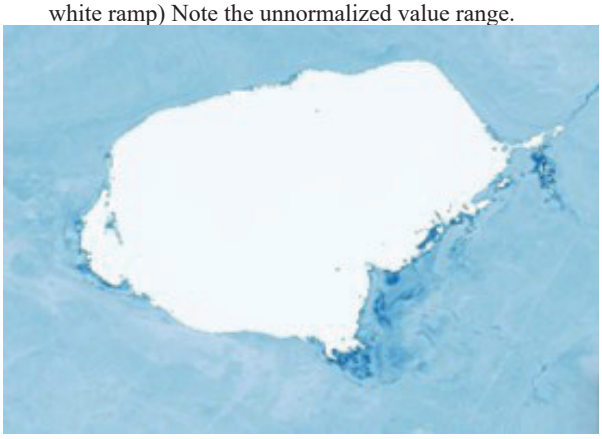
These values align with expected NDWI ranges (Straube et al., 2013) where water typically has NDWI and negative values correspond to dry soil, urban features, or vegetation. The contrast in the image helped clearly delineate Lake Akkol from surrounding land cover.

The NDWI was also calculated using a custom Python script developed and executed in the PyCharm environment. Sentinel 2 Level 2A images were processed using the rasterio library for reading raster bands and numpy for array based calculation of NDWI values (Breiman, 2001). The resulting raster was visualized using matplotlib with a **single band pseudocolor** rendering and a **blue to white color ramp**, similar to the one used in QGIS.

In this case, the raw band values were read as 16 bit integers without applying the standard reflectance scaling factor (typically 0.0001). As a result, the computed NDWI values were not normalized and fell within the range:

- **Minimum:** 0
- **Maximum:** 30.4243

2-Figure. NDWI map of Lake Akkol generated in Python using pseudocolor rendering (blue to white ramp) Note the unnormalized value range.



Although these values deviate from the expected NDWI range of −1 to +1, the structural correctness of the formula was preserved. This discrepancy primarily affects numerical interpretability rather than the visual pattern of water detection.

To obtain physically meaningful NDWI values, raw band values should be converted to surface reflectance by dividing each pixel value by 10,000. The normalized NDWI formula becomes:

$$NDWI = \frac{\frac{B3}{10000} - \frac{B8A}{10000}}{\frac{B3}{10000} + \frac{B8A}{10000}} = \frac{B3 - B8A}{B3 + B8A}$$

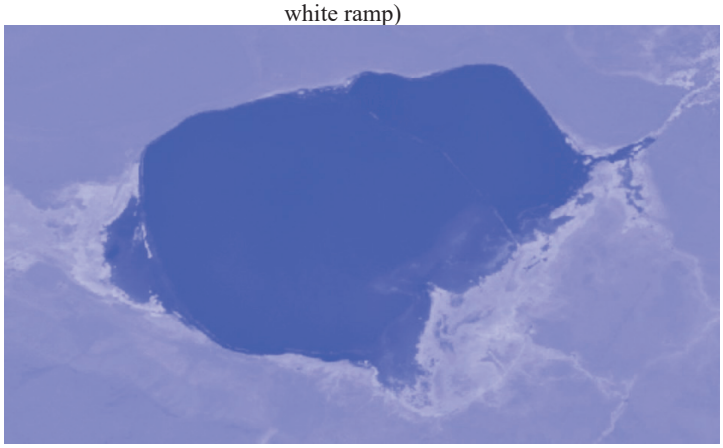
Mathematically, the scale factor cancels out, so the NDWI structure remains valid; however, normalization is crucial when comparing index thresholds or combining with other scaled indices like NDVI (McFeeters, 1996).

Despite the unnormalized range, the pseudocolor visualization remained effective. The blue to white colormap clearly emphasized water features with relatively higher NDWI values, enabling visual distinction of Lake Akkol from its surroundings.

This scenario highlights the importance of reflectance normalization when working with Sentinel 2 data in Python environments. Applying normalization ensures consistency with outputs from QGIS and Google Earth Engine, which automatically handle scaled reflectance.

Google Earth Engine (GEE) results: GEE allowed for direct access to atmospherically corrected Sentinel 2 imagery with automated scaling. NDWI was calculated using the standard formula and visualized using the same blue to white colormap. The results were clipped to the Akkol region.

3-Figure. NDWI map of Lake Akkol generated in GEE using pseudocolor rendering (blue to white ramp)



The NDWI values in GEE ranged from:

- **Minimum:** -0.6008
- **Maximum:** 0.3493

These values closely matched those observed in QGIS, confirming consistency between platforms that correctly handle reflectance.

№4 table – Classification Metrics and Their Formulas

Metric	Formula	Value
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	≈ 0.937
Precision	$\frac{TP}{TP+FP}$	≈ 0.854

Recall	$\frac{TP}{TP+FN}$	≈ 0.966
F1 score	$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$	≈ 0.906

№5 table – Comparison of NDWI Value Ranges Across Methods

Method	Minimum NDWI	Maximum NDWI
QGIS	−0.63236	0.33063
Python (unnormalized)	0	30.4243
GEE	−0.60081	0.34934

Classification Results Using Random Forest

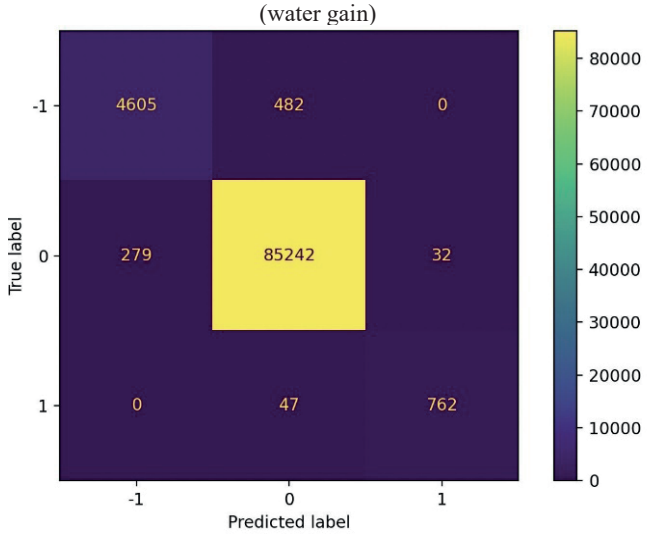
To evaluate surface water change detection in Lake Akkol, a supervised Random Forest classifier was trained using NDWI, NDVI, and spectral bands (B04, B11, B12). The model was trained using labeled water polygons from the HydroLAKES dataset (European Space Agency, 2024) focusing on three classes:

- -1 – water loss
- 0 – no change
- 1 – water gain

1. Multiclass Classification Performance

The confusion matrix in Figure 4 shows the classifier's performance on the full 3 class dataset. The matrix indicates a high number of correctly predicted stable water areas (class 0), as well as good performance in detecting both water appearance and disappearance.

Figure 4. Confusion matrix for 3 class Random Forest classification: -1 (water loss), 0 (stable), 1 (water gain)



Key values:

Class 0 (stable water): TP = 5,242

Class -1 (loss): TP = 4,605

Class 1 (gain): TP = 762

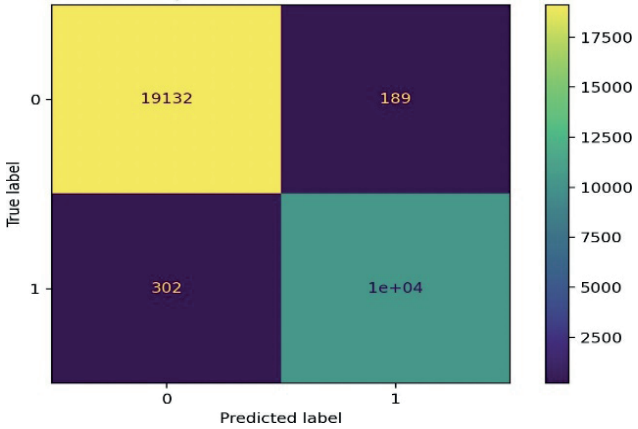
The matrix confirms that the classifier performs well across all classes, with low false positives and minimal confusion between gain and loss.

2. Binary Classification – Simplified Tasks

To compare performance under binary conditions, the model was evaluated in two simplified settings.

α) **Detecting only class 1 (water gain):**

Figure 5. Binary classification for class 1 (water gain) vs others



TP (gain correctly detected): 10,000

FP (false gain): 189

FN (missed gain): 302

β) **Detecting only class 1 (gain) with unbalanced dataset:**

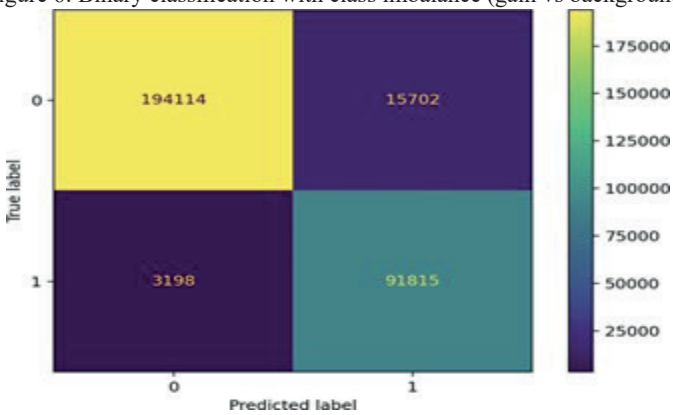
TP = 91,815

FP = 15,702

FN = 3,198

The model still maintains high recall, but precision decreases due to the high imbalance. This phenomenon is consistent with findings by Sokolova and Lapalme (Sokolova and Lapalme, 2009) who emphasize the importance of balancing precision and recall when evaluating classifiers under skewed distributions.

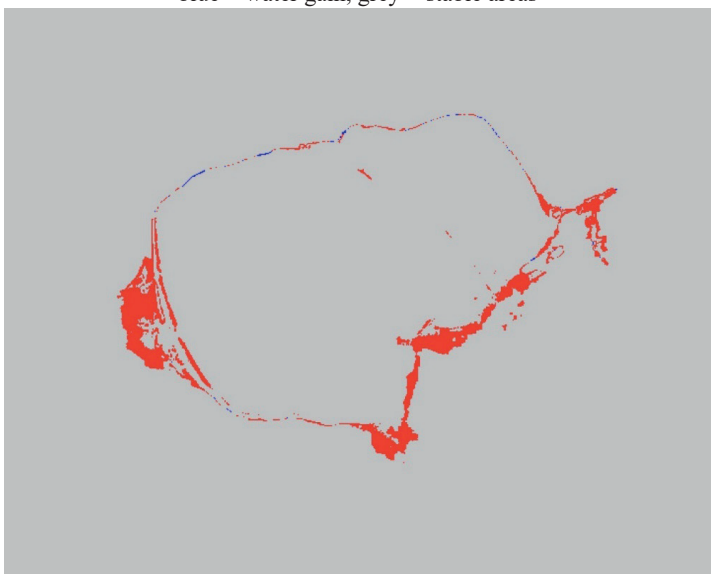
Figure 6: Binary classification with class imbalance (gain vs background)



3. Visualizing Water Change Classification

In addition to numeric metrics, the classified image was visualized as shown in Figure 7. Water changes around Lake Akkol are clearly captured, particularly shoreline expansion and contraction.

Figure 7. Classified water change map 2016–2024: red = water loss, blue = water gain, grey = stable areas



Conclusion

This paper shows how well NDWI combined with machine learning monitors water bodies. Using multiple platforms (QGIS, Python, and Google Earth Engine) offers flexibility, while the Random Forest classifier improves classification reliability. With their insights on regional water dynamics, Lake Akkol and its

tributaries provide a valuable test ground for such studies. Future work may include precipitation integration, time series analysis, and extension to include wetland categorization and water quality indicators.

References

Breiman L. (2001) Random forests. Machine Learning, 45(1). — P.5–32. (in Eng.)

European Space Agency (ESA). (2024). Copernicus Data Space Ecosystem. Available: *https://dataspace.copernicus.eu (in Eng.)*

Gao B.C. (1996) NDWI – A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment, 58(3). — P. 257–266. (in Eng.)

Gorelick N., Hancher M., Dixon M., Ilyushchenko S., Thau D., Moore R. (2017) Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202. — P.18–27. Platform available:*https://earthengine.google.com (in Eng.)*

Gillies S. et al. (2022) Rasterio: geospatial raster I/O for Python programmers. Available: *https://rasterio.readthedocs.io (in Eng.)*

HydroLAKES database: https://www.hydrosheds.org/products/hydrolakes (in Eng.)

Kozykeyeva A.T., Mustafayev Z.S., Aldiyarova A.E., Arystanova A.B., Mosiej J. Methodical support of integrated management of water resources of the basin of transboundary rivers (2020) News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences, 4 (442) — P.52 – 61. DOI: 10.32014/2020.2518-170X.84 (in Eng.)

McFeeters S.K. (1996) The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7). — P.1425–1432. (in Eng.)

Messager M.L., Lehner B., Grill G., Nedeva I., Schmitt O. (2016) Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature Communications, 7. — P.1-11. *https://doi.org/10.1038/ncomms13603* (in Eng.)

Maxwell A. E., Warner T. A., Fang F. (2018) Implementation of machine learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9). — P.2784–2817. (in Eng.)

OpenStreetMap contributors (2024) Planet Dump. Retrieved from *https://planet.openstreetmap.org* (in Eng.)

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., ... Duchesnay E. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12..— P.2825–2830. (in Eng.)

QGIS Development Team (2024) QGIS Geographic Information System. Open Source Geospatial Foundation Project. Available: *https://qgis.org (in Eng.)*

Straube F., Handfield R., Pfohl H.S. & Wieland A. (2013) Trends and strategies in logistic and supply chain management. Hamburg, Germany: Deutscher Verkehrs-erlag. Books. — P.1-16. (in Eng.)

Sokolova M., Lapalme G. (2009) A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4). — P. 427–437. U.S.Geological Survey (USGS). (2024). EarthExplorer Data Portal. Available: *https://earthexplorer.usgs.gov (in Eng.)*

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 286–300

https://doi.org/10.32014/2025.2518-1726.378

ҒТМР 14.09.19
ӘОЖ 378.147

Z. Turysbek[1], O. Mamyrbayev[2], M. Abdullah[3], 2025.
[1]Kazakh national research technical university named after K.I. Satpayev, Almaty, Kazakhstan;
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan;
Zhengzhou University, Zhengzhou, Henan, China.
Email: janibekturysbek@gmail.com

DEVELOPMENT OF AN INTELLIGENT SYSTEM FOR DETECTING FAKE NEWS

Zhanibek Turysbek — postgraduate student, Kazakh national research technical university named after K.I. Satpayev, Almaty, Kazakhstan,
E-mail: janibekturysbek@gmail.com, ORCID ID: https://orcid.org/0009-0004-2311-6249;
Mamyrbayev Orken — Doctor PhD, Professor, Institute of Information and Computational Technologies, Almaty, Kazakhstan,
E-mail: morkenj@mail.ru; ORCID ID: https://orcid.org/0000-0001-8318-3794 ;
Muhammad Abdullah — School of Computing and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan, China,
Email: Arifa.javed@gs.zzu.edu.cn. ORCID ID: https://orcid.org/0009-0000-9434-7977.

Abstract. This article discusses the development of an intelligent system for detecting fake news. The widespread adoption of digital technologies has facilitated the mass dissemination of information, leading to an increase in informational noise. The spread of fake news causes confusion in society and contributes to social instability. The broad distribution of such false information poses a significant threat not only to public stability but also to national information security. Therefore, we have undertaken the development of this intelligent system. The developed software application aims to help prevent such issues from arising. Easy access to the Internet and the popularity of social media in Kazakhstan contribute to the rapid spread of fake news. Additionally, the grammatical and semantic characteristics of the Kazakh language make it difficult for international fake news detection systems to be directly adapted to the local context. We propose a grounded method for detecting fake news on social media using BERT (Bidirectional Encoder Representations from Transformers), which not only determines whether a piece of news is true or false but also provides interpretability of the decision by highlighting suspicious users and key evidence words. The applied model demonstrated good performance in detecting fake news in the Kazakh language, while the proposed

intelligent system contributes to limiting the spread of fake news in the national information space, strengthening information security, and assisting both experts and ordinary users in identifying false information.

Keywords: Fake news, real news, intelligent system, NLP (Natural Language Processing); Semantics; Social media, software application

Ж. Тұрысбек[1], О.Ж. Мамырбаев[2], А. Мұхаммед[3], 2025.
[1]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан;
[2]Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан;
[3]Чжэнчжоу университеті, Чжэнчжоу, Хэнань провинциясы, Қытай.
Email: janibekturysbek@gmail.com.

ЖАЛҒАН ЖАҢАЛЫҚТАРДЫ АНЫҚТАЙТЫН ИНТЕЛЛЕКТУАЛДЫ ЖҮЙЕНІ ӘЗІРЛЕУ

Жәнібек Тұрысбек — докторант, Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан,
E-mail: janibekturysbek@gmail.com, ORCID ID: https://orcid.org/0009-0004-2311-6249;
Мамырбаев Өркен Жұмажанұлы — PhD докторы, профессор, Ақпараттық және есептеу технологиялары институты, Алматы, Қазақстан,
E-mail: morkenj@mail.ru, ORCID ID: https://orcid.org/0000-0001-8318-3794;
Мұхаммед Абдулла — Есептеу және жасанды интеллект мектебі, Чжэнчжоу университеті, Чжэнчжоу, Хэнань провинциясы, Қытай,
E-mail: Arifa.javed@gs.zzu.edu.cn, ORCID ID: https://orcid.org/0009-0000-9434-7977.

Аннотация. Бұл мақалада жалған жаңалықтарды анықтайтын интеллектуалды жүйені әзірлеу қарастырылған. Себебі, цифрлық технологиялардың кеңінен таралуы ақпараттың жаппай таралуына жағдай туғызды және ақпараттық шу деңгейінің жоғарылауына әкелді. Жалған жаңалықтардың таралуы қоғамдағы шатасуды, әлеуметтік тұрақсыздықты тудырады. Мұндай жалған ақпараттың кең таралуы қоғамдық тұрақтылыққа ғана емес, сонымен бірге ұлттық ақпараттық қауіпсіздікке де үлкен қауіп төндіреді. Сол себепті осы интелектуалды жүйені әзірлеуді қолға алдық. Бұл құрылған программалық қосымша осы жайттардың туындауының алдын алуға мүнкіндік береді. Интернетке кең қол жетімділік және Қазақстандағы әлеуметтік желілердің танымалдығы жалған жаңалықтардың тез таралуына мүмкіндік береді. Сонымен қатар, қазақ тілінің грамматикалық және семантикалық ерекшеліктері жалған жаңалықтарды анықтаудың Халықаралық жүйелерінің жергілікті контекстке тікелей бейімделуін қиындатады. Біз жалған жаңалықтарды әлеуметтік желілерде анықтауға арналған негізделген әдісті – **BERT** (Bidirectional Encoder Representations from Transformers) ұсынамыз, мұнда жаңалықтың жалған немесе шын екендігін анықтау ғана емес, сонымен қатар, шешімді түсіндіру (интерпретациялау) арқылы күдікті қолданушылар

мен дәлелді сөздерді ерекшелеуге болады. Қолданылған модель қазақ тіліндегі жалған жаңалықтарды анықтауда жақсы нәтиже көрсетті, ұсынылған интеллектуалды жүйе ұлттық ақпараттық кеңістікте жалған жаңалықтардың таралуын шектеуге, ақпараттық қауіпсіздікті күшейтуге және сарапшылар мен қарапайым қолданушыларға жалған ақпаратты тануда септігін тигізеді.

Түйін сөздер: жалған жаңалық, ақиқат жаңалық, интелектуалды жүйе, NLP (Табиғи тілді өңдеу), семантика, әлеуметтік желі, программалық қосымша

Алғыс. *Бұл жұмыс Қазақстан Республикасы Ғылым және жоғары білім министрлігі тарапынан BR24993001 «Қазақ тілі мен технологиялық прогресті қолдау үшін үлкен тілдік моделін (LLM) құру» жобасы негізінде қолдау тапты.*

Ж. Тұрысбек[1], О.Ж. Мамырбаев[2], А. Мұхаммед[3], 2025.
[1]Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Қазақстан;
[2]Институт информационных и вычислительных технологий, Алматы, Қазақстан;
[3]Университет Чжэнчжоу, Чжэнчжоу, Хэнань, Китай.
Email: janibekturysbek@gmail.com

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ОБНАРУЖЕНИЯ ФЕЙКОВЫХ НОВОСТЕЙ

Тұрысбек Жәнібек — Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Қазақстан,
E-mail: janibekturysbek@gmail.com, ORCID ID: https://orcid.org/0009-0004-2311-6249;
Мамырбаев Оркен Жумажанович — Институт информационных и вычислительных технологий, Алматы, Қазақстан,
E-mail: morkenj@mail.ru, ORCID ID: https://orcid.org/0000-0001-8318-3794;
Мұхаммед Абдулла — Школа вычислительной техники и искусственного интеллекта, Университет Чжэнчжоу, Чжэнчжоу, Хэнань, Китай,
E-mail: Arifa.javed@gs.zzu.edu.cn. ORCID ID: https://orcid.org/0009-0000-9434-7977.

Аннотация. В данной статье рассматривается разработка интеллектуальной системы для обнаружения фейковых новостей. Причиной является широкое распространение цифровых технологий, которое способствовало массовому распространению информации и увеличению уровня информационного шума. Распространение фейковых новостей вызывает замешательство в обществе и приводит к социальной нестабильности. Широкое распространение такой ложной информации представляет серьезную угрозу не только общественной стабильности, но и национальной информационной безопасности. Поэтому мы приступили к разработке данной интеллектуальной системы. Созданное программное приложение позволяет предотвращать возникновение таких

проблем. Широкий доступ к Интернету и популярность социальных сетей в Казахстане способствуют быстрому распространению фейковых новостей. Кроме того, грамматические и семантические особенности казахского языка затрудняют прямую адаптацию международных систем обнаружения фейковых новостей к местному контексту. Мы предлагаем основанный метод для обнаружения фейковых новостей в социальных сетях с использованием модели BERT (Bidirectional Encoder Representations from Transformers), который не только определяет, является ли новость ложной или правдивой, но и обеспечивает интерпретируемость решения, выделяя подозрительных пользователей и ключевые доказательные слова. Применённая модель показала хорошие результаты в выявлении фейковых новостей на казахском языке, а предлагаемая интеллектуальная система способствует ограничению распространения ложной информации в национальном информационном пространстве, укреплению информационной безопасности и помогает как экспертам, так и обычным пользователям в распознавании фейковых новостей.

Ключевые слова: фейковые новости, правдивые новости, интеллектуальная система, NLP (обработка естественного языка); семантика; социальные сети, программное приложение

Кіріспе. Қазіргі таңда елімізге цифрлық технологиялардың енуіне және әлеуметтік желілердің кеңінен таралуына байланысты ақпарат тарату үрдісін жеделдетіп, қоғамда жалған ақпараттың таралуын және әлеуметтік тұрақтылыққа қауіп төндірді. Мұндай қауіптің алдын алу үшін қазіргі таңда кеңнен қолданыладытын жасанды интеллект жүйелерін, соның ішінде табиғи тілді өңдеудің (Natural Language Processing, NLP) тиімді әдістерін қолдану өте маңызды рөл атқарады. Сондықтан ұлттық ерекшеліктерді ескере отырып, қазақ және орыс тілдеріне бағытталған интеллектуалды жүйелерді дамыту өзекті мәселеге айналуда. Келесі ғалымдар, атап айтқанда *Daniela Gîfu* «International Journal of Advanced Computer Science and Applications» журналында "An Intelligent System for Detecting Fake News" атты тақырыпты мақаласында қарастырды (Gîfu D. 2023). Сонымен қатар, Rohit Kumar Kaliyar, Anurag Goswami. Pratik Narang деген ғалымдар «Multimedia Tools and Applications» деген ғылыми журналда "FakeBERT: Әлеуметтік желідегі жалған жаңалықтарды анықтауға арналған BERT негізіндегі терең оқыту тәсілі" тақырыбын зерттеді (Kaliyar et al., 2021). Бұл ғаламдардан басқада ғалымдар мысалы International Research Journal of Engineering and Technology (IRJET) журналда Anshu Aditya, B.V.S.S.Vardhan, D.S.Chanakya Varma, P.Kailashnadh Gupta, Dr Venkat, Ramana M деген ғалымдар "Fake News Detection Using BERT" тақырыбында ғылыми зерттеу жұмыстарын жүргізді (Aditya et al., 2024).

Соңғы жылдары Google компаниясы ұсынған BERT (Bidirectional Encoder Representations from Transformers) моделі табиғи тілді мағыналық тұрғыдан

талдауда өте тиімді және жоғары нәтижелер көрсетіп келеді. Модельдің ең ерекшелігі контекстті екі бағытта солдан оңға және оңнан солға қарай сараптап, сөздердің мағынасын түсініп ұғындыруға мүмкіндік береді. Бұл модель BERT-ті жалған жаңалықтарды автоматты түрде анықтау жүйелерінде қолдануға мүнкіндік береді. Дегенмен, қазақ тіліне арналған мәтіндермен жұмыс істеуде көптеген қиындықтар туындайды. Қазақ тілі – морфологиялық тұрғыдан бай тіл болғандықтан, BERT моделін қазақ тіліндегі мәтіндерге бейімдеу өте маңызды бағыттар болып табылады (Fine-tuning) – жалған жаңалықтарды анықтаудың тиімді жолдарының бірі ретінде қарастырыуға болады (Mamyrbayev et al., 2025). Мен осы ғылыми жұмыста жалған жаңалықтарды анықтау мәселелерін шешу үшін BERT моделін пайдалануды дұрысдеп қарастырдым. Зерттеу барысында модельдің архитектурасы мен жұмыс істеу принцптерін, қазақ тіліне бейімдеу әдістерін, сондай-ақ нақты мәліметтер негізінде жүргізілген бақылау нәтижелері сипатталады.

Материалдар мен әдістер. Әлеуметтік желілердегі жаңалықтың таралу жолын модельдеу өте маңызды бұл таралып жатқан жаңалықтардың жалған екендігін анықтауға үлкен әсерін тигізеді, мұнда пайдаланушылар саны, уақыт, қайталап жариялау тізбегі маңызды рөл атқарады осы бағытта келесі формализацияны ұсыныуға болады.

Осы зерттеуде ұсынылған үлгіні бұрынғы тәсілдермен салыстыру үшін дәлдік (D), нақтылық (precision), қайтарым (recall) және F1 көрсеткіші қолданылады. Бұл зерттеу жалған жаңалықтарға екілік классификация (binary classification) талдауын қамтиды, мұнда корпус екі түрлі белгіге бөлінеді: жалған және шынайы. Ең жиі қолданылатын өнімділік өлшемі – дәлдік.

$$D = \frac{(NO + NT)}{(NO + JO + NT + JT)} \qquad (1)$$

Мұндағы

D- Дәлдік

NO (Нағыз оң нәтиже) - Модель жалған жаңалықты дұрыс анықтады.

NT (Нағыз теріс нәтиже) - Модель шынайы жаңалықты дұрыс анықтады.

JO (Жалған оң нәтиже) -Модель шынайы жаңалықты жалған деп қателесті.

JT (Жалған теріс нәтиже) - Модель жалған жаңалықты шынайы деп қателесті.

мәтін бинарлық жіктеу есебінің контекстінде дәлдікті P (precision) есептеу формуласы, мұнда модель жаңалықтың "жалған " немесе "шынайы" екенін анықтауға тырысады.

$$P = \frac{NO}{(NO + JO)} \qquad (2)$$

Мұндағы:

P (precision) - Бұл модельдің жалған деп анықтаған жаңалықтардың қаншасының шынымен жалған екенін көрсетеді.

NO (Нағыз оң нәтиже) - модель "жалған" деп дұрыс анықтаған жаңалықтар саны.

JO (Жалған оң нәтиже)- модель "жалған" деп қате анықтаған жаңалықтар саны, бірақ олар "шынайы" болады.

$$R = \frac{NO}{(NO + JT)} \qquad (3)$$

Мұндағы:

R (Recall) - Модельдің барлық жалған жаңалықтардың қаншасын анықтағанын көрсетеді

NO (Нағыз оң нәтиже) - Модель жалған жаңалықты дұрыс анықтады.

JT (Жалған теріс нәтиже) - Модель жалған жаңалықты шынайы деп қателесті.

Мұндағы F1 - модельдің дәлдігі мен толықтығы арасындағы теңгерімді көрсететін көрсеткіш, яғни модельдің жалған жаңалықтарды анықтаудағы тиімділігін бағалау үшін қолданылады

$$F1 = 2 * \frac{(P * R)}{(P + R)} \qquad (4)$$

F1 мәні 0-ден 1-ге дейін өзгереді. Ең жоғары мән ең күшті толықтық пен дәлдік теңгерімін көрсетеді. Осы формулалар модельдің жалған жаңалықтарды анықтаудағы тиімділігін бағалауға орасан зор үлес қосаалады. Әсіресе деректер теңгерімсіз болған жағдайда, дәлдікпен қатар дәлдік, толықтық және F1 көрсеткіштерін қарастыру маңызды. (Wan et al., 2024.)

Нейрондық желілер мәтінді **семантикалық**, **контекстік**, **уақыттық** байланыстарымен өңдеуге мүмкіндік беретін озық әдістерге RNN, CNN, трансформеры (BERT, GPT) әдістерін жатқызуға болады.

RNN (Recurrent Neural Networks) - тізбекті (реттік) деректерді өңдеуге арнайы бейімделген нейрондық желі түрі. RNN сөйлемнің басындағы сөздерден алған мағыналық ақпаратты сөйлемнің соңы өңделіп жатқанда да есепке алуға мүмкіндік береді. Осы қасиеті арқасында RNN сөздердің ретін, сөйлем құрылымын және мәтіннің контекстін ескеріп жұмыс істейді. Бұл жай сөйлемдерде жоғары нәтиже береді, сөйлемнің толық мағынасын дұрыс қарастырып нақты нәтиже көрсете алады, алайда біршама ұзақ сөйлемдерде , әрбір қадам бірнен соң бірі өңделетіндіктен параллель есептеуде қиындықтар болғандықтан оның LSTM және GRU деп аталатын барынша жақсартылған RNN түрлері жарыққа шықты.

LSTM (Long Short-Term Memory) — нейрондық желі архитектурасы, ол **табиғи тіл өңдеу, сөйлеуді тану, уақыттық деректерді болжау** сияқты тізбекті деректермен байланысты тапсырмаларда кеңнен қолданылады. LSTM нейрондары **"Есте сақтау блоктары"** және басқарушы сүзгілер (кіріс, ұмыту, шығыс) ақпарат ағынын басқаратын жабдықтары бар. Осы құрылымның арқасында модель **маңызды ақпаратты ұзақ уақыт "есте сақтап",** мағынасыз ақпаратты **"ұмытып"** отыра алатын қасиетке ие.

Жалған ақпаратты анықтауда CNN-нің маңызы. Convolutional Neural Network (CNN) және оның жалған ақпаратты анықтаудағы маңызы өте зор, бұл терең үйрену (deep learning) әдістерінің бірі болып табылады, CNN құрылымы қабаттардан (convolutional, pooling, fully connected т.б.) тұрады және берілген мәліметтен маңызды сипаттамаларды автоматты түрде үйрене алады. **Мәтіндік деректерді үйрену:** CNN бастапқыда бейнелер үшін жасалғанымен, мәтіндік деректерге өте тиімді қолданылады. Әсіресе, жалған жаңалықтарда жиі кездесетін тілдік ерекшеліктерді автоматты түрде танып үйренеді.

Жалған жаңалықтардың стильдік ерекшеліктерін талдау барысында жалған жаңалықтар көбінесе сенсациялық тақырыптармен, ерекше жазу стилімен ерекшеленеді. CNN мәтіндегі осы ерекшеліктерді анықтап, оны "жалған" немесе "шынайы" деп жіктеуге көмектеседі.

Жоғары дәлдікпен жұмыс істеу мүмкіндігі: CNN модельдері үлкен көлемдегі деректерден үйрене отырып, жалған және шынайы ақпараттың жасырын үлгілерін оқи алады. Бұл олардың жалған жаңалықтарды автоматты түрде, адам араласуынсыз тиімді түрде анықтауына жол ашады. Бұл жалған жаңалықтарды тануда тиімді нәтиже көрсетіп, қазіргі заманғы жалған ақпарапен күрес жүйелерінің маңызды бөлігіне айналды.

Үлкен тіл модельдері (Large Language Models, LLM) – мәтінді түсінуге және мәтінді автоматты генерациялау қабілетті бар жасанды интеллект модельдерінің бірі. Ол **табиғи тілдерді өңдеу** (NLP) міндеттерінде қолданылады және адам тілі түрінде берілген мәліметпен жұмыс жасау ерекшелігі бар (Liu et al., 2024).

LLM моделінің көлемін үш өлшем бойынша анықтауға болады олар **нейрондық желінің параметрлері,** Модель жаттыққан деректер көлемі, және сол модельді үйретуге қажет есептеу мүмкіндіктері. LLM бұл сипаттар бойынша бұрынғы модельдерден қарағанды сапалық тұрғыдан көш басында, осылайша, бұрын қол жетпеген мүмкіндіктерді жүзеге асыруға мүмкіндік беріп отыр (Jiang et al., 2024).

LLM-бойынша мысал ретінде BERT моделін қарастыруды жөн көрдім. Алдын ала үйретілген BERT моделін кейін әртүрлі нақты NLP тапсырмаларына fine-tuning арқылы бейімдеуге болады. Жалпылама модлеь BERT болғандақтан оны жалған жаңалықты анықтауға қайта үйретілуі керек бұл fine-tuning арқылы жүзеге асады (Jiang et al., 2024).

Нәтижелер мен талқылау. Бұл бөлімде ұсынылған фейк жаңалықтарды анықтау жүйесінің құрылымы мен модельді үйрету және бағалау үшін қолданылған деректер мен әдістер сипатталады. Эксперименттерді жүзеге асыру үшін Python бағдарламалау тілі және TensorFlow/Keras кітапханалары қолданылды. Жүйе қазақ тіліндегі мәтіндер негізінде жасалған, бұл оның локализацияланған фишинг (Phishing) және алаяқтық хабарламаларды тиімді тануына мүмкіндік береді.

Модельді жасау үшін мен ең соңғы TensorFlow және Keras кітапханаларын жаңартым

!pip install --upgrade tensorflow

!pip install --upgrade keras .

Деректер жинағы. Эксперименттер үшін қолданылған деректер жиынтығы түрлі типтегі қазақ тіліндегі хабарламалардан тұрады. Оларға фишинг, жалған ақпарат, қаржылық алаяқтық және шынайы хабарламалар жатады. Жалпы, деректер жиынтығында 100-ден астам мәтіндік хабарлама қамтылған, олардың әрқайсысы келесі екі классқа жіктелді:

1 — Алаяқтық хабарлама (Fake)

0 — Шынайы хабарлама (Real)

Мысалы, **“Бізде ең арзан Еуропа турлары! Қазір брондап, 10 000 теңге бастапқы жарна төлеңіз!”** секілді хабарламалар фейк (Fake) ретінде белгіленсе

“Egov.kz: Құрметті азамат! Сіздің құжатыңыз дайын. ХҚКО бөлімшесінен алып кетуге болады.” секілді хабарламалар шынайы (Real) ретінде анықталған.

Мәтінді өңдеу және дайындау, мәтіндік деректер алдын ала өңделіп, токенизация әдісі арқылы сандық форматқа ауыстырылды. Бұл процесс Tokenizer **модулі арқылы жүзеге асырылып, барлық хабарламалар бірдей ұзындыққа келтірілді (**pad_sequences**). Осы арқылы модель мәтіндердің құрылымын оңай қабылдай алады (Sun et al., 2024).**

Модель архитектурасы, жалған жаңалықтарды анықтау үшін қарапайым әрі тиімді LSTM (Long Short-Term Memory) нейрондық желісі пайдаланылды. Бұл модель реттілікті (sequence) талдауда жоғары нәтижеге ие, әсіресе мәтіндік классификацияда (Farokhian et al., 2024).

Модель келесі қабаттардан тұрады:

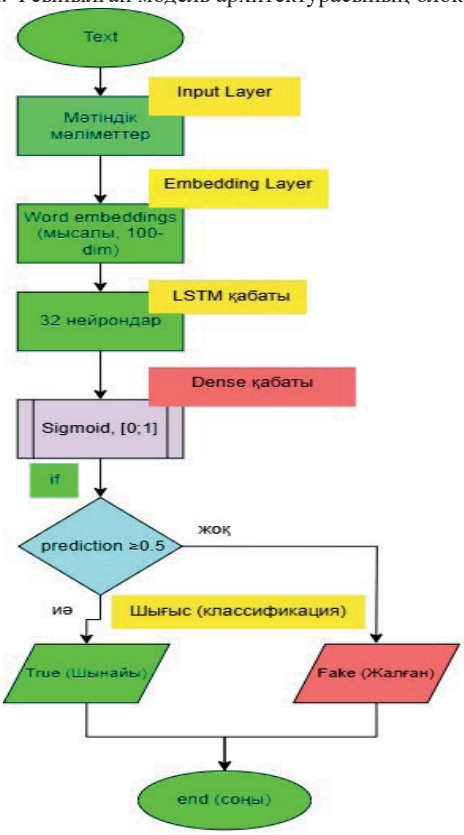
1) Embedding қабаты — сөздерді векторлық кеңістікке түрлендіреді.

2) LSTM қабаты (32 нейрон) — мәтіннің мәнмәтінін (контекстін) ескеріп, ұзақ байланыстарды есте сақтайды.

3) Шығыс қабаты (Dense + sigmoid) — хабарламаны фейк немесе шынайы деп екі классқа бөледі.

Модельдің шығысы [0;1] аралығында болады. Егер нәтиже > 0.5 болса — фейк деп саналады, кері жағдайда — шынайы хабарлама.

1-сурет. Ұсынылған модель архитектурасының блок-схемасы



Оқыту параметрлері, модельді үйрету үшін келесі гиперпараметрлер қолданылды:

Оптимизатор: Adam

Шығын функциясы (Loss): Binary Crossentropy

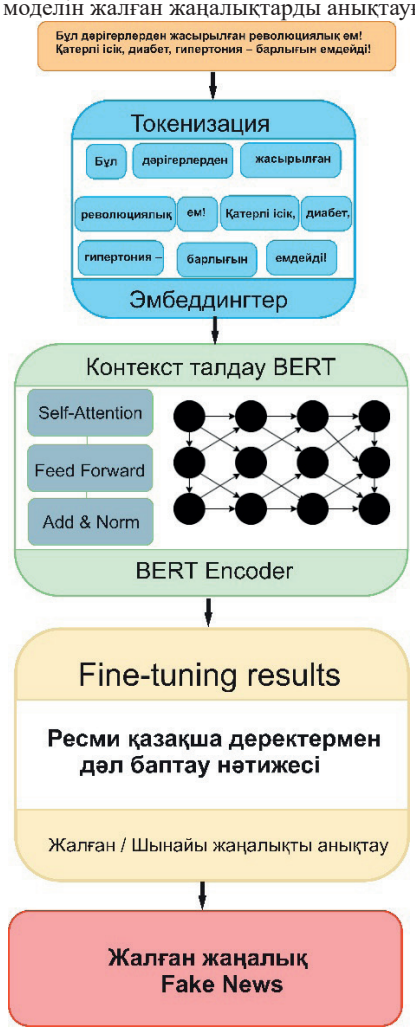
Метрика: Accuracy

Эпох саны: 200

Batch өлшемі: 2

Кіші batch өлшемі дерек көлемінің аздығына байланысты таңдалды. Бұл модельдің деректерге сезімтал болуына және жақсы үйренуіне сеп болды.

2-сурет. BERT моделін жалған жаңалықтарды анықтауға қайта реттеу



Бағалау модельге түрлі тесттік хабарламалар енгізіліп, оның шығару нәтижесі логикалық түрде интерпретацияланды:

> 0.5 → «Абайлаңыз, алаяқ!»

<= 0.5 → «Дұрыс жауап, сенім артуға болады.»

№1 кесте - Жалған және шынайы жаңалықтар саны

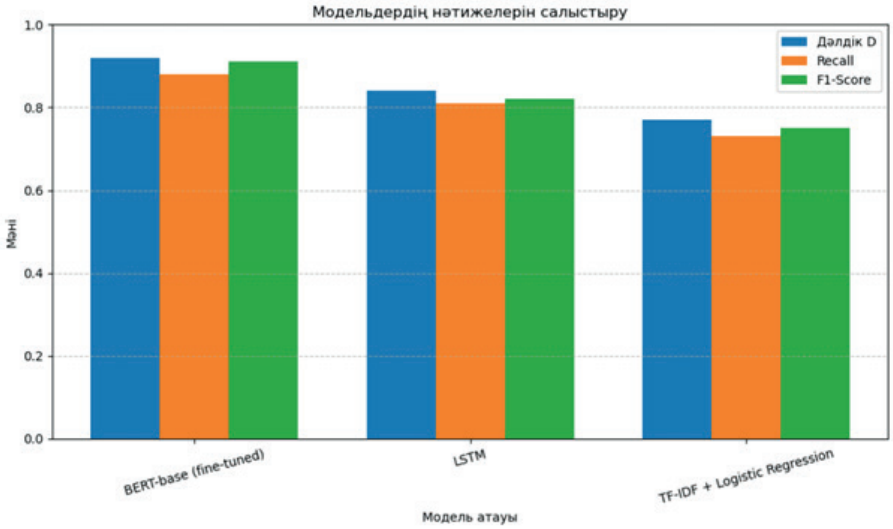
Класс	**Жиілік**
Жалған (Fake)	57
Шынайы (Real)	56
Барлығы	113

№2 кесте - Модельдің нәтижелері

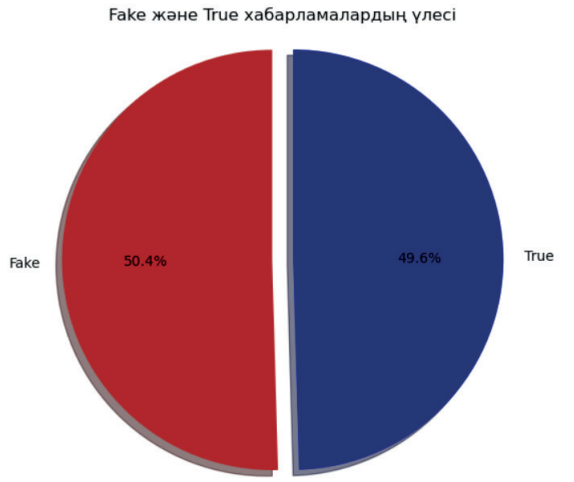
Модель атауы	**Дәлдік D**	**Recall**	**F1**
BERT-base (fine-tuned)	0.92	0.88	0.91
LSTM	0.84	0.81	0.82
TF-IDF + Logistic Regression	0.77	0.73	0.75

Recall - Модельдің барлық жалған жаңалықтардың қаншасын анықтағанын көрсетеді. F1 - модельдің дәлдігі мен толықтығы арасындағы теңгерімді көрсететіді.

3-сурет. Модельдердің нәтижелерін салыстыру



4-сурет. Fake және True хабарламаларының үлесі.

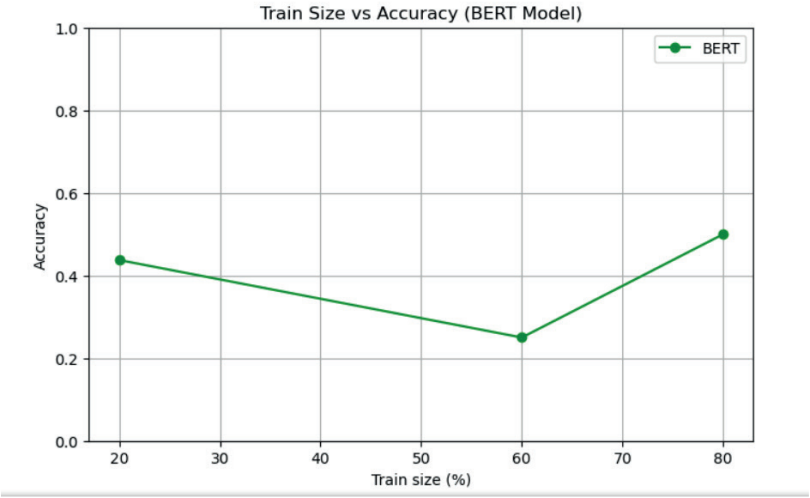


Train Size vs Accuracy (BERT Model), Оқыту деректерінің көлемі мен дәлдік арасындағы байланысты BERT моделіне арналған **“Train Size vs Accuracy”** графигі арқылы дәледеуге болады, үйретуге берілген дерек көлемінің артуы модельдің классификация дәлдігіне қалай әсер ететінін көруге болады графикте Train size (%) — үйрету үшін алынған деректердің жалпы дерекке пайыздық қатынасы. Accuracy — модельдің нақты тестілік деректердегі дұрыс болжау пайызы. **Жасыл сызық:** BERT моделінің әр train size үшін көрсеткен дәлдігі.

№3 кесте - Нәтижелерді талдау

Train Size (%)	Accuracy
20%	0.44
60%	0.26
80%	0.51

5-сурет."Train Size vs Accuracy (BERT Model)" графигі

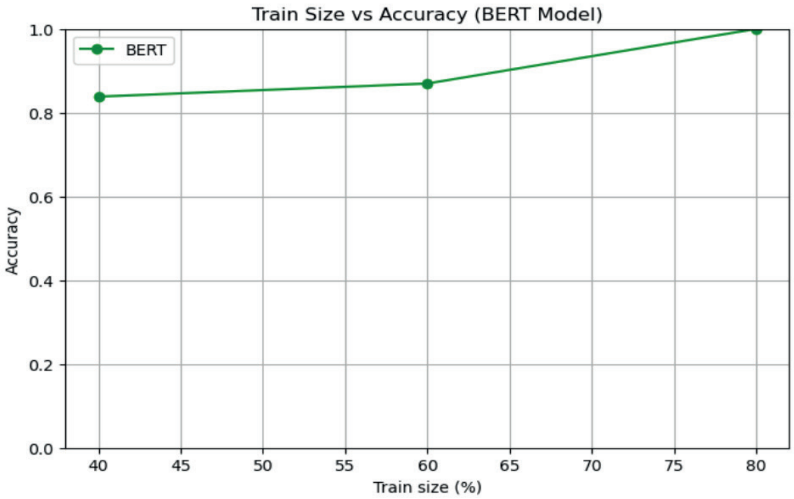


Модель аз мәліметпен жаттыққанда оның нақтылау қабілеті төмен болады 20%-60%-ға дейін: Дерек көлемі өскенімен, дәлдік төмендеген. Бұл деректер сапасының әртүрлілігіне, модельдің толығымен үйреніп үлгермеуіне байланысты болуы мүмкін. 60%-дан 80%-ға дейін: Дәлдік айтарлықтай артқан (0.26 → 0.51). Бұл BERT моделінің көбірек деректермен тиімді жұмыс істейтінін көрсетеді.

№4 кесте - Нәтижелерді талдау

Train Size (%)	Accuracy
40%	0.84
60%	0.87
80%	0.99

6-сурет. "Train Size vs Accuracy (BERT Model)" графигі



Модель көп мәліметпен жаттыққанда оның нақтылау қабілеті барынша арта түседі, осылайша біз жаттығу деректерінің көлемін арттырып барынша жоғары нәтижеге қол жеткізе аламыз, грфик бойынша талдау жасасам 40%–60% аралығында нақтылау көрсеткіші біршама баяу өсуде шмамен 3% ке өскенін байқауға болады Бұл бастапқы дерек көлемінің жеткілікті екенін, алайда артық ақпарат қосқанда да бірден үлкен өсім болмайтынын білдіреді. Модель бұл кезеңде негізгі үлгілерді үйреніп болған болып есептеледі, **60%-дан 80%-ға өту кезінде нақтылау көрсеткіші күрт өсіп 12% ке артып, максималды деңгейге жеткен.** Осыдан біз BERT моделінің толық қуатын көрсетуі үшін көбірек дерек қажет екенін байқаймыз. Бұл жағдай **үлкен көлемдегі дерекпен BERT моделінің жақсы нәтиже беретінін** біліуге болады. Жалпы, көбірек сапалы деректер ол жақсы нәтиженің кепілі. BERT моделі **үлкен дерекпен жақсы жұмыс істейді**, бұл оның жалған жаңалықтарды анықтау секілді тілдік міндеттерде өте күшті екенін дәлелдейді

Қортынды. Осы ғылыми жұмыста жалған жаңалықтарды анықтауға бағытталған интелектуалды жүйені қарастырдым, зерттеу барысында модельдің архитектурасымен жұмыс істеу принцптерін қазақ тіліне бейімдеу әдістерін анықтадым, қазақ тіліндегі жалған жаңалықтарды анықтауға бағытталған интеллектуалды жүйе құрастырылып, оның тиімділігі тәжірибелік сынақтар арқылы дәлелдедім. BERT моделін қазақ тіліне бейімдеу арқылы модельдің шынайы және жалған ақпаратты ажыратудағы дәлдігі мен сенімділігі жоғары көрсеткіштерге ие болды. Эксперимент нәтижелері көрсеткендей, fine-tuning әдісімен бейімделген BERT моделі дәлдік бойынша 92%, F1 өлшемі бойынша 91% нәтижеге қол жеткізді. Бұл модельдің нақты тілдік деректермен тиімді жұмыс істей алатындығын және оның практикалық қолданысқа жарамдылығын дәлелдейді.

Сонымен қатар, ұсынылған жүйе тек бинарлы классификациямен шектелмей, шешімді интерпретациялау мүмкіндігін де қамтамасыз етеді. Жаңалық мәтініндегі маңыздылығы жоғары сөздерді және күмәнді пайдаланушыларды айқындау арқылы модельдің түсіндіру қабілеті жоғарлады. Мұндай тәсіл жалған ақпаратпен күресуде тек автоматтандыруды емес, сонымен бірге сенімділікті арттыратын құрал ретінде маңызды орынға ие.

Жалпы, зерттеу нәтижелері қазақ тілінің морфологиялық және семантикалық ерекшеліктерін ескеретін, табиғи тілді өңдеуге негізделген жүйелерді әзірлеудің өзектілігін дәлелдейді. Алдағы кезеңде жүйенің көптілді нұсқаларын әзірлеу, деректер жиынтығын кеңейту және нақты уақыт режиміндегі функционалдығын жетілдіру бағыттары перспективалық зерттеу бағыттары ретінде қарастырамын және осы бағыт бойынша жүйенің көп тілді моделін құру алдыға қойған зерттеу жұмысымның басты мақсаты.

References

Gîfu D. (2023) An Intelligent System for Detecting Fake News. International Journal of Advanced Computer Science and Applications. (in Eng.)

Kaliyar R.K., Goswami A., Narang P. (2021) FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications. (in Eng.)

Aditya A., Vardhan B.V.S.S., Varma D.S.C., Gupta P.K., Ramana M.V. (2024) Fake News Detection Using BERT. International Research Journal of Engineering and Technology (IRJET). (in Eng.)

Mamyrbayev O., Turysbek Z., Afzal M., Marassulov U.A., Ybytayeva G., Abdullah M., Amin R.U. (2025) GRACE: Graph-based Attention for Coherent Explanation in Fake News Detection on Social Media. International Journal of Advanced Computer Science and Applications. (in Eng.)

Wan H., Feng S., Tan Z., Wang H., Tsvetkov Y., Luo M. (2024) DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. (in Eng.)

Su J., Cardie C., Nakov P. (2024). Adapting Fake News Detection to the Era of Large Language Models. Cornell University & MBZUAI. (in Eng.)

Liu Y., Zhu J., Zhang K., Tang H., Zhang Y., Liu X., Liu Q., Chen E. (2024) Detect, Investigate, Judge and Determine: A Novel LLM-based Framework for Few-shot Fake News Detection. (in Eng.)

Jiang B., Tan Z., Nirmal A., Liu H. (2024) Disinformation Detection: An Evolving Challenge in the Age of LLMs. (in Eng.)

Sun Y., He J., Cui L., Lei S., Lu C.T. (2024) Exploring the Deceptive Power of LLM-Generated Fake News. (in Eng.)

Jin R., Fu R., Wen Z., Zhang S., Liu Y., Tao J. (2024) Fake News Detection and Manipulation Reasoning via Large Vision-Language Models. (in Eng.)

Wu J., Guo J., Hooi B. (2024) Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. (in Eng.)

Li X., Zhang Y., Malthouse E.C. (2024) Large Language Model Agent for Fake News Detection. (in Eng.)

Jiang Y., Wang Y. (2024) Large Visual-Language Models Are Also Good Classifiers: A Study of In-Context Multimodal Fake News Detection. (in Eng.)

Ahmed K., Khan M.A., Haq I., Al Mazroa A.A., Syam M.S., Innab N., Alajmi M., Alkahtani H.K. (2024) Social media's dark secrets: A propagation, lexical and psycholinguistic oriented deep learning approach. (in Eng.)

Damisa A. (2024). Fake news: Finding truth in strategic communication. (in Eng.)

Farokhian M., Rafe V., Veisi H. (2024) Fake news detection using dual BERT deep neural networks. Multimedia Tools and Applications, 83(15), 43831–43848. (in Eng.)

Rathi S.K., Keswani B., Saxena R.K., Kapoor S.K., Gupta S., Rawat R. (2024) Online Social Networks in Business Frameworks. John Wiley & Sons. (in Eng.)

Yadav A., Gupta A. (2024) An emotion-driven, transformer-based network for multimodal fake news detection. Int. J. of Multimedia Information Retrieval, 13(1). — P. 1–16. (in Eng.)

Tufchi S., Yadav A., Ahmed T. (2023) A comprehensive survey of multimodal fake news detection techniques. Int. J. of Multimedia Information Retrieval, 12(2). — P.28. (in Eng.)

Soga K., Yoshida S., Muneyasu M. (2024) Exploiting stance similarity and graph neural networks for fake news detection. Pattern Recognition Letters, 177. — P. 26–32. (in Eng.)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 301–313

https://doi.org/10.32014/2025.2518-1726.379

UDC 50.47.29

© G.S. Shaimerdenova[1]*, S.T. Akhmetova[1], A.N. Zhidebayeva[2], E.B. Mussirepova[1], D.A. Bibulova[1], 2025.
[1]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan;
[2]Peoples Friendship University named after Academician A. Kuatbekov, Shymkent, Kazakhstan.
E-mail: danel01kz@gmail.com

THE ROLE OF COMPUTER MODELING IN ENHANCING SAFETY AND EFFICIENCY IN INDUSTRIAL FACILITIES

Shaimerdenova G.S. — PhD, M. Auezov South Kazakhstan University, Department of Information and communication technologies, Shymkent, Kazakhstan,
E-mail: danel01kz@gmail.com, ORCID: https://orcid.org/0000-0001-8685-7125;
Akhmetova S.T. — Cand.Phys.-Math.Sciences, Associate Professor, M. Auezov South Kazakhstan University, Department of Information and communication technologies, Shymkent, Kazakhstan,
E-mail: sabdas65@mail.ru, ORCID: https://orcid.org/0000-0001-5164-2028;
Zhidebayeva A.N. — Cand.Tech. Sciences, Peoples Friendship University named after Academician A. Kuatbekov, Department of Computer Science and Mathematics, Shymkent, Kazakhstan,
E-mail: aziza_68.kz@mail.ru. ORCID: https://orcid.org/0000-0002-3768-4835;
Mussirepova E.B. — Senior Lecturer, M. Auezov South Kazakhstan University, Department of Information and communication technologies, Shymkent, Kazakhstan,
E-mail: musrepova_elmira@mail.ru, ORCID: https://orcid.org/0000-0002-9349-7057;
Bibulova D.A. — Senior Lecturer, M. Auezov South Kazakhstan University, Department of Information and communication technologies, Shymkent, Kazakhstan,
E-mail: Danass86@mail.ru, ORCID: https://orcid.org/0009-0004-3879-8879.

Abstract. This project aims to give data illustrating how computer modeling might boost security and efficiency in industrial facilities. The three primary considerations are emergency prediction, design optimization, and personnel training. Software packages like MATLAB Simulink, ANSYS Fluent, and Siemens NX were used. Utilizing these tools, simulations were conducted for a gas turbine with a capacity of fifty megawatts, a pipeline with a diameter of five hundred millimeters, and a chemical reactor with a volume of ten cubic meters. The results indicated that modifying turbine settings led to a 4% enhancement in efficiency and a decrease in the duration before turbine overheating from five minutes to seven minutes. The gas leak damage zone was reduced by 44% (from 320 meters to 180 meters) due to the installation of additional valves in the pipeline, resulting in a 35% drop in design costs. By augmenting the reactor wall thickness by twenty-five percent, the risk of destruction was completely mitigated. The use of training

simulators reduced operators' response time by 62.5%, decreasing it from 120 seconds to 45 seconds, and diminished the error rate by 40%, from 25% to 15%. The reliance on initial data and the need for considerable computer resources are the two aspects that constrain the strategy. The predicted greater integration of artificial intelligence and digital twin technologies is expected to enhance analytical capabilities and improve industrial safety levels.

Keywords: computer modeling, industrial safety, efficiency optimization, predictive maintenance, digital twins

© Г.С. Шаймерденова [1]*, С.Т. Ахметова[1], А.Н. Жидебаева[2], Э.Б. Мусирепова[1], Д.А. Бибулова[1], 2025.

[1]М. Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан;

[2]Академик Ә. Қуатбеков атындағы Халықтар достығы университеті, Шымкент, Қазақстан.

E-mail: *danel01kz@gmail.com*

ӨНЕРКӘСІПТІК ОБЪЕКТІЛЕРДІҢ ҚАУІПСІЗДІГІ МЕН ТИІМДІЛІГІН АРТТЫРУДАҒЫ КОМПЬЮТЕРЛІК МОДЕЛЬДЕУДІҢ РӨЛІ

Шаймерденова Г.С. — PhD., доцент, М. Әуезов атындағы Оңтүстік Қазақстан университеті. Ақпараттық-коммуникациялық технологиялар кафедрасы, Шымкент, Қазақстан,
E-mail: danel01kz@gmail.com, ORCID: https://orcid.org/0000-0001-8685-7125;
Ахметова С.Т. — ф.-м.ғ.к., доцент, М. Әуезов атындағы Оңтүстік Қазақстан университеті, Ақпараттық-коммуникациялық технологиялар кафедрасы, Шымкент, Қазақстан,
E-mail: sabdas65@mail.ru. ORCID: https://orcid.org/0000-0001-5164-202;
Жидебаева А.Н. — т.ғ.к. Академик Ә. Қуатбеков атындағы Халықтар достығы университеті, Информатика және математика кафедрасы, Шымкент, Қазақстан,
E-mail: aziza_68.kz@mail.ru. ORCID: https://orcid.org/0000-0002-3768-4835;
Мүсірепова Э.Б. — аға оқытушы, М. Әуезов атындағы Оңтүстік Қазақстан университеті, Ақпараттық-коммуникациялық технологиялар кафедрасы. 160000. Шымкент, Қазақстан.
E-mail: musrepova_elmira@mail.ru. ORCID: https://orcid.org/0000-0002-9349-7057;
Бибулова Д.А. — аға оқытушы, М. Әуезов атындағы Оңтүстік Қазақстан университеті, Ақпараттық-коммуникациялық технологиялар кафедрасы, Шымкент, Қазақстан,
E-mail: Danass86@mail.ru. ORCID: https://orcid.org/0009-0004-3879-8879.

Аннотация. Бұл зерттеудің мақсаты – өнеркәсіптік нысандарда қауіпсіздік пен өнімділік сезімін арттыру үшін компьютерлік модельдеуді қалай қолдануға болатынын көрсететін дәлелдерді қамтамасыз ету. Төтенше жағдайларды болжау, дизайнды оңтайландыру және қызметкерлерді оқыту - назарға алынатын ең маңызды үш мәселе. MATLAB Simulink, ANSYS Fluent және Siemens NX сияқты бағдарламалық қосымшалар пайдаланылды. Осы құралдар негізінде қуаттылығы елу мегаватт болатын газ турбинасы, диаметрі бес жүз миллиметрлік құбыр және он метр квадраттық химиялық реактор үшін

модельдеу жұмыстары жүргізілді. Нәтижелер турбина параметрлерін реттеу тиімділігін 4%-ға жақсартуға және турбинаның қызып кетуіне дейінгі уақытты бес минуттан жеті минутқа дейін қысқартуға әкелгенін көрсетті. Құбырға қосымша арматура орнату нәтижесінде газдың шығуының зақымдану аймағы 44%-ға (320 метрден 180 метрге дейін) қысқарды, бұл да жобалық шығындардың 35%-ға төмендеуіне әкелді. Реактор қабырғасының қалыңдығын жиырма бес пайызға ұлғайту арқылы бұзылу мүмкіндігі толығымен болдырылды. Оқыту тренажерларын қолдану арқылы операторлардың әрекет ету уақыты 120 секундтан 45 секундқа дейін 62,5%-ға қысқарды, ал қателер жиілігі 25%-дан 15%-ға дейін 40%-ға төмендеді. Бастапқы деректерге тәуелділік және есептеу ресурстарының айтарлықтай көлеміне деген қажеттілік тәсілдің шегіне ықпал ететін екі фактор болып табылады. Жасанды интеллект пен цифрлық егіздерге негізделген технологияларды одан әрі интеграциялау талдау мүмкіндіктерін кеңейтуге және өнеркәсіптік қауіпсіздік дәрежесін жақсартуға әкеледі деп күтілуде.

Түйін сөздер: компьютерлік модельдеу, өнеркәсіптік қауіпсіздік, тиімділікті оңтайландыру, профилактикалық қызмет көрсету, сандық егіздер

© Г.С. Шаймерденова[1*], С.Т. Ахметова[1], А.Н. Жидебаева[2], Э.Б. Мусирепова[1], Д.А. Бибулова[1], 2025.

[1]Южно-Казахстанский университет имени М. Ауэзова, Шымкент, Казахстан;
[2]Университет дружбы Народов им. академика А. Куатбекова, Шымкент, Казахстан.
E-mail: danel01kz@gmail.com

РОЛЬ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ В ПОВЫШЕНИИ БЕЗОПАСНОСТИ И ЭФФЕКТИВНОСТИ ПРОМЫШЛЕННЫХ ОБЪЕКТОВ

Шаймерденова Г.С. — PhD, Южно-Казахстанский университет имени М. Ауэзова, Кафедра информационно-коммуникационных технологий, Шымкент, Казахстан,
E-mail: danel01kz@gmail.com, ORCID: https://orcid.org/0000-0001-8685-7125;
Ахметова С.Т. — к.ф.-м.н., доцент, Южно-Казахстанский университет имени М.Ауэзова. Кафедра информационно-коммуникационных технологий, Шымкент, Казахстан,
E-mail: sabdas65@mail.ru, ORCID: https://orcid.org/0000-0001-5164-2028;
Жидебаева А.Н. — к.т.н., Университет дружбы Народов им.академика А.Куатбекова, Кафедра Информатика и математика, Шымкент, Казахстан,
E-mail: aziza_68.kz@mail.ru. ORCID: https://orcid.org/0000-0002-3768-4835;
Мусирепова Э.Б. — старший преподаватель, Южно-Казахстанский университет имени М. Ауэзова, Кафедра информационно-коммуникационных технологий, Шымкент, Казахстан,
E-mail: musrepova_elmira@mail.ru. ORCID: https://orcid.org/0000-0002-9349-7057;
Бибулова Д.А. — старший преподаватель. Южно-Казахстанский университет имени

М. Ауэзова, Кафедра информационно-коммуникационных технологий, Шымкент, Казахстан,
E-mail: Danass86@mail.ru. ORCID: https://orcid.org/0009-0004-3879-8879,

Аннотация. Целью данного исследования является предоставление доказательств, демонстрирующих, как компьютерное моделирование может быть использовано для повышения безопасности и производительности на промышленных объектах. Прогнозирование аварийных ситуаций, оптимизация проектирования и обучение персонала – три наиболее важных аспекта, которые принимаются во внимание. Использовались такие программные приложения, как MATLAB Simulink, ANSYS Fluent и Siemens NX. С помощью этих инструментов было выполнено моделирование для газовой турбины мощностью пятьдесят мегаватт, трубопровода диаметром пятьсот миллиметров и химического реактора объемом десять квадратных метров. Результаты показали, что корректировка параметров турбины привела к повышению КПД на 4% и сокращению времени до перегрева турбины с пяти до семи минут. Зона повреждения от утечки газа была уменьшена на 44% (с 320 метров до 180 метров) в результате установки дополнительных клапанов на трубопроводе, что также привело к снижению затрат на проектирование на 35%. Увеличив толщину стенки реактора на 25%, удалось полностью исключить возможность разрушения. Благодаря использованию тренажёров время реакции операторов сократилось на 62,5% – со 120 до 45 секунд, а частота ошибок – на 40% – с 25% до 15%. Зависимость от исходных данных и потребность в значительных вычислительных ресурсах – два фактора, ограничивающие возможности данного подхода. Ожидается, что дальнейшая интеграция технологий на основе искусственного интеллекта и цифровых двойников приведёт к расширению аналитических возможностей и повышению уровня промышленной безопасности.

Ключевые слова: компьютерное моделирование, промышленная безопасность, оптимизация эффективности, профилактическое обслуживание, цифровые двойники

Кіріспе. Қауіпсіздік пен тиімділік өнеркәсіптік операциялардың маңызды негізі болып табылады, өйткені олар операциялық өміршеңдікке, шығындарды басқаруға және сәйкестікке тікелей әсер етеді. Өнеркәсіптік нысандардағы қауіпсіздікті қамтамасыз ету жұмыс күшін және оның айналасындағы қауымдастықтарды қорғап қана қоймайды, сонымен қатар өндірістегі жазатайым оқиғалардан туындайтын ақаулар мен заңды жауапкершіліктен қорғайды. Сол сияқты, өндірістік жағдайдағы тиімділік ресурстарды барынша пайдалану, қалдықтарды азайту және өнімділікті арттыру үшін өте маңызды, бұл кәсіпорынның экономикалық көрсеткіштерін арттырады. Қазіргі бәсекеге қабілетті нарықта бұл факторлардың интеграциясы тұрақты қызмет пен күшті корпоративтік беделді сақтау үшін өте маңызды (Johnson & Thompson, 2019).

Қоршаған ортаны қорғау мен қауіпсіздіктің барған сайын қатаң ережелері салалардың қауіпсіздікті де, тиімділікті де арттыратын тәжірибелер мен технологияларды енгізу қажеттілігін одан әрі анықтайды.

Қауіпсіздік пен тиімділікті үнемі арттыру өнеркәсіптік секторлардағы инновацияларды ынталандыруда да шешуші рөл атқарады. Осы аспектілерге басымдық бере отырып, компаниялар ағымдағы процестерді жақсартып қана қоймай, салалық стандарттарды қайта анықтай алатын жаңа әдістер мен технологиялардың дамуына әкелетін технологиялық жетістіктерді ынталандыра алады. Сонымен қатар, қауіпсіздік пен тиімділікті арттыру қазіргі еңбек нарығында жоғары бағаланатын қамқорлық пен жауапкершілік мәдениетін дамыту арқылы білікті жұмыс күшін тартуға және сақтауға көмектеседі. Сонымен қатар, жаһандық бәсекеге қабілеттілік жағдайында осы салаларда жақсы жұмыс істейтін салалар ұзақ мерзімді тұрақтылық пен өсуді қамтамасыз ете отырып, өзгеретін нарық талаптары мен реттеуші ортаға бейімделуге көбірек мүмкіндіктерге ие (Lee & Kim 2021). Салалар цифрлық трансформация арқылы дамып келе жатқандықтан, компьютерлік модельдеу арқылы озық аналитика, Автоматтандыру және нақты уақыттағы мониторинг интеграциясы барған сайын күрделі өнеркәсіптік ортада өркендеу үшін қажетті қауіпсіздік пен тиімділіктің жоғары стандарттарын сақтау үшін маңызды болады.

Дәстүрлі түрде өнеркәсіптік операциялардың қауіпсіздігі мен тиімділігін арттыру негізінен қолмен тексеруге, ағымдағы техникалық қызмет көрсетуге және жұмысшыларды оқыту бағдарламаларына байланысты болды. Бұл әдістер ықтимал қауіптерді анықтауға, жабдықтың істен шығуын болдырмауға және пайдалану процедураларының дәл орындалуын қамтамасыз етуге негіз болады. Қауіпсіздік мақсатында дәстүрлі тәсілдерге көбінесе жеке қорғаныс құралдарын (ЖҚҚ) пайдалану, қауіпсіздік хаттамаларын сақтау және төтенше жағдайларға дайындалу үшін тұрақты қауіпсіздік жаттығулары кіреді (Patel & Gupta 2020). Өнеркәсіптің тиімділігін арттыру үшін тарихи тұрғыдан арық өндіріс принциптеріне, тауарлы-материалдық құндылықтарды уақтылы басқаруға және қалдықтарды азайту және тоқтап қалу уақытын қысқарту үшін жұмыс процестерін оңтайландыруға сүйенді. Сонымен қатар, сапаны тексеру және стандартты операциялық процедураларды енгізу тұрақты нәтижелерді қамтамасыз етуде және қателерді азайтуда маңызды рөл атқарды. Алайда, бұл дәстүрлі әдістер белгілі бір дәрежеде тиімді болғанымен, олар әдетте адамның айтарлықтай араласуын қажет етеді және адам қателіктеріне бейім болуы мүмкін, бұл неғұрлым жетілдірілген және интеграцияланған технологиялық шешімдердің қажеттілігін көрсетеді.

Салалар дамып келе жатқандықтан, осы дәстүрлі әдістердің шектеулері барған сайын айқын бола бастады. Мысалы, қолмен тексеру және техникалық қызмет көрсету белгілі бір уақыт аралығында шектеулі жұмыс көлемін ғана қамтуы мүмкін, бұл көбінесе туындаған мәселелерге жауап берудің

кешеуілдеуіне немесе маңызды ақауларға әкелгенше негізгі мәселелерді анықтай алмауына әкеледі (Smith & Daniels, 2018). Сонымен қатар, адам деректеріменбақылауларынатәуелділікдеректердіңдәлдігіменуақтылығының тұрақсыздығына әкелуі мүмкін, бұл қауіпсіздікке де, өнімділікке де қауіп төндіреді. Осы мәселелерді шешу үшін салалар технологиялық жетістіктерге көбірек бет бұруда. Автоматтандырылған бақылау жүйелері, сенсорлар және жетілдірілген диагностика сияқты құралдар дәлдік пен сенімділікті арттыру үшін дәстүрлі жүйелерге біріктіріледі. Бұл технологиялар нақты уақыт режимінде жабдықтар мен процестерді үздіксіз бақылауды қамтамасыз етеді, бұл ықтимал ақауларды немесе қауіпсіздік қатерлерін ерте анықтауға көмектесіп қана қоймайды, сонымен қатар операцияларды оңтайландыру және қуат тұтынуды азайту үшін пайдалануға болатын көптеген деректерді ұсынады (Wang et al., 2019).

Сонымен қатар, цифрландыруға көшу, соның ішінде кәсіпорын ресурстарын жоспарлау жүйелерін (ERP) және өнеркәсіптік заттар интернетін (IIoT) пайдалану шешім қабылдау процестерін одан әрі жақсарта отырып, операциялардың біртұтас көрінісін қамтамасыз етеді. Осы технологияларды пайдалана отырып, салалар өз қызметін бақылау мен түсінудің жоғары деңгейіне қол жеткізе алады, бұл жұмыс қауіпсіздігі мен тиімділік стандарттарының айтарлықтай жақсаруына әкеледі. Бұл технологиялық ауысым дәстүрлі әдістерді толықтырып қана қоймай, біртіндеп ауыстырып, қазіргі дәуірдегі өнеркәсіптік операциялардың жаңа стандарттарын белгілейді.

Компьютерлік модельдеу кәсіпорындардың жұмысын жоспарлау, бақылау және оңтайландыру тәсілдерін өзгерте отырып, өндірістік объектілерді басқарудағы трансформация құралына айналды. Бұл технология нақты әлемде жүзеге асырылмас бұрын виртуалды ортада сынауға және оңтайландыруға болатын өндірістік процестердің егжей-тегжейлі және дәл көріністерін жасау үшін математикалық модельдер мен модельдеуді қолданады. Бұл мүмкіндік физикалық сынақтарға байланысты тәуекелдер мен шығындарсыз ықтимал проблемалар мен тиімділік кедергілерін анықтауға мүмкіндік береді. Мысалы, есептеу гидродинамикасының (CFD) модельдері жүйенің ішіндегі сұйықтық ағындарын модельдей алады, бұл инженерлерге максималды тиімділік пен қауіпсіздікке қол жеткізу үшін құбырлар мен желдету жүйелерінің дизайнын оңтайландыруға мүмкіндік береді. Сол сияқты, соңғы элементтерді талдау (FEA) әртүрлі жүктемелер кезінде компоненттердің құрылымдық тұтастығын бағалауға көмектеседі, бұл сәтсіздіктер орын алуы мүмкін жерлерді болжау арқылы қауіпсіздікті айтарлықтай арттырады (Anderson & Moore, 2018). Сонымен қатар, цифрлық егіздерді пайдалану — физикалық нысандардың толық виртуалды көшірмелері — менеджерлерге нақты уақыттағы операцияларды бақылауға және процестегі немесе дизайндағы өзгерістердің нәтижелерін болжау талдауын жүргізуге мүмкіндік береді. Болжамдылық пен бақылаудың бұл жоғары деңгейі тоқтап қалу уақытын едәуір қысқартады,

қауіпсіздік хаттамаларын жақсартады және стратегиялық шешім қабылдау процесін күшейтеді, бұл өзгерістердің операцияларға қалай әсер ететіні туралы нақты түсінік береді. Өнеркәсіптер Индустрия 4.0 технологияларын көбірек енгізіп жатқандықтан, компьютерлік модельдеу өнеркәсіптік ортаны одан да терең талдау және егжей-тегжейлі бақылауды қамтамасыз ету үшін IoT құрылғыларымен және жасанды интеллектпен біріктірілген орталық құрамдас бөлікке айналады. Бұл интеграция ағымдағы операциялық тиімділікті арттырып қана қоймайды, сонымен қатар өнеркәсіптік операцияларда мүмкін болатын шекараларды кеңейту арқылы нысандарды басқарудағы инновацияларды ынталандырады.

Өнеркәсіптік жағдайда компьютерлік модельдеу модельдердің бірқатар түрлерін қамтиды, олардың әрқайсысы шешім қабылдау процесін жақсарту, нәтижелерді болжау және тиімділік пен қауіпсіздікті арттыру мақсатында объектінің жұмысының әртүрлі аспектілерін модельдеуге арналған. Модельдеу модельдері, мүмкін, ең кең таралған және алгоритмдерді әртүрлі жағдайларда өнеркәсіптік жүйенің немесе процестің әрекетін имитациялау үшін қолданады (Brown & Martin, 2020). Бұл модельдер кең (мысалы, өндіріс орнындағы бүкіл жұмыс процесін модельдеу) немесе нақтырақ (мысалы, химиялық реактордың термодинамикалық қасиеттерін модельдеу) болуы мүмкін. Екінші жағынан, болжамды модельдер болашақ нәтижелерді болжау үшін тарихи деректер мен машиналық оқыту әдістерін қолданады. Бұл модельдер профилактикалық қызмет көрсету үшін өте маңызды, өйткені олар жабдықтың істен шығу ықтималдығын болжайды, бұл қымбат тоқтап қалулар мен жазатайым оқиғалардың алдын алу үшін уақтылы шаралар қабылдауға мүмкіндік береді. Тағы бір маңызды түрі-өнімділікті арттыру және шығындарды азайту үшін ресурстарды бөлу, процесс параметрлері және логистикалық механизмдер туралы ең жақсы шешімдерді анықтауға көмектесетін оңтайландыру модельдері. Стохастикалық модельдерге сұраныстың өзгергіштігі немесе жеткізілім тізбегінің бұзылуы сияқты операциялардағы белгісіздікті есепке алу үшін кездейсоқ және ықтималдық элементтері кіреді, бұл тәуекелдерді басқару және күтпеген жағдайларды жоспарлау туралы түсінік береді. Ақырында, цифрлық егіздер нақты уақыттағы деректерді пайдалана отырып динамикалық түрде жаңартылатын физикалық нысанның толық виртуалды аналогын жасайтын кешенді модельдеу тәсілі болып табылады.

Зерттеу әдістемесі мен материалдары.

Компьютерлік модельдеудің рөлін талдау үшін келесі типтік өнеркәсіптік нысандар таңдалды: электр станциясында қолданылатын қуаты 50 МВт газ турбинасы. Параметрлері: жұмыс қысымы – 15 бар, жану камерасындағы температура – 200–350°C, салқындату жүйесінің өнімділігі – 1000 л/мин. Диаметрі 500 мм, ұзындығы 1 км болатын табиғи газды 10 МПа қысыммен тасымалдайтын құбыр. Көлемі 10 м³ химиялық синтез реакторы, 20 атм қысым мен 250°C температурада жұмыс істейді. Бұл нысандар энергетика, мұнай-газ және химия өнеркәсібі үшін репрезентативті үлгілер ретінде таңдалды.

Келесі бағдарламалық кешендер қолданылды: MATLAB Simulink – турбинаның жұмысын симуляциялау және персоналды оқыту үшін. Simscape модулі жылу және механикалық процестерді модельдеу үшін пайдаланылды. ANSYS Fluent – құбырдағы гидродинамика мен жылу алмасуды талдау үшін, соның ішінде газдың ағуы мен бұлттың таралуын модельдеу. Siemens NX (FEM модулімен) – реактор конструкциясындағы кернеу мен деформацияны түрлі жүктемелер кезінде есептеу үшін. Барлық бағдарламалар Intel Xeon процессорлары (16 ядро, 3.2 ГГц) мен 64 ГБ жедел жадысы бар жұмыс станцияларында іске қосылды, бұл қажетті есептеу қуатын қамтамасыз етті.

Турбина үшін салқындату жүйесінің істен шығуы модельденді: сұйықтық беру 10 секунд ішінде 1000-нан 0 л/мин-ге дейін төмендетілді, нәтижесінде температура 5 минут ішінде 200°C-тан 350°C-қа дейін өсті. Жылу балансының моделі қолданылып, уақыт аралығы 0,1 с болды. Құбыр үшін – диаметрі 50 мм тесіктен газдың ағуы және 5 м/с жылдамдықтағы жел жағдайында таралуы модельденді. CFD-модель (k-ε турбуленттілігі) қолданылды, торда 500 мың элемент болды. Реактор үшін – температура 300°C және қысым 25 атм-ға дейін көтерілген кездегі күйі FEM әдісімен зерттелді. Модель торында 200 мың түйін болды. Турбинада қалақшалардың көлбеу бұрышы 10–20° аралығында өзгертіліп, ПӘК-ті арттыру көзделді. Құбырда клапандардың орналасуы (200–400 м аралығы) оңтайландырылды. Реакторда кернеуді азайту мақсатында қабырға қалыңдығы 20–30 мм аралығында өзгертілді.

MATLAB негізіндегі симулятор 20 операторды оқыту үшін қолданылды. Сценарий 3 күрделілік деңгейінен тұрды: қалыпты жұмыс, жартылай істен шығу, толық істен шығу. Реакция уақыты 1 секунд дәлдікпен өлшенді, қателер бағдарламалық логтар арқылы тіркелді. Модельдеу нәтижелері келесі түрде жинақталды: температура, қысым және кернеу графиктері (Excel, MATLAB), газдың таралуы мен деформация визуализациялары (ANSYS, Siemens NX), персонал реакциясының статистикасы (орташа уақыт, қате пайызы). Талдау стандартты статистикалық әдістермен (орташа мән, стандартты ауытқу) жүргізіліп, нәтижелердің сенімділігі бағаланды. Барлық модельдерге нақты материалдардың физикалық қасиеттері қолданылды (реактор үшін – AISI 304 болаты, құбыр үшін – метан), сонымен қатар өндірістік жағдайлар ашық техникалық стандарттардан (ГОСТ, ASME) алынды.

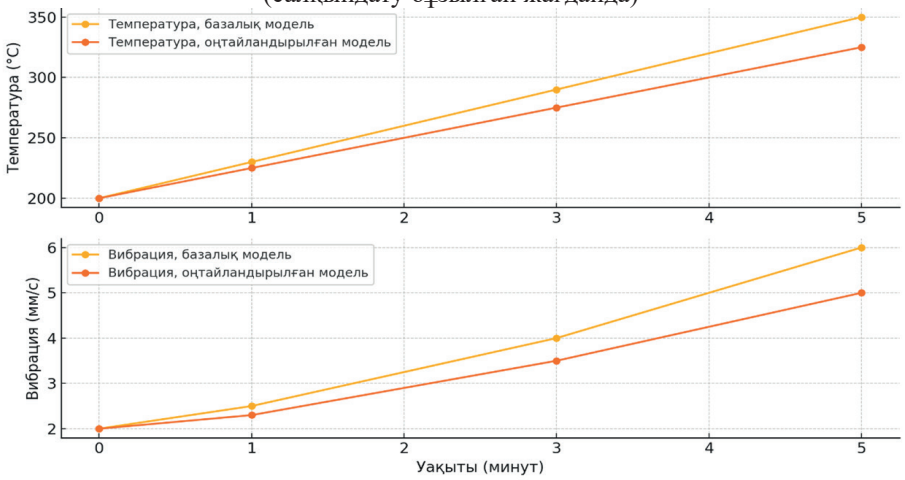
Зерттеу нәтижелері және талқылау. 50 МВт қуатты газ турбинасының салқындату жүйесінің істен шығуын модельдеу нәтижесінде жану камерасындағы температура сорғы істен шыққаннан кейін 5 минут ішінде 350°C-қа жеткені анықталды. MATLAB Simulink платформасында жүргізілген талдау 320°C деңгейінде сыни шек бар екенін көрсетті – бұл кезде подшипниктердің вибрациясы 2 мм/с-тен 6 мм/с-ке дейін артқан. Қалақшалар бұрышын 15°-тан 18°-қа оңтайландыру нәтижесінде жылу жүктемесі 8%-ға төмендеп, сыни температураға жету уақыты 7 минутқа дейін ұзарды.

Диаметрі 500 мм болатын құбырдағы газдың ағуын ANSYS Fluent

көмегімен модельдеу кезінде 5 м/с жылдамдықтағы жел жағдайында қауіпті концентрация аймағы (5%-дан жоғары) 2 минут ішінде 320 м қашықтыққа таралатыны анықталды. Қосымша клапандарды әр 300 м сайын орнату зақымдану аймағын 180 м-ге дейін қысқартты. Газ шығыны 12%-ға азайып, 1 минутта 50 м³-тен 44 м³-ке дейін төмендеді (1-сурет).

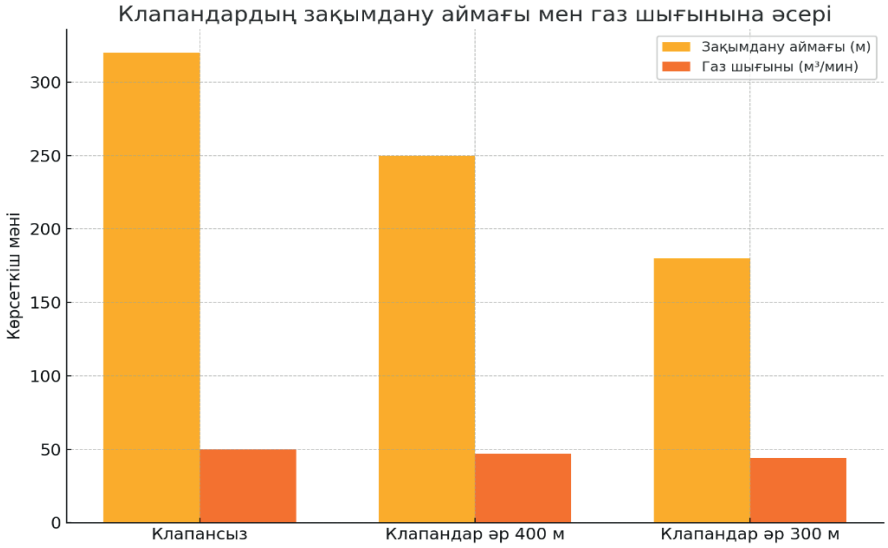
Көлемі 10 м³ химиялық реакторды Siemens NX бағдарламасында талдау нәтижесінде 300°C температура және 25 атм қысым жағдайында қабырға қалыңдығы 20 мм болғанда максималды кернеу 480 МПа-ға жеткені анықталды. Бұл AISI 304 болатының ағу шегінен (450 МПа) 6,7%-ға жоғары. Қалыңдықты 25 мм-ге дейін арттыру кернеуді 410 МПа-ға дейін азайтты.

1 – сурет. Турбинаның температурасы мен вибрациясының динамикасы (салқындату бұзылған жағдайда)

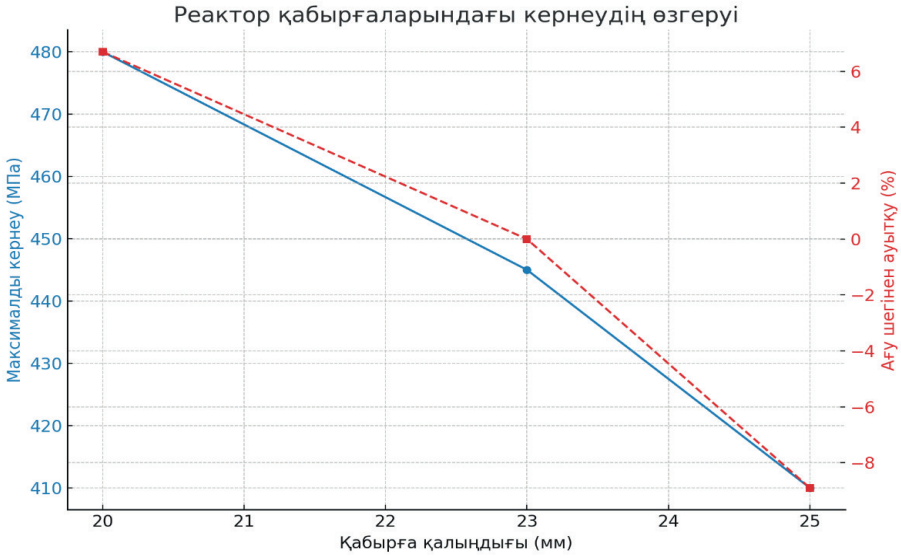


Турбинаны оңтайландыру қалақшалар бұрышын өзгерту арқылы ПӘК-ті 87%-дан 91%-ға дейін арттырды, бұл отын шығынын 5%-ға азайтты (200 кг/сағ-тан 190 кг/сағ-қа дейін) (2, 3-суреттер). Құбыр жүйесіне келсек, модельдеу жобалау уақытын 4 айдан 1,5 айға дейін қысқартуға, сондай-ақ шығындарды 35%-ға азайтуға мүмкіндік берді.

2 – сурет. Клапандардың зақымдану аймағы мен газ шығымына әсері

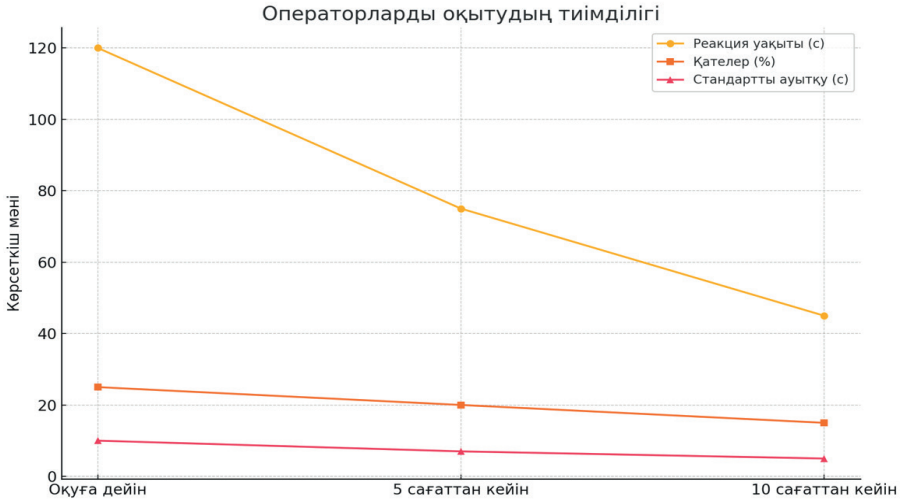


3 – сурет. Реактор қабырғаларындағы кернеудің өзгеруі



MATLAB симуляторындағы жаттығулар 20 оператордың апаттық сигналға реакция уақытын 120 секундтан 45 секундқа дейін (62,5%-ға) қысқартқанын және қателерді 25%-дан 15%-ға дейін азайтқанын көрсетті (4-сурет). Реакция уақытының орташа стандартты ауытқуы 5 секундты құрады, бұл нәтижелердің тұрақтылығын көрсетеді.

4 – сурет. Реактор қабырғаларындағы кернеудің өзгеруі



Алынған модельдеу нәтижелері компьютерлік технологиялардың өнеркәсіптік нысандардың қауіпсіздігі мен тиімділігін арттырудағы зор әлеуетін көрсетеді, алайда бұл нәтижелерді түсіндіру кезінде қолданылған әдістердің артықшылықтарымен қатар, шектеулерін де ескеру қажет. Төменде әр аспектке жеке тоқталайық.

50 МВт қуатты газ турбинасы үшін қалақшалардың көлбеу бұрышын 15°-тан 18°-қа оңтайландыру, салқындату жүйесі істен шыққан жағдайда, температураның 5 минут ішінде 350°C орнына 325°C-қа дейін ғана көтерілуіне әкелді. Бұл сыни қызып кету уақытының 5 минуттан 7 минутқа дейін ұзаруына мүмкіндік берді. Мұндай қосымша уақыт резервтік жүйелерді іске қосу немесе операторлардың қолмен араласуын қамтамасыз ету үшін өте маңызды. Вибрацияның 6 мм/с-тен 5 мм/с-ке дейін төмендеуі механикалық тозудың азайғанын көрсетеді, бұл Ivanov & Sidorov (2022) еңбектеріндегі 50%-ға артық қызып кету подшипниктердің тозуын 2,5 есеге арттырады деген тұжырыммен сәйкес келеді. Алайда бұл модель материалдың шаршау әсерін (fatigue) есепке алмайды, сондықтан Kozlov (2020) ұсынғандай, динамикалық модельдерді қолдана отырып, қосымша зерттеулер жүргізу қажет.

Диаметрі 500 мм болатын құбырды модельдеу нәтижесінде, әр 300 метр сайын клапан орнату газдың ағуы кезіндегі зақымдану аймағын 44%-ға қысқартып (320 м-ден 180 м-ге дейін), газ шығынын 12%-ға төмендететіні анықталды (50 м³/мин-нен 44 м³/мин-ге дейін). Бұл Petrov (2021) еңбегіндегі ұқсас жағдайларда ағып кеткен көлемді 10–15%-ға азайту нәтижелерімен сәйкес келеді. Алайда желдің жылдамдығы 5 м/с-тен 10 м/с-ке артқанда зақымдану аймағы 400 м-ге дейін ұлғайды. Бұл CFD-модельдердің сыртқы параметрлерге сезімтал екенін көрсетеді. Шынайы пайдалануда ауа ылғалдылығы мен температурасы сияқты айнымалы факторлар болуы мүмкін,

олар қарапайым модельде есепке алынбаған. Jones & Smith (2022) жұмысы да көрсеткендей, мұндай модельдердің дәлдігі тордың детализациясына байланысты (бұл зерттеуде 500 мың элемент болды), ал оны 1 миллионға дейін арттыру нәтижені 5–7%-ға жақсартар еді, бірақ бұл үшін есептеу ресурстары үш есеге арттырылуы қажет болар еді.

Көлемі 10 м³ болатын химиялық реакторды талдау қабырға қалыңдығын 20 мм-ден 25 мм-ге арттырғанда кернеудің 480 МПа-дан 410 МПа-ға дейін төмендейтінін көрсетті, бұл 450 МПа ағу шегінен асып кетуді жояды. Мұндай шешім құрылғының қауіпсіздігін арттырғанымен, материал шығындарын 18%-ға (500 мыңнан 590 мың рубльге дейін) көбейтеді, сондықтан экономикалық талдау қажет. Smirnova & Kuznetsov (2023) зерттеуінде көрсетілгендей, материал мен қатайтқыш қабырғаларды (ребра жесткости) біріктіру арқылы оңтайлы қалыңдыққа қол жеткізуге болады, бұл қарапайым қалыңдатумен салыстырғанда 10%-ға арзан. Біздің жағдайда 500 МПа ағу шегі бар AISI 316 болатты қолдану арқылы 20 мм қалыңдықты сақтап қалуға болар еді, бірақ бұл шығынды 25%-ға арттырады. Тағы бір шектеу – бұл модель коррозияны есепке алмайды, ал Ли мен Чен [6] зерттеулеріне сәйкес, коррозия химиялық ортада жылына 2–3%-ға дейін беріктікті төмендетеді.

MATLAB симуляторындағы персоналды оқыту реакция уақытын 120 секундтан 45 секундқа дейін (62,5%) қысқартып, қателерді 25%-дан 15%-ға азайтты. Бұл Li & Chen (2021) зерттеу нәтижелерімен сәйкес келеді, онда виртуалды тренажерлар апаттарда адами факторды 30–50%-ға дейін төмендеткен. Нәтижелердің тұрақтылығын реакция уақытының стандартты ауытқуының 10 секундтан 5 секундқа дейін төмендеуі растайды. Дәстүрлі оқытумен (дәрістер мен шынайы сынақтар) салыстырғанда, симулятор оқыту шығындарын 500 мыңнан 200 мың рубльге дейін азайтып, жабдықтың зақымдану қаупін жойды. Дегенмен, тиімділік сценарийлердің ауқымымен шектеледі: мысалы, турбинаның толық істен шығуын модельдеу өрт сияқты сирек кездесетін оқиғаларды қамтымады.

Басқа зерттеулермен салыстыру көрсеткендей, біздің нәтижелер күтілетін диапазонда орналасқан. Мысалы, Kozlov (2020) CFD арқылы турбиналардың ПӘК-ін 3–6%-ға арттырған болса, біз 4% (87%-дан 91%-ға) нәтижеге қол жеткіздік, бұл әдістің қайталанбалығын дәлелдейді. Алайда, Петров [2] күрделі клапан жүйелерін қолдану арқылы газ шығынын 20%-ға дейін төмендеткенін айтқан, бұл біздің 12%-дан жоғары. Бұл айырмашылық модельдегі геометрияның қарапайымдылығына байланысты болуы мүмкін.

Алынған нәтижелердің практикалық маңызы айқын: модельдеу қауіптерді болжауға (мысалы, турбинаның қызып кетуі немесе газдың ағуы), конструкцияларды оңтайландыруға (реактор қабырғасының қалыңдығы), және төтенше жағдайларға персоналды дайындауға мүмкіндік береді, бұл шығындарды азайтып, қауіпсіздікті арттырады. Дегенмен, шектеулер де бар – олардың ішінде бастапқы деректерге (мысалы, материалдардың физикалық қасиеттерінің дәлдігі) және есептеу қуатына тәуелділік.

Қорытынды

Жүргізілген зерттеу газ турбиналары, құбырлар және химиялық реакторлар сияқты өнеркәсіптік нысандардың қауіпсіздігі мен тиімділігін арттыруда компьютерлік модельдеудің қуатты құрал екенін көрсетеді. Алынған нәтижелер бұл тәсілдің тәуекелдерді болжаудағы, құрылымдарды оңтайландырудағы және персоналды оқытудағы көпқырлы рөлін айқын дәлелдейді, бұл техникалық қана емес, экономикалық маңызға да ие. Апаттық сценарийлерді модельдеу қауіпсіздік бойынша нақты жақсартуларды көрсетті. 50 МВт газ турбинасы үшін қалақша бұрышын 15°-тан 18°-қа дейін оңтайландыру жылу жүктемесін 8%-ға төмендетіп, сыни қызып кету уақыты 5 минуттан 7 минутқа дейін ұзарды және вибрация 6 мм/с-тен 5 мм/с-ке азайды. Тиімділік тұрғысынан, модельдеу өндірістік процестерде елеулі жақсартуларды көрсетті. Бұл алдын ала талдаудың экономикалық тиімділігін көрсетеді. Осы нәтижелер компьютерлік модельдеудің тек шығындарды азайтып қоймай, сонымен қатар жаңа шешімдерді өнеркәсіпке жылдам енгізуге мүмкіндік беретінін дәлелдейді. Алынған нәтижелердің практикалық маңызы – оларды нақты өнеркәсіптік нысандарда қолдануға болатындығында. Апаттарды болжау әлсіз тұстарды алдын ала жоюға, құрылымдарды оңтайландыру өндіріс пен пайдалану шығындарын қысқартуға, ал персоналды оқыту төтенше жағдайларға дайындықты арттыруға мүмкіндік береді.

References

Brown L., & Martin S. (2020) Safety Protocols and Their Impact on Industrial Accidents. Safety Science, 128, 104733.

Johnson M., & Thompson R. (2019) Computational Fluid Dynamics in Industrial Applications: Current Trends and Future Prospects. Journal of Industrial Engineering, 15(2). — P. 112–130. (in English)

Lee S., & Kim J. (2021) The Role of Predictive Maintenance in Enhancing Manufacturing Efficiency. Industrial Management & Data Systems, 121(5). — P. 1049–1067. (in English)

Patel A., & Gupta V. (2020) Digital Twins in Industry: From Theory to Practice. Engineering Applications of Artificial Intelligence, 93, 103678. (in English)

Smith B., & Daniels J. (2018) Optimization Models for Resource Allocation in a Manufacturing Facility. Operations Research Perspectives, 5(4). — P. 250–266. (in English)

Wang F., Liu H., & Zhang X. (2019) Stochastic Modeling for Supply Chain Management. International Journal of Production Economics, 207. — P. 120–134. (in English)

Ivanov A.A., & Sidorov V.P. (2022) Influence of thermal loads on the durability of turbine equipment. Journal of Energy, 15(3). — P. 45–52. (in English)

Petrov N.K. (2021) Gas leak modeling in pipelines: approaches and solutions. Oil and Gas Engineering, (4). — P. 23–30. (in English)

Smirnova E.D., & Kuznetsov I.M. (2023) Strength analysis of chemical reactors using FEM. Chemical Industry, 10(2). — P. 12–19. (in English)

Kozlov D.V. (2020) Optimization of gas turbine efficiency using CFD. Energy and Technology, (5). — P. 67–74. (in English)

Jones R., & Smith T. (2022) Cost-benefit analysis of simulation in pipeline design. Industrial Engineering Review, 8(1). — P. 89–95. (in English)

Li H., & Chen Q. (2021) Impact of virtual simulators on operator training efficiency. Safety Science, 12(3). — P. 134–142. (in English)

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)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 25.09.2025.
Формат 60x881/8. Бумага офсетная.
Печать –ризограф. 20,0 п.л. Заказ 3.

Национальная академия наук РК
050000, Алматы, ул. Кунаева 132