МРНТИ 20.53.19

МРНТИ 81.93.29

**ПРОГРАММНО-АППАРАТНЫЙ КОМПЛЕКС ДАКТИЛОСКОПИЧЕСКОЙ ИДЕНТИФИКАЦИИ ЧЕЛОВЕКА**

**1,2 Т.С. Шорманов, 2А.Т. Мазакова, 1 М.С. Алиаскар, 2А.Д. Бургегулов, 2А.А. Саметова,**

**1Н.Т. Исимов, 2Ш.А. Джомартова, 1,2Т.Ж. Мазаков🖂**

1 Международный инженерно-технологический университет, Алматы, Казахстан,

2 Казахский национальный университет имени аль-Фараби, Алматы, Казахстан

**🖂**Корреспондент-автор: tmazakov@mail.ru

Дактилоскопическая идентификация основывается на уникальности кожных рисунков на пальцах рук, которые формируются в процессе эмбрионального развития и не меняются на протяжении жизни человека. Каждый отпечаток пальца имеет свои особенности, включая линии, завитки, поры, выступы и другие микроскопические характеристики.

Современные **программно-аппаратные комплексы** дактилоскопической идентификации человека обеспечивают высокую точность и скорость обработки отпечатков пальцев, что делает их востребованными в различных сферах, таких как правоохранительные органы, система безопасности, а также в биометрических системах доступа.

В работе рассматривается разработка системы дактилоскопической идентификации, объединяющей программное обеспечение и аппаратные компоненты. Предложенная система, основанная на использовании микроконтроллера Arduino и сканера отпечатков пальцев FPM10A, предназначена для выполнения операций по хранению, обработке, идентификации и визуализации данных отпечатков пальцев. Для идентификации личности были выбраны особенности структуры папиллярных линий на пальцах рук.

Экспериментальные исследования подтвердили устойчивость разработанной системы к изменениям масштаба, поворотам изображений и небольшим искажениям.

**Ключевые слова:** биометрические технологии, дактилоскопическая идентификация, программно-аппаратные системы, микроконтроллер.

**АДАМДЫ САУАҚ ІЗІМЕН АНЫҚТАУ ҮШІН БАҒДАРЛАМАЛЫҚ-АППАРАТТЫҚ КЕШЕН**

**1,2 Т.С. Шорманов, 2Ә.Т. Мазақова, 1 М.С. Әлиасқар, 2А.Д. Бургегулов, 2А.А. Саметова,**

**1 Н.Т. Исимов, 2Ш.А. Джомартова, 1,2Т.Ж. Мазақов🖂**

1Халықаралық инженерлік және технология университеті,

2Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

e-mail:tmazakov@mail.ru

Саусақ ізін анықтау эмбриональды даму кезінде қалыптасатын және адамның өмір бойы өзгермейтін саусақтардағы тері үлгілерінің бірегейлігіне негізделген. Әрбір саусақ ізінің өзіндік сипаттамалары бар, соның ішінде сызықтар, бұйралар, тесіктер, жоталар және басқа микроскопиялық сипаттамалар.

Адамның саусақ ізін сәйкестендіруге арналған заманауи бағдарламалық-аппараттық жүйелер саусақ іздерін өңдеудің жоғары дәлдігі мен жылдамдығын қамтамасыз етеді, бұл оларды құқық қорғау органдары, қауіпсіздік жүйелері сияқты әртүрлі салаларда, сондай-ақ биометриялық қол жеткізу жүйелерінде сұранысқа ие етеді.

Жұмыста бағдарламалық және аппараттық құрамдас бөліктерді біріктіретін саусақ ізін сәйкестендіру жүйесінің дамуы талқыланады. Arduino микроконтроллерін және FPM10A саусақ ізі сканерін пайдалануға негізделген ұсынылған жүйе саусақ ізі деректерін сақтау, өңдеу, анықтау және визуализациялау операцияларын орындауға арналған. Жеке тұлғаны анықтау үшін саусақтардағы папиллярлық сызықтардың құрылымының ерекшеліктері таңдалды. Эксперименттік зерттеулер әзірленген жүйенің масштабтағы өзгерістерге, кескіннің айналуына және шамалы бұрмалануына тұрақтылығын растады.

**Түйін сөздер:** биометриялық технологиялар, саусақ іздерін сәйкестендіру, бағдарламалық-аппараттық жүйелер, микроконтроллер.

**HARDWARE AND SOFTWARE COMPLEX FOR HUMAN DACTYLOSCOPIC IDENTIFICATION**

**1,2 T.S. Shormanov, 2A.T. Mazakova, 1M.S. Aliaskar, ²A.D.Burgegulov, 2A.A. Sametova,**

**1N.T. Isimov, 2Sh.A. Jomartova, 1,2T.Zh. Mazakov🖂**

1 International Engineering and Technology University, Almaty, Kazakhstan,

2 Kazakh National University named after Al-Farabi, Almaty, Kazakhstan,

e-mail:tmazakov@mail.ru

Fingerprint identification is based on the uniqueness of skin patterns on the fingers, which are formed during embryonic development and do not change throughout a person's life. Each fingerprint has its own characteristics, including lines, curls, pores, protrusions and other microscopic characteristics.

Modern software and hardware systems for fingerprint identification of a person provide high accuracy and speed of fingerprint processing, which makes them in demand in various areas, such as law enforcement agencies, security systems, as well as in biometric access systems.

The paper considers the development of a fingerprint identification system that combines software and hardware components. The proposed system, based on the Arduino microcontroller and the FPM10A fingerprint scanner, is designed to perform operations on storing, processing, identifying and visualizing fingerprint data. The features of the structure of the papillary lines on the fingers were selected for personal identification. Experimental studies confirmed the stability of the developed system to changes in scale, image rotation and minor distortions.

**Keywords:** biometric technologies, fingerprint identification, hardware and software systems, microcontroller.

**Введение.** Современные методы защиты информации и объектов требуют высокой степени надежности, которая зависит от специфики и уровня безопасности, необходимого в конкретной ситуации. Одним из наиболее эффективных решений являются биометрические системы, в частности технологии идентификации личности по отпечаткам пальцев. Эти системы получили широкое распространение благодаря своей адаптивности, высокой точности и удобству использования. Применение биометрических технологий, таких как сканирование отпечатков пальцев, не только усиливает уровень защиты, но и устраняет необходимость в традиционных средствах доступа, таких как ключи или карты, заменяя их уникальным, неизменным биометрическим признаком [1].

Дактилоскопические системы функционируют на основе сравнения полученных отпечатков с хранимыми в базе данных. Методика сопоставления определяется областью применения технологии. Уникальность отпечатков пальцев обусловлена анатомическими особенностями строения кожного рисунка, который формируется под влиянием генетических и экологических факторов [2].

Идентификационные признаки папиллярных узоров делятся на глобальные и локальные. Глобальные признаки (например, тип узора, направление линий) видимы невооруженным глазом, в то время как локальные признаки, или минуции (раздвоения, разрывы и окончания линий), требуют специального анализа. Методы, основанные на локальных признаках, обладают большей надежностью и детализированностью, что делает их предпочтительными для биометрической идентификации. Однако, такие факторы, как давление при сканировании, влажность кожи и возрастные изменения, могут влиять на качество изображения отпечатков, что требует использования устойчивых к искажениям алгоритмов [3].

**Материалы и методы.** Программно-аппаратные комплексы для идентификации по отпечаткам пальцев состоят из нескольких ключевых элементов, которые взаимодействуют друг с другом для реализации технологии.

*Аппаратные компоненты включают:*

1) Сканеры отпечатков пальцев:

* + Оптические сканеры: используют светодиоды (LED) для подсветки пальца и камеры для захвата изображения. Это один из наиболее популярных типов, однако он может быть подвержен загрязнениям и повреждениям.
  + Полимерные (сенсорные) сканеры: используют электрическое поле для захвата изображения. Они обеспечивают более высокую точность и защищенность от грязи, а также могут быть более компактными.
  + Ультразвуковые сканеры: передают ультразвуковые волны на поверхность пальца, анализируя отраженные сигналы для создания детализированного изображения. Эти сканеры обеспечивают высокую точность, могут работать с влажными и поврежденными пальцами.
  + Термические сканеры: регистрируют тепловое излучение поверхности пальца и могут быть использованы в условиях низкой видимости или при загрязненных сканерах.

1. Процессоры и устройства обработки: Обработка данных с сенсора требует быстрого и мощного оборудования для извлечения, анализа и сравнения отпечатков пальцев в реальном времени. Это может быть либо специализированное аппаратное обеспечение, либо вычислительные мощности встроенных процессоров.

Программные компоненты состоят из:1) Модуля захвата изображений: Программное обеспечение, отвечающее за захват и обработку изображений отпечатков пальцев с сенсора. Этот модуль включает в себя алгоритмы для нормализации изображений, улучшения их качества (например, удаление шумов, повышение контраста) и подготовки их для дальнейшего анализа.

2) Модуля извлечения признаков: На основе изображения отпечатка пальца извлекаются ключевые биометрические признаки, такие как:

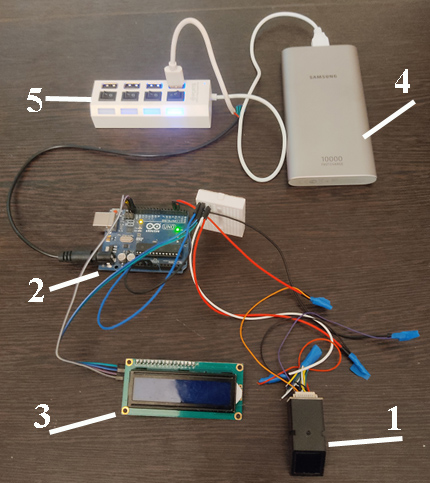
3) Модуля сопоставления: Алгоритмы сравнивают текущий отпечаток с базой данных зарегистрированных отпечатков. В случае совпадения система может подтвердить идентичность, и пользователь получит доступ. Обычно используются методы, основанные на анализе статистических признаков или методах машинного обучения для повышения точности и устойчивости к ошибкам.

4) Модуля идентификации: Этот компонент проверяет, соответствует ли результат сопоставления зарегистрированным данным пользователя и принимает решение о подтверждении или отклонении идентификации [4-5].

**Обсуждение и результаты.** Традиционные методы контроля доступа, такие как использование паролей или ключей, часто неудобны и подвержены риску утраты [6].

В ходе данной работы разработана биометрическая система идентификации, основанная на использовании сканера отпечатков пальцев и микроконтроллера Arduino. Применение библиотеки Adafruit для Arduino позволило интегрировать аппаратные и программные компоненты, обеспечив точное распознавание и идентификацию.

На рисунке 1 представлена фотография разработанного программно-аппаратного комплекса системы: 1) сканер FPM10A, 2) микроконтроллер Arduino, 3) средство визуализации результата распознавания, 4) внешний аккумулятор (Power Bank), 5) мульти-разветвитель USB.

****

**Рис.1- Программно-аппаратный комплекс идентификации личности по отпечаткам пальцев**

Программно-аппаратный комплекс состоит из нескольких ключевых компонентов:

*1) аппаратная часть:*

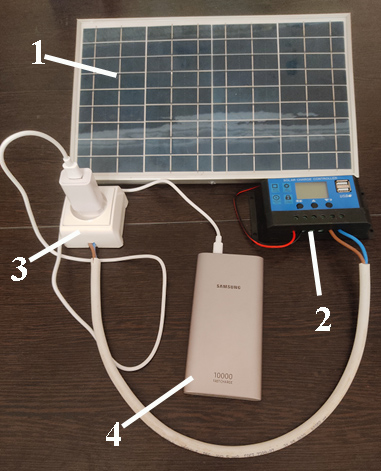
* + оптический сканер FPM10A, способный сохранять до 1000 отпечатков пальцев:
  + микроконтроллер Arduino, обеспечивающий обработку данных в реальном времени.

*2) программная часть:*

* + библиотека Adafruit\_Fingerprint, позволяющая выполнять все этапы обработки отпечатков;
  + модули для извлечения ключевых точек и анализа биометрических данных.

*3) система электрообеспечения:*

* + внешний аккумулятор Power Bank емкостью 10 000 мА, обеспечивающий бесперебойную работу комплекса;
  + мульти-разветвитель USB, позволяющий снизить энергетическую нагрузку на микроконтроллер;
  + солнечная панель с контроллером заряда, обеспечивающие зарядку внешнего аккумулятора.



**Рис.2- Схема электрического обеспечения: 1- солнечная панель, 2) контроллер заряда, 3) розетка для заряда внешнего аккумулятора, 4) внешний аккумулятор**

Сенсор использует алгоритмы, позволяющие хранить, обрабатывать и сопоставлять отпечатки пальцев. Для этого применяется библиотека Adafruit\_Fingerprint, которая включает функции:

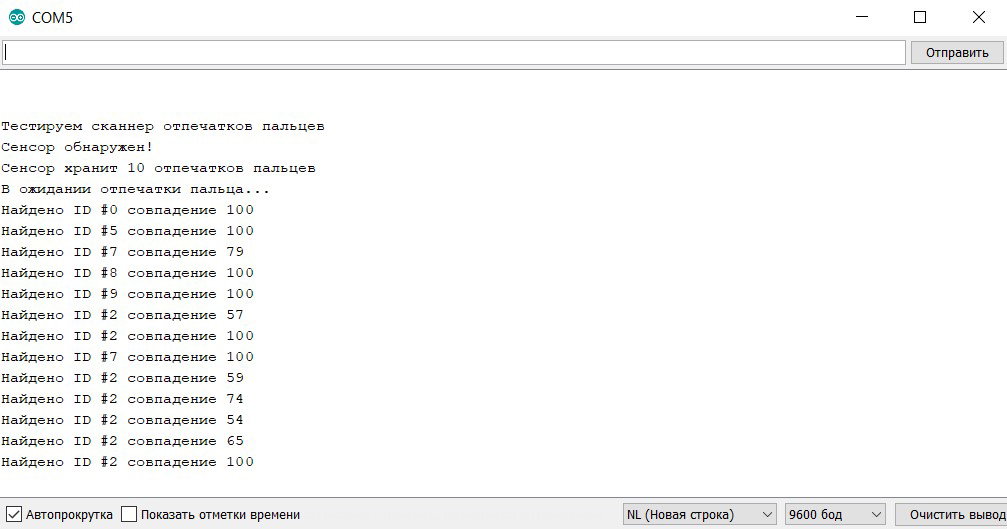
* Захвата изображения.
* Извлечения ключевых признаков (локальных точек).
* Сравнения текущего отпечатка с сохранёнными образцами.

Программные инструменты позволяют записывать новые отпечатки, присваивая каждому уникальный ID (который сохраняется в памяти сенсора для дальнейшего сравнения), и проверять работу системы через окно серийного монитора Arduino.

Запись новых отпечатков через программу для Windows

Для записи новых данных в память оптического датчика отпечатков пальцев рекомендуется использовать специализированную программу для Windows.

На рисунке 3 продемонстрирована работа программно-аппаратного комплекса по идентификации личности по отпечаткам пальцев.



**Рис.3 - Результаты работы программно-аппаратного комплекса**

Внешний аккумулятор Samsung EB-P1100C емкостью 10000 мАч позволяет работать программно-аппаратному комплексу больше 48 часов (двое суток), что подтверждается дальнейшими расчетами.

Расчет времени работы Arduino Uno от внешнего аккумулятора Power Bank на 10000 мАч, необходимо учитывать несколько факторов:

Arduino Uno обычно потребляет от 30 до 50 мА при обычной работе (без дополнительных устройств). При использовании дополнительных модулей, датчиков или экранов потребление может увеличиться, но для базового расчета возьмем среднее значение 50 мА.

Power Bank на 10000 мАч, как правило, имеет выходное напряжение 5 В (при использовании USB-выхода), и его реальная емкость может быть немного меньше из-за потерь при преобразовании напряжения. Обычно можно рассчитывать на эффективность около 85%. Поэтому примем что эффективная емкость Power Bank будет примерно 8500мАч

Время работы рассчитывается по формуле:

Время=Емкость/Потребление=8500 мАч/50 мА=170 часов.

Учитывая, что Arduino подключает дополнительные устройства (датчики, экраны и т.д.) потребление энергии может увеличиться втрое. Поэтому мы берем за гарантийное время - 48 часов.

**Выводы.** В рамках данной работы получены следующие результаты:

Проведен анализ методов биометрической идентификации, включающих использование сканеров отпечатков пальцев и алгоритмов сопоставления.

Разработан программно-аппаратный комплекс на основе микроконтроллера и оптического сканера, обеспечивающий хранение, обработку и идентификацию отпечатков пальцев. Основой идентификации послужила структура папиллярных узоров.

Экспериментальные исследования продемонстрировали, что разработанная система обеспечивает:

1) Устойчивость к изменениям масштаба и поворотам изображения.

2) Инвариантность к незначительным искажениям и изменению уровня освещения до 50–70%.

3) Надежное распознавание даже при использовании частичных отпечатков пальцев.

Преимущества применения дактилоскопической информации:

1) Высокая точность: Отпечатки пальцев имеют уникальные биометрические признаки, которые остаются неизменными на протяжении всей жизни человека, что делает их надежным средством для идентификации.

2) Удобство использования: Сканирование отпечатков пальцев требует всего лишь прикосновения к сенсору, что делает процесс аутентификации быстрым и удобным.

3) Высокий уровень безопасности: Трудно подделать отпечатки пальцев, что делает этот метод биометрической идентификации более безопасным по сравнению с паролями или PIN-кодами.

Особую перспективу представляет разработка алгоритмов для поиска по неполным отпечаткам пальцев, что особенно важно в реальных условиях, когда доступна лишь часть изображения для идентификации.

***Финансирование****. Работа выполнена за счет средств НИИ математики и механики при КазНУ имени аль-Фараби и грантового финансирования научных исследований на 2023–2025 годы по проекту AP19678157*.

**Литература**

1. Болл Р.М., Коннел Дж.Х., Панканти Ш. и др. Руководство по биометрии. – Москва: Техносфера, 2007.-368 с. [ISBN](https://www.libex.ru/qna/ref/isbn/): 978-5-94836-109-3.

URL: https://www.technosphera.ru/lib/book/187

2. Мазур Е.С. Дерматоглифика в исследованиях личности криминалистический и судебно-медицинский аспекты. – Томск: Изд. дом ТГУ, 2014, - 150с. [ISBN](https://www.libex.ru/qna/ref/isbn/): 978-5-94621-450-6. –

URL: https://ibooks.ru/bookshelf/380802/reading

3. Лепихова Д.Н., Гудков В.Ю., Кирсанова А.А. Обзор современных моделей представления дактилоскопических изображений // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика.- 2018.-Т.7(1)- с.40-59.

DOI 10.14529/cmse180104.

4. Гридчин А.В. Микродатчики и микросистемы. – Москва, Вологда: Инфра-Инженерия, 2023. -184 с. [ISBN](https://www.libex.ru/qna/ref/isbn/) 978-5-9729-1220-9.

URL: https://www.iprbookshop.ru/133049.html

5. Апейников А.Ф., Гридчин В.А., Цапенко М.П. Датчики (перспективные направления развития). – Новосибирск: Изд-во НГТУ, 2001.-176.

[ISBN](https://www.libex.ru/qna/ref/isbn/) 5-7782-0300-4

6. Бриллиант К. Цифровая модель человека. – М.: Кудиц-образ, 2004.- 400 с. ISBN 5-7782-0300-4

**References**

1. Boll R.M., Konnel Dzh.H., Pankanti Sh. i dr. Rukovodstvo po biometrii. – Moskva: Tehnosfera, 2007.-368 s. ISBN: 978-5-94836-109-3.

URL: <https://www.technosphera.ru/lib/book/187>. [in Russian]

2. Mazur E.S. Dermatoglifika v issledovanijah lichnosti kriminalisticheskij i sudebno-medicinskij aspekty. - Tomsk: Izd. dom TGU, 2014, - 150s. ISBN: 978-5-94621-450-6.

URL: <https://ibooks.ru/bookshelf/380802/reading>/ [in Russian]

3. Lepihova D.N., Gudkov V.Ju., Kirsanova A.A. Obzor sovremennyh modelej predstavlenija daktiloskopicheskih izobrazhenij // Vestnik JuUrGU. Serija: Vychislitel'naja matematika i informatika.- 2018.-T.7(1)- s.40-59. DOI 10.14529/cmse180104. [in Russian]

4. Gridchin A.V. Mikrodatchiki i mikrosistemy. – Moskva, Vologda: Infra-Inzhenerija, 2023. -184 s. ISBN 978-5-9729-1220-9. URL: https://www.iprbookshop.ru/133049.html.

5. Apejnikov A.F., Gridchin V.A., Capenko M.P. Datchiki (perspektivnye napravlenija razvitija). -Novosibirsk: Izd-vo NGTU, 2001.-176. ISBN 5-7782-0300-4. [in Russian]

6. Brilliant K. Cifrovaja model' cheloveka. – M.: Kudic-obraz, 2004.- 400 s. ISBN 5-7782-0300-4. [in Russian]

***Сведение об авторах***

Т.С. Шорманов– докторант КазНУ имени аль-Фараби, старший преподаватель МИТУ, Алматы, Казахстан,

e-mail[shormanov@gmail.com](mailto:shormanov@gmail.com);

А.Т. Мазақова – докторант КазНУ им.аль-Фараби, e-mail: [aigerym97@mail.ru](mailto:aigerym97@mail.ru);

М.С. Әлиасқар - старший преподаватель МИТУ, Алматы, Казахстан, e-mail: [m.alyasqar@gmail.ru](mailto:m.alyasqar@gmail.ru);

А.Д. Бургегулов – докторант КазНУ им.аль-Фараби, e-mail: [dizel\_kz@bk.ru](mailto:dizel_kz@bk.ru);

Саметова А.А. – докторант КазНУ им.аль-Фараби, e-mail: [sametova\_aygerim@mail.ru](mailto:sametova_aygerim@mail.ru);

Н.Т. Исимов - заведующий кафедрой МИТУ, Алматы, Казахстан, e-mail: [int\_nurdaulet@mail.ru](mailto:int_nurdaulet@mail.ru);

Ш.А. Джомартова - доктор технических наук, доцент, КазНУ им. аль-Фараби, Алматы, Казахстан,

e-mail: jomartova@mail.ru;

Т.Ж. Мазаков – доктор физико-математических наук, профессор, КазНУ им. аль-Фараби, Алматы, Казахстан, e-mail: tmazakov@mail.ru.

***Information about the authors***

Shormanov T.S. - PhD student of the Al-Farabi Kazakh National University, Lecturer at the International University of Engineering and Technology, Almaty, Kazakhstan, e-mail: [shormanov@gmail.com](mailto:shormanov@gmail.com);

Mazakova A.T. - PhD student of the Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [aigerym97@mail.ru](mailto:aigerym97@mail.ru);

Aliaskar M.S. - Lecturer at the International University of Engineering and Technology, Almaty, Kazakhstan, e-mail: [m.alyasqar@gmail.ru](mailto:m.alyasqar@gmail.ru);

Burgegulov A.D. - PhD student of the Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [dizel\_kz@bk.ru](mailto:dizel_kz@bk.ru);

Sametova A.A.- PhD student of the Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [sametova\_aygerim@mail.ru](mailto:sametova_aygerim@mail.ru);

Issimov N.T. - Head of Department at the International University of Engineering and Technology, Almaty, Kazakhstan, e-mail: [int\_nurdaulet@mail.ru](mailto:int_nurdaulet@mail.ru);

Jomartova Sh.A. - Doctor of Technical Sciences, Associate Professor, Al-Farabi Kazakh National University,Almaty, Kazakhstan, e-mail: jomartova@mail.ru;

Mazakov T.Zh. – Doctor of Physical and mathematical sciences, professor, Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [tmazakov@mail.ru](mailto:tmazakov@mail.ru)

IRSTI 50.43.15

**EVALUATION OF THE USE OF AUTOMATED CONTROL SYSTEMS FOR SOLVING TRAFFIC CONTROL PROBLEMS**

**1A.Tulegulov🖂, 1K. Akishev, 1D.Zhamangarin, 2N.Yurkov, 3L. Akisheva,**

**1S.A.Altynbek, 1N.S. Smakova**

1K. KulazhanovKazakh University of Technology and Business, Astana, Kazakhstan,

2Penza State University, Penza, Russia,

3Nazarbayev Intellectual School, Astana, Kazakhstan,

**🖂**Correspondent-author: [tad62@ya.ru](mailto:tad62@ya.ru)

One of the most pressing problems of megacities is traffic jams. With the growth of the urban population, there are more and more of them. Given the fact that there is no metro in Astana, the number of vehicles is growing quite quickly and urban public transport is also replenished with new buses every year. Attempts to introduce bus lanes partially eliminate the problem for public transport, but at the same time creates problems for private individual transport.

Theoretically, there are various solutions to the problem of congestion on urban highways, but in practice we face a number of additional problems. Attempts to create intelligent systems for regulating urban transport do not yield tangible results due to a chaotic approach to the main problem.

Transport system management is a set of various measures aimed at the effective functioning of this system through coordination, organization, and ordering of the elements of this system, both among themselves and with the external environment. The Intelligent Transport System (ITS) is a management system that implements innovative developments for managing traffic flows. As a result of using such systems, we get the so-called "smart roads". The article examines one of the components of ITS – the smart traffic light system.

**Keywords:** algorithms, information systems, traffic jams, integrated systems, information modeling, program code.

**ЖОЛ ҚОЗҒАЛЫСЫН БАҚЫЛАУ МІНДЕТТЕРІН ШЕШУ ҮШІН АВТОМАТТАНДЫРЫЛҒАН БАСҚАРУ ЖҮЙЕЛЕРІН ПАЙДАЛАНУДЫ БАҒАЛАУ**

**1А.Д. Тулегулов🖂,1К.М. Акишев, 1Д.С. Жамангарин, 2Н.К. Юрков, 3Л.К. Акишева,**

**1С. А.Алтынбек, 1Н.С. Смакова**

1Қ. Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан,

2Пенза мемлекеттік университеті, Пена, Ресей,

3Назарбаев Зияткерлік мектебі, Астана, Қазақстан,

e-mail: [tad62@ya.ru](mailto:tad62@ya.ru)

Мегаполистердің ең ауыр проблемаларының бірі-көлік кептелісі. Қала халқының өсуімен олардың саны артып келеді. Астанада метро жоқ екенін ескерсек, автокөлік саны тез өсуде және қалалық қоғамдық көлік жыл сайын жаңа автобустармен толықтырылып отырады. Автобус жолақтарын енгізу әрекеттері қоғамдық көлік мәселесін ішінара жеңілдетеді, бірақ сонымен бірге жеке жеке көлікке қиындық тудырады.

Теориялық тұрғыдан қалалық магистральдардың кептелісі мәселесін шешудің әртүрлі нұсқалары бар, бірақ іс жүзінде біз бірқатар қосымша мәселелерге тап боламыз. Қалалық көлікті реттеудің интеллектуалды жүйелерін құру әрекеттері негізгі проблемаларға хаотикалық көзқарасқа байланысты айтарлықтай нәтиже бермейді.

One of the problems is that congestion and situations are created at the main intersections of roads when flows of cars from one direction do not have time to complete the maneuver when turning and do not allow cars to move from the other direction accordingly. Attempts to create prohibited "islands" have not yielded any special results, since they do not allow automating the operation of traffic lights in terms of setting the glow time of the warning yellow color. There have even been illegal attempts to impose fines for driving through yellow traffic lights. The purpose of this article is to evaluate the use of automated control systems to solve traffic control problems.

**Түйін сөздер:** алгоритмдер, ақпараттық жүйелер, көлік кептелісі, интеграцияланған жүйелер, ақпараттық модельдеу, бағдарлама коды.

**ОЦЕНКА ИСПОЛЬЗОВАНИЯ АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ КОНТРОЛЯ ДОРОЖНОГО ДВИЖЕНИЯ**

**1А.Д. Тулегулов🖂 , 1К.М. Акишев, 1Д.С. Жамангарин, 2Н.К. Юрков, 3Л.К. Акишева, 1С.А.Алтынбек, 1Н.С.Смакова**

1Казахский университет технологии и бизнеса имени К. Кулажанова, Астана, Казахстан

2Пензенский государственный университет, Пенза, Россия

3Назарбаев интеллектуальная школа, Астана, Казахстан,

e-mail: [tad62@ya.ru](mailto:tad62@ya.ru)

Одной из наиболее наболевших проблем мегаполисов являются автомобильные пробки. С ростом городского населения их становится все больше и больше. Учитывая тот факт, что в Астане отсутствует метро количество автотранспорта растет довольно быстро и городской общественный транспорт также ежегодно пополняется новыми автобусами. Попытки введения автобусных полос частично снимают проблему для общественного транспорта, но в то же время создает проблем для частного индивидуального транспорта.

Теоретически существуют различные варианты решения проблемы загруженности городских магистралей, но на практике мы сталкиваемся с рядом дополнительных проблем. Попытки создать интеллектуальные системы регулирования городского транспорта не дают ощутимых результатов по причине хаотичного подхода к основной проблем.

Одной из проблем является то, что на основных перекрестках дорог создаются заторы и ситуации, когда потоки автомашин с одного из направлений при повороте не успевают завершить маневр и не позволяют соответственно двигаться автомашинам с другого направления. Попытки создавать запрещенные «островки» не дали особых результатов, так как не позволяют автоматизировать работу светофоров с точки зрения установки времени свечения предупреждающего желтого цвета. Даже были незаконные попытки введения штрафов за проезд на желтый свет светофора. В данной статье поставлена цель сделать оценку использования автоматизированных систем управления для решения задач контроля дорожного движения.

Ключевые слова: автоматизированные системы управления, алгоритмы, информационные системы, транспортные пробки, интегрированные системы, информационное моделирование, код программы.

**Introduction.** Traffic in conditions of congestion of roads with traffic and pedestrian flows requires innovative ways of regulation. Automated traffic management systems, which are a complex of technical means implementing certain technological algorithms for traffic flow management, have recently become increasingly relevant. The algorithm assumes the implementation of the principle of decomposition of decision-making, which can be considered generally accepted for network adaptive management methods. The decomposition of management is based on the division of the area into mutually overlapping zones. The center of each zone is an adjustable intersection, and the zone itself covers all intersections adjacent to the central one. The main purpose of the introduction of automated traffic management systems is to reduce the total delays of vehicles on a certain section of the road network, in the area of operation of this system – at an intersection, in an area or city[1-2].

Microcontroller intellectualization of traffic lights is a modern and promising direction for improving traffic light facilities. The presence of a digital reprogrammable microcontroller, such as the Siemens LOGO! 230rce, and small microcontrollers with built-in ESP32 Cam video cameras open up wide opportunities for the use of modern software and hardware that allow interactively intellectualize the process of optimizing traffic flows without significant material costs.

**Materials and methods.** The Siemens LOGO! 230RCE microcontroller in combination with the updated DC-2 linear controller allows you to optimize the algorithm of the entire information intelligent system. The technology of improvement of the DC-2 linear controller does not present great electrical complexity. Figure 1 shows the internal arrangement of the elements of the DC-2 controller. The device of the 230RCE programmable logic controller (PLC) is also presented here[3].

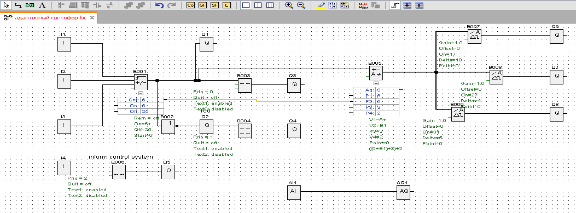


а б

a – DC-2 controller; b – LOGO!230RCE

**Figure 1 - The layout of the DC-2 and the general view of the Siemens LOGO! 230RCE microcontroller**

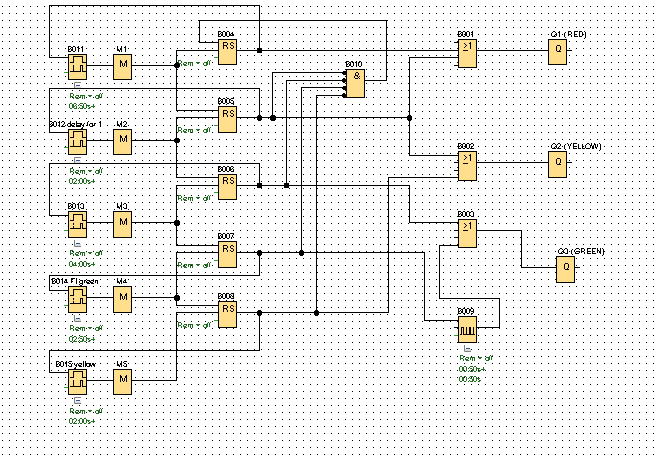
The dimensions of the DC-2 and the length of the DIN rail PLC LOGO! they are compact enough. Please note that the 230rce allows you to attach the rail without significant structural changes. In practice, an additional relay with the function of reprogramming and data exchange over Ethernet networks is installed in the housing of the DC-2. An important feature of the new hardware is the constantly expanding libraries of application programs and software LOGO! and the presence of a simulator in Soft Comfort. This is the personal logo of the microcontroller! The Web editor can be supplemented with a web server, which allows you to process network data streams, images obtained from video cameras specific to the OOP application. Figure 2 shows the program code written in the FBD graphical language [4-5].



**Figure 2 - The code of the adaptive traffic light program**

The algorithm of the traffic light with a recognition sensor located in the traffic light zone of cars takes into account their number and if the specified number of cars is exceeded (in our case, 15 cars), the green light turns on[6].

Figure 3 shows the standard program for switching on the temporary phases of green, yellow and red light.



**Figure 3 - The standard operating program of the traffic light in the "green wave" mode**

The special feature of the code is that it can be reloaded taking into account traffic jams, emergencies and night mode (yellow, which lights up again).

Various software solutions are used to eliminate the phenomena of congestion and forced unreasonable delays of cars at intersections in front of traffic lights[7]. The algorithm and Program for calculating the number of vehicles can be adapted to a pedestrian traffic light, which is of particular relevance for streets with divided lanes near the centerline of the bus (for example, on Timiryazev Street in Almaty).

In this regard, an upgraded computer traffic management system in Almaty is needed, which should calculate online the optimal and coordinated time intervals for switching on the phases of red, yellow, and green traffic lights at adjacent intersections, ensuring traffic flow along the "green" wave. Among such intelligent control systems, we note the automated control system "Smart traffic lights" of New York City.

The intelligent traffic light control system of large cities is a complex and expensive automated system, however, reducing travel time, reducing vehicle downtime at intersections, reducing harmful emissions into the atmosphere, increasing the mobility of the population as a whole compensate for these costs. So, traffic management of a megalopolis is a complex and multilevel technological process.

Complex automation requires a powerful computer network with central controllers of supercomputers, on the periphery of which the IoT device constantly transmits data to the traffic control center via communication channels.

A promising task of modern automation of the traffic management information system is research on "advanced" traffic management and traffic light regulation. The implemented software products should calculate indicators interactively, taking into account the requirements of local government systems installed at the nodal points of the city's road network. Also promising tasks are the creation of network simulators for microcontroller control of "smart intersections", and then their localization into more complex networks on the scale of a polycentric district and the city as a whole.

In urban conditions, the following methods of traffic management are used:

- Control by stopping vehicles. For example, by means of traffic lights. Each traffic flow moving through the intersection is affected by three light signals — green, yellow and red.

- The direction of traffic flow bypassing the area where congestion is formed.

- Information management. The driver receives information about the traffic situation on his way, and decides on the further route on his own.

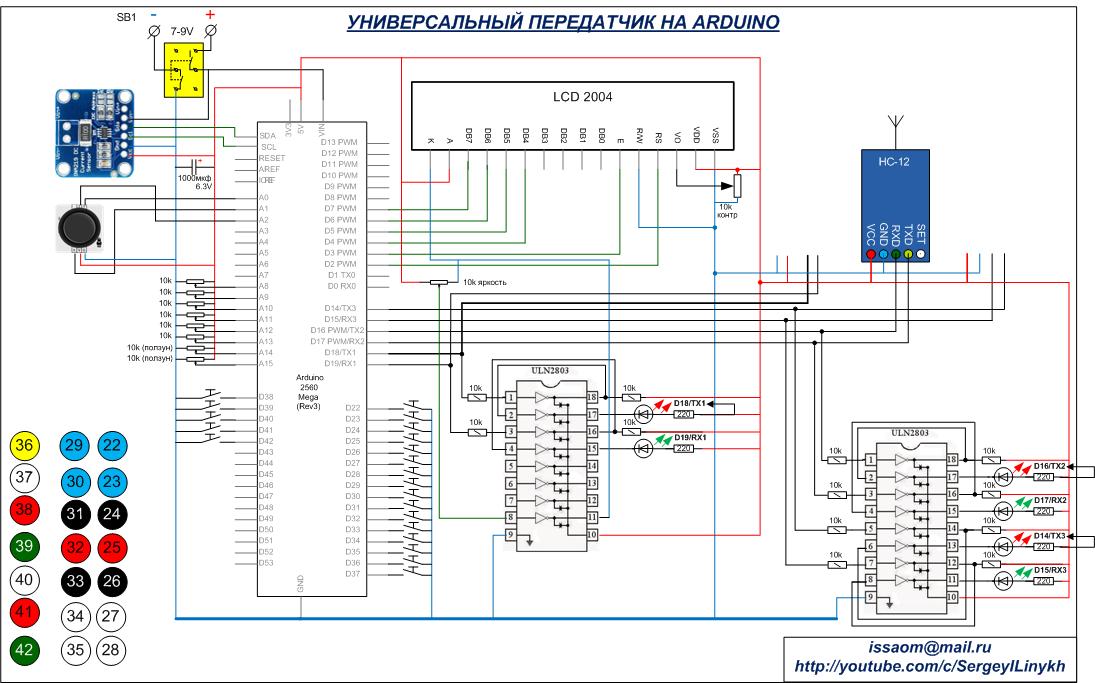
- Coordinated management. It allows you to increase the capacity of roads and reduce the risk of accidents by streamlining traffic. The following methods are used to implement coordinated management

- Rigid software control. Coordination programs and the order of their change are calculated in advance based on movement data collected, as a rule, by field measurements.

- Software management with forecast. The selection of the program activation points, the duration of their operation and their adaptation to the speed of traffic flow is performed based on data from transport detectors.

-Adaptive control. Each cycle is implemented with the calculation of coordination programs based on traffic flow data at each intersection and in each control phase.

**Results and discussions.** As a result of the conducted research, the software and hardware of the ESP32 Cam interface was developed on the principle of a "green wave" for special services vehicles using a QR code. Similarly, the hardware and software of the ESP32 QR Code Reader interface for medical vehicles was developed. The "smart" HR software appeared during the implementation of a library program in which, with standard QR code recognition, a text message transcribing the text of the QR code was displayed on the screen of the COM port of the Arduino IDE environment [8, 9] (Fig.4).



**Figure 4 - Electrical diagram of an intelligent interface for controlling a "smart" traffic light using a QR code**

In the case of fire trucks, often driven by a small column, the "Green Wave" has a longer turn-on time, and this information is entered into the QR code. Below is a program that implements an algorithm for adaptive activation of the "green wave" in the codes of the Arduino IDE environment:

Void setup()

{

TM1283 mod;

Byte num();

int pinA = 3;

int pinD = 5;

pinMode (pinA, input);

digitalWrite(pin1);

Boolean dSW(pinC);

}

Int encoderPosCount(stepsPerrevo5){

myStepper = 7;

int pinDLast = 8;

Serial.begin (362);

Value\_Y = analogRead (axis-X);

Bool isvalid = false;

MFRC522 mfrc522(DD\_PIN;

Byte uidCar;

delay(1);

mode = cloc() / 5;

if X=4 then {

go to 10(“.”);

delay(100);

}

As can be seen from the text of the application, the webcam reads the QR code, analyzes the text and selects from it the number specified in the text format and converts it to a digital format. Then the coil of the intelligent traffic light relay turns on to turn on the yellow light and then the green light.

**Conclusion.** In conclusion, it should be noted that the assessment of the informativeness of adaptive control algorithms makes it possible to optimally select the necessary operating mode for traffic flows, which in turn will significantly reduce the number of congestion on the streets of megacities.

The use of this system has made it possible to increase the efficiency of the automated control system by 18%. The developed software and hardware of the ESP32 Cam interface based on the "green wave" principle for special services vehicles using a QR code has shown its viability and practicality. The created program code for the Arduino IDE environment is adapted to these requirements. The article is based on the results of research carried out within the framework of the funded project "A neural computer view of the smart traffic light of megacities of the country".

**References**

1.Badaguev B.T. Jekspluatacija transportnyh sredstv (organizacija i bezopasnost' dvizhenija). – M.: Al'fa-Press, 2012. - 239 s. ISBN 978-5-94280-556-2 [in Russian]

2. Dudko N.I., Petrovec V.R., Bershadskij V.F. Bezopasnost' dvizhenija mehanicheskih transportnyh sredstv: posobie - Gorki: BGSHA, 2014. - 238 s. ISBN 978-985-467-490-2 [in Russian]

3. Blinkin M.Ja. Bezopasnost' dorozhnogo dvizhenija: istorija voprosa, mezhdunarodnyj opyt, bazovye institucii.-M.: Izd. dom Vysshej shkoly jekonomiki, 2013.- 240 s. ISBN: 978-5-7598-1086-5 [in Russian]

4. Volkov V.S. Jelektrooborudovanie transportnyh i transportno-tehnologicheskih mashin: uchebnoe posobie - M.: Akademija, 2010. - 208 s. ISBN 978-5-7695-5749-1 [in Russian]

5. Gorev, A. Je.Osnovy teorii transportnyh sistem: uchebnoe posobie / A. Je. Gorev;SPbGASU. – SPb., 2010. - 214 s. ISBN 978-5-9227-0266-9 [in Russian]

6. Majboroda M.E., Bednarskij V.V. Gruzovye avtomobil'nye perevozki. - M.: Feniks, 2008.- 442 s. ISBN 978-5-222-14364-3 [in Russian]

7. Klinkovshtejn G.I. Organizacija dorozhnogo dvizhenija: uchebnik dlja vuzov. – 5-e izd., pererab. i dop.- M: Transport, 2001 - 247. ISBN 9785277022405

8. Urykov V.A., Zelenina L.I. Modeli transportnogo infrastrukturnogo kompleksa // [https://web.snauka.ru/issues/2014. 18.06.2020](https://web.snauka.ru/issues/2014.%2018.06.2020). [in Russian]

9. Morozov I.I. i dr. Chislennoe issledovanie transportnyh potokov na osnove gidrodinamicheskih modelej // Komp'juternye issledovanija i modelirovanie.-2011. -T. 3(4).-S. 389-412. [in Russian]

***Сведения об авторах***

Тулегулов А.Д. - кандидат физико-математических наук, ассоциированный профессор, Казахский университет технологии и бизнеса имени К. Кулажанова, Астана, Казахстан, e-mail: tad62@ya.ru;

Акишев К.М. - кандидат технических наук, ассоциированный профессор, Казахский университет технологии и бизнеса имени К. Кулажанова, Астана, Казахстан, e-mail: akmail04@ya.ru;

Жамангарин Д. С.- PhD, Казахский университет технологии и бизнеса имени К. Кулажанова, Астана, Казахстан, е-mail: [Dus\_man89@mail.ru](mailto:Dus_man89@mail.ru);

Юрков Н.К. - доктор технических наук, профессор, Пензенский государственный университет, Пенза, Россия, е-mail: [kipra@mail.ru](mailto:kipra@mail.ru);

Акишева Л.К.-исследователь, Назарбаев интеллектуальная школа, Астана, Казахстан, е-mail: akmail04@ya.ru

Алтынбек С. А. **-** PhD, ассоциированный профессор, Казахский университет технологии и бизнеса, Астана, Казахстан, е-mail: [serik\_aa@bk.ru](mailto:serik_aa@bk.ru);

Смакова Н.С. - PhD, ассоциированный профессор, Казахский университет технологии и бизнеса, Астана, Казахстан, е-mail:nuri\_5@mail.ru

***Information about the authors***

Tulegulov A.D. - Candidate of Physical and Mathematical Sciences, Associate Professor, Kazakh University of Technology and Business named after K. Kulazhanov, Astana, Kazakhstan, e-mail: [tad62@ya.ru](mailto:tad62@ya.ru);

Akishev K.M. - Candidate of Technical Sciences, Associate Professor, Kazakh University of Technology and Business named after K. Kulazhanov, Astana, Kazakhstan, e-mail: [akmail04@ya.ru](mailto:akmail04@ya.ru);

Zhamangarin D. S.- PhD, K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan,

e-mail: [Dus\_man89@mail.ru](mailto:Dus_man89@mail.ru);

Akisheva L.K.-researcher, Nazarbayev Intellectual School, Astana, Kazakhstan, e-mail: akmail04@ya.ru

Yurkov N.K. - Doctor of Technical Sciences, Professor, Penza State University, Penza, Russia, е-mail: [kipra@mail.ru](mailto:kipra@mail.ru);

Altynbek S. A. PhD, Associate Professor, Kazakh University of Technology and Business, Astana, Kazakhstan, е-mail: [serik\_aa@bk.ru](mailto:serik_aa@bk.ru)

Smakova N.S.- PhD, Associate Professor, Kazakh University of Technology and Business, Astana, Kazakhstan, е-mail: nuri\_5@mail.ru

МРНТИ 20.53.19

**ИДЕНТИФИКАЦИИ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ МЕТОДОМ КВАЗИЛИНЕАРИЗАЦИИ**

**1,2Т.Ж. Мазаков🖂, 1Ш.А. Джомартова, 1А.Т. Мазакова, 3Г.Ч. Тойкенов,**

**1,2М.С. Алиаскар, 3У.Г. Тойкенова**

1Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,

2Международный инженерно-технологический университет, Алматы, Казахстан,

3Казахский национальный женский педагогический университет, Алматы, Казахстан

**🖂**Correspondent-author: [tmazakov@mail.ru](mailto:tmazakov@mail.ru)

Целью данной работы является определение параметров математической модели, описывающей прорывы селевых потоков через гидротехнические объекты, такие как дамбы и плотины. Достижение этой цели позволит соотнести разработанную модель с реальными наблюдениями и данными. Важность исследования обусловлена увеличением частоты и интенсивности катастрофических наводнений, вызванных разрушением таких сооружений, что подчеркивает необходимость разработки эффективных методов для прогнозирования и предотвращения подобных инцидентов. Методы и подходы, предложенные в данной работе, обладают высоким практическим значением для оценки вероятных рисков, организации эвакуационных мероприятий и снижения ущерба от аварий, связанных с разрушением дамб и плотин.

**Ключевые слова:** гидротехнические сооружения (ГТС), параметры, идентификация, прорыв дамбы, селевой поток, математическое моделирование.

**МАТЕМАТИКАЛЫҚ МОДЕЛЬДЕРДІ КВАЗИЛИНЕАРИЗАЦИЯ ӘДІСІМЕН СӘЙКЕСТЕНДІРУ**

**1,2Т.Ж. Мазаков🖂, 1Ш.А. Джомартова, 1А.Т. Мазақова, 3Г.Ч Тойкенов,**

**1,2М.С. Әлиасқар, 3Тойкенова У.Г.**

¹Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,  
²Халықаралық инженерлік-технологиялық университет, Алматы, Қазақстан,  
³Қазақ ұлттық қыздар педагогикалық университеті, Алматы, Қазақстан,  
e-mail: [tmazakov@mail.ru](mailto:tmazakov@mail.ru)

Бұл зерттеудің мақсаты — гидротехникалық құрылыстар, соның ішінде бөгеттер мен дамбалар арқылы болатын сел ағындарының жарылуын сипаттайтын математикалық модельдің параметрлерін анықтау. Бұл мақсатқа қол жеткізу әзірленген модельді нақты бақылаулармен және деректермен салыстыруға мүмкіндік береді. Зерттеудің өзектілігі осындай құрылыстардың бұзылуы салдарынан болатын апатты су тасқындарының жиілігі мен қарқындылығының артуымен ерекшеленеді, бұл осындай оқиғаларды болжау және алдын алу үшін тиімді әдістерді әзірлеу қажеттігін көрсетеді. Осы жұмыста ұсынылған әдістер мен тәсілдер ықтимал тәуекелдерді бағалауға, эвакуация шараларын ұйымдастыруға және дамбалар мен бөгеттердің бұзылуымен байланысты апаттардан келетін шығындарды азайтуға үлкен практикалық мәнге ие.

**Түйін сөздер**: гидротехникалық құрылыстар (ГТҚ), параметрлер, сәйкестендіру, дамбаның жарылуы, сел ағыны, математикалық модельдеу.

**IDENTIFICATION OF MATHEMATICAL MODELS USING THE QUASILINEARIZATION METHOD**

**1,2 T.Zh. Mazakov🖂, ¹S 1h.A. Jomartova, 1A.T. Mazakova, 3G.Ch. Toykenov,  
 1,2M.S. Aliaskar, 3U.G. Toykenova**

¹Al-Farabi Kazakh National University, Almaty, Kazakhstan,  
²International Engineering and Technology University, Almaty, Kazakhstan,  
³Kazakh National Women's Teacher Training University, Almaty, Kazakhstan,  
e-mail: [tmazakov@mail.ru](mailto:tmazakov@mail.ru)

The aim of this study is to determine the parameters of a mathematical model that describes debris flow breaches through hydraulic structures, such as dams and embankments. Achieving this goal will allow the developed model to be correlated with real observations and data. The importance of the research is underscored by the increasing frequency and intensity of catastrophic floods caused by the failure of such structures, emphasizing the need to develop effective methods for predicting and preventing such incidents. The methods and approaches proposed in this work hold significant practical value for assessing potential risks, organizing evacuation measures, and reducing the damage from accidents related to dam and embankment failures.

**Keywords**: hydraulic structures (HS), parameters, identification, dam breach, debris flow, mathematical modeling.

**Введение**. За последнее столетие зарегистрировано более тысячи разрушений гидротехнических сооружений по всему миру, где основными причинами выступают как природные, так и антропогенные факторы. Среди основных факторов аварий на ГТС [1] выделяются:

* прохождение экстремальных объемов воды, что может вызвать переполнение водохранилища и нарушить нормальную работу сбросных сооружений, приводя к переливу через гребень плотины и формированию прорыва;
* износ основных компонентов плотин и гидромеханического оборудования вследствие длительной эксплуатации, что создает риск возникновения прорывов;
* ошибки персонала, недостаток мониторинга опасных ситуаций и недостаточные прогнозные данные по паводкам;
* террористические акты, направленные на разрушение плотин.

Примеры значимых происшествий:

1. В Италии, в 1963 году, прорыв дамбы Вайонт был вызван оползнем объёмом около 260 млн куб. метров, который обрушился в водохранилище, вызвав перелив и уничтожение нескольких деревень, погибло около 2000 человек.
2. В Китае, прорыв дамбы Банкиао в 1975 году, спровоцированный чрезмерными осадками, привёл к гибели более 170 000 человек, а миллионы остались без жилья.
3. Аварии в Бразилии на хвостохранилищах горнодобывающих предприятий, включая прорыв дамбы в 2019 году в Брумадинью, вызвали масштабные экологические катастрофы и многочисленные жертвы.

Кроме того, техногенная катастрофа на Саяно-Шушенской ГЭС в России, произошедшая 17 августа 2009 года, привела к гибели 75 человек и нанесению серьезного ущерба объекту. Прорыв дамбы в селе Кызылагаш Алматинской области в 2010 году также привел к человеческим жертвам и разрушениям.

**Математические модели селевых потоков** служат важным инструментом для анализа и прогнозирования этих сложных и разрушительных природных явлений. Селевые потоки представляют собой массу воды, грунта и камней, которая перемещается под действием гравитации по склонам и руслам рек, требуя тщательного анализа и моделирования.

Основные подходы к моделированию селевых потоков:

1. Эмпирические модели, опирающиеся на статистические данные прошлых событий.
2. Физические модели, воссоздаваемые в лабораторных условиях для изучения поведения селевых потоков.
3. Численные модели, использующие уравнения гидродинамики и механики грунтов.

Применение математического моделирования позволяет:

1. Оценивать риски и уязвимость:
   * Определение зон потенциального воздействия селевых потоков.
   * Использование при планировании территорий и разработке мер по снижению рисков.
2. Проектировать инженерные сооружения:
   * Поддержка при проектировании дамб, каналов и барьеров.
   * Оценка эффективности защитных сооружений.
3. Создавать системы раннего предупреждения:
   * Интеграция моделей с метеорологическими данными для прогнозирования селевых потоков.
   * Обеспечение своевременного оповещения населения и служб реагирования.

Примеры математических моделей для анализа селевых потоков и оползней [2-3]:

***1) Модель DEBRIS-2D***

DEBRIS-2D — это специализированная двумерная численная модель, предназначенная для симуляции движения селевых потоков и грязекаменных лавин по сложному рельефу местности. Она позволяет анализировать динамику селевых масс с учётом их реологических характеристик и взаимодействия с топографией. Модель учитывает различные реологические подходы, такие как модели Бингема, Гершеля-Балкли и турбулентного потока, что помогает описывать неньютоновские свойства селевых потоков, включая зависимость вязкости от напряжения сдвига.  
DEBRIS-2D использует высокоточные цифровые модели рельефа (DEM), что обеспечивает детальное представление поверхности и точное прогнозирование путей движения потоков. Модель решает двухмерные уравнения глубинно-усредненного потока с использованием методов конечных объемов или конечных разностей, что позволяет эффективно моделировать нестационарные процессы. Она применяется для определения вероятных маршрутов распространения селевых потоков и зон повышенного риска, а также при планировании мер по снижению рисков, эвакуации и размещению защитных сооружений.

***2) Модель FLO-2D***

FLO-2D — двумерная гидродинамическая модель, предназначенная для моделирования поверхностного течения жидкости на обширных территориях. Изначально разработанная для прогнозирования паводков, модель FLO-2D была адаптирована для симуляции селевых потоков, лахаров и других гидрологических явлений с учётом неньютоновских свойств. Модель решает полные уравнения мелкой воды без упрощений, что позволяет учитывать сложные потоки с переменными скоростями и направлениями. Она использует квадратную расчётную сетку, что упрощает моделирование и повышает его эффективность при обработке больших объемов данных.  
FLO-2D поддерживает различные реологические модели для вязкопластичных материалов, таких как селевые потоки, и учитывает влияние концентрации частиц и вязкости на движение потока. Модель совместима с ГИС, что облегчает подготовку данных и анализ результатов. FLO-2D широко используется для прогнозирования распространения паводковых вод, оценки риска затопления и моделирования высококонцентрированных смесей воды и твёрдых частиц, характерных для селевых потоков. Она также применяется при проектировании гидротехнических сооружений, таких как дамбы и каналы.

***3) Модель DAN (Динамический анализ оползней)***

DAN (Dynamic Analysis of Landslides) — одномерная численная модель, разработанная для анализа быстродвижущихся оползней, селевых потоков и обвалов. Эта модель ориентирована на анализ глубинно-усредненного движения массы вдоль заданной траектории с учётом изменения реологических характеристик материала в процессе движения.  
DAN поддерживает несколько реологических моделей, таких как:

* **Модель сухого трения** — используется для материалов, у которых сопротивление движению пропорционально нормальному напряжению;
* **Модель Бингема** — учитывает начальное напряжение сдвига и постоянную вязкость;
* **Модель Гершеля-Балкли** — обобщает модель Бингема, вводя показатель степени для напряжения сдвига;
* **Модель Воеллми** — сочетает сухое трение и турбулентное сопротивление, часто применяемая для снежных лавин и обвалов.

Модель DAN позволяет изменять реологические параметры по траектории движения, учитывая процессы разжижения или уплотнения материала. Она помогает прогнозировать дальность, скорость и время прибытия оползневых масс, а также оценивает потенциальный ущерб зданиям, дорогам и другим объектам. Модель активно используется в планировании защитных инженерных мер и разработке систем раннего предупреждения.

**Материалы и методы.** Хотя все упомянутые модели описываются уравнениями в частных производных, в работе [4] был предложен алгоритм, позволяющий свести их к системе обыкновенных дифференциальных уравнений. После того как структура математической модели определена (т.е. задан вид уравнений с неуточнёнными параметрами), встает задача параметрической идентификации — нахождения численных значений параметров на основе известных входных и выходных данных.

Ввиду сложности моделируемого объекта применимы лишь пассивные методы идентификации, предполагающие использование данных, полученных в процессе нормального функционирования объекта.

Рассмотрим следующую математическую модель, описываемую системой обыкновенных дифференциальных уравнений:

|  |  |
| --- | --- |
| , | (1) |
|  |  |
| . | (2) |

Здесь – ny-мерный вектор состояния модели, p – np-мерный вектор параметров.

В качестве меры близости системы и модели выберем функционал:

|  |  |
| --- | --- |
| , | (3) |

где – значение выходного сигнала в момент времени в год j, полученное с использованием математической модели.

Таким образом, задача параметрической идентификации сводится к задаче минимизации функционала *S* по параметрам *p*:

|  |  |
| --- | --- |
|  | (4) |

где *P* – пространство допустимых значений параметров *p*.

Метод квазилинеаризации, предложенный Беллманом и Калабой [5] для решения краевых задач нелинейных дифференциальных уравнений, может быть также использован для задачи параметрической идентификации при условии, что параметры постоянны. Этот метод характеризуется высокой степенью сходимости.

**Результаты и обсуждение.** Дополним систему обыкновенных дифференциальных уравнений (1)-(2) приобретает следующий вид с добавлением уравнения:

|  |  |
| --- | --- |
| , | (5) |

где – параметры модели, подлежащие идентификации.

Введем обозначения: , , где - вектор размерности – *).* Тогда систему дифференциальных уравнений (1) и (5) можно записать как:

|  |  |
| --- | --- |
|  | (6) |

Предположим, что *k*-я оценка вектора состояния , известна. Разлагая правую часть (6) в окрестности траектории , в ряд Тейлора и ограничиваясь линейной частью, получаем систему дифференциальных уравнений для -го приближения вектора:

|  |  |
| --- | --- |
|  | (7) |

где

,

.

Решение уравнения (7) имеет вид:

|  |  |
| --- | --- |
| ,  , | (8) |

где - является решением следующей матричной системы дифференциальных уравнений:

|  |  |
| --- | --- |
| , | (9) |

а – решением неоднородного дифференциального уравнения:

|  |  |
| --- | --- |
| . | (10) |

Здесь E – единичная матрица.

Разобьем матрицу на блочные матрицы:

Подставим решение (8) в функционал (3):

|  |  |
| --- | --- |
| , | (11) |

где – значение -го приближения выходного сигнала в момент времени , для года j. С учётом обозначений:

, для m= и ∀ k, j,

, для m= и ∀ k, j,

получим следующий функционал, зависящий от неизвестных :

|  |  |
| --- | --- |
| . | (12) |

Приравняем к нулю частные производные от по :

|  |  |
| --- | --- |
|  | (13) |

которую решим методом Гаусса для определения . Это позволит найти траекторию , на интервале , тем самым определяя - ю оценку вектора состояния . Процесс продолжается, пока очередные оценки и не станут близки по заданному критерию.

Таким образом, алгоритм идентификации на основе квазилинеаризации можно представить следующим образом:

Шаг 1. Задаем начальные значения , инициализируем .

Шаг 2. Для определения -го приближения интегрируем систему:

при начальных условиях

, для m=,

, для m=.

Шаг 3. Решаем задачи Коши (9) и (10).

Шаг 4. Для получения -й оценки параметров *p* решаем систему линейных алгебраических уравнений (13).

Шаг 5. Если , процесс идентификации завершается; иначе — возврат к шагу 2 (где – требуемая точность).

Таким образом, математические модели для анализа угроз прорыва дамб обеспечивают безопасность гидротехнических объектов и позволяют: прогнозировать последствия возможных аварий, разрабатывать меры для снижения рисков и повышать осведомлённость населения к ЧС. Развитие вычислительных технологий и интеграция с геоинформационными системами делают модели более точными, хотя их результативность зависит от качества исходных данных и опыта специалистов.

Предложенный алгоритм реализован программно и проходит экспериментальные расчёты.

**Выводы.** Модели DEBRIS-2D, FLO-2D и DAN представляют собой ключевые инструменты в геотехническом и гидрологическом моделировании опасных природных процессов. Их использование позволяет: а) углубленно изучить механизмы движения селевых потоков и оползней; б) точно прогнозировать зоны потенциального воздействия и пути распространения опасных потоков; в) разрабатывать эффективные стратегии управления рисками и планирования территорий. Выбор конкретной модели определяется спецификой задачи, доступными данными и уровнем детализации, необходимым для анализа. Интеграция результатов моделирования с практическими мерами по снижению рисков способствует повышению уровня безопасности населения и устойчивости инфраструктуры перед воздействием опасных геологических процессов.

В данной работе разработана схема идентификации параметров моделей, позволяющая адаптировать их к конкретным водоемам. Этот алгоритм предоставляет возможность более точного моделирования процессов разрушения плотин и распространения селевых волн, что является важным аспектом для предотвращения катастрофических наводнений. Результаты численных расчетов показывают, что предложенная схема эффективно применяется для идентификации параметров выбранных математических моделей. Программное обеспечение, реализующее данный алгоритм, демонстрирует высокую точность и надежность расчетов, что подтверждается сравнением с существующими моделями и экспериментальными данными.

Перспективы дальнейших исследований связаны с интеграцией разработанной схемы идентификации параметров моделей с современными системами мониторинга и предупреждения о чрезвычайных ситуациях. Это позволит повысить эффективность мер по защите населения и инфраструктуры от последствий селевых прорывов и других гидрологических угроз.

***Финансирование.*** *Работа выполнена при финансовой поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан в рамках гранта AP19678157 «Разработка программно-аппаратного комплекса мониторинга состояния уровня заполняемости водоёма» (2023-2025)*

**Литература**

1.Стихийные бедствия и техногенные катастрофы: Превентивные меры/ The World Bank and The United Nations; пер. с англ.-М.: Альпина Паблишер.-2011.-312 с.

ISBN 978-5-9614-1527-8

2. [Francesca Aureli](https://www.researchgate.net/profile/Francesca-Aureli?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), [Paolo Mignosa](https://www.researchgate.net/profile/Paolo-Mignosa?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), [Massimo Tomirotti](https://www.researchgate.net/profile/Massimo-Tomirotti?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19). Numerical simulation and experimental verification of Dam-Break flows with shocks//[Journal of Hydraulic Research](https://www.researchgate.net/journal/Journal-of-Hydraulic-Research-1814-2079?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19).-2010.-Vol.2000(3).-P.197-206. DOI [10.1080/00221680009498337](http://dx.doi.org/10.1080/00221680009498337)

3.[Tomaž Podobnikar](https://www.researchgate.net/profile/Tomaz-Podobnikar?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) Methods for visual quality assessment of a digital terrain model// S.A.P.I.EN.S [Surveys and Perspectives Integrating Environment and Society](https://www.researchgate.net/journal/Surveys-and-Perspectives-Integrating-Environment-and-Society-1993-3819?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), Special issue /-2009.-Vol.2(2).-P.1-10. URL http://journals.openedition.org/sapiens/738

4. Mazakova A.T., Jomartova Sh.A., Mazakov T.Zh., Ziyatbekova G.Z., Begaliyeva K.B.

Digital Determination of the Thermal Conductivity of a Square Bar by Reduction to a System of Integral Equations//Ecological Footprint of the Modern Economy and the ways to Reduce it Advances in Science, Technology&amp;Innovation.-2024.-P.167-171.

DOI:10.1007/978-3-031-49711-7\_29

5. Гроп Д. Методы идентификации систем.// Перевод с англ. В.А. Васильева, В.И. Лопатина. - Под ред. Е.И. Кринецкого М.: Мир, 1979. – 302 с.

**References**

1.Stihijnye bedstvija i tehnogennye katastrofy: Preventivnye mery/ The World Bank and The United Nations; per. s angl.-M.: Al'pina Pablisher.-2011.-312 s.

ISBN 978-5-9614-1527-8. [in Russian]

2. [Francesca Aureli](https://www.researchgate.net/profile/Francesca-Aureli?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), [Paolo Mignosa](https://www.researchgate.net/profile/Paolo-Mignosa?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), [Massimo Tomirotti](https://www.researchgate.net/profile/Massimo-Tomirotti?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19). Numerical simulation and experimental verification of Dam-Break flows with shocks//[Journal of Hydraulic Research](https://www.researchgate.net/journal/Journal-of-Hydraulic-Research-1814-2079?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19).-2010.-Vol.2000(3).-P.197-206. DOI [10.1080/00221680009498337](http://dx.doi.org/10.1080/00221680009498337)

3.[Tomaž Podobnikar](https://www.researchgate.net/profile/Tomaz-Podobnikar?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) Methods for visual quality assessment of a digital terrain model// S.A.P.I.EN.S [Surveys and Perspectives Integrating Environment and Society](https://www.researchgate.net/journal/Surveys-and-Perspectives-Integrating-Environment-and-Society-1993-3819?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19), Special issue /-2009.-Vol.2(2).-P.1-10. URL http://journals.openedition.org/sapiens/738

4. Mazakova A.T., Jomartova Sh.A., Mazakov T.Zh., Ziyatbekova G.Z., Begaliyeva K.B.

Digital Determination of the Thermal Conductivity of a Square Bar by Reduction to a System of Integral Equations//Ecological Footprint of the Modern Economy and the ways to Reduce it Advances in Science, Technology&amp;Innovation.-2024.-P.167-171.

DOI:10.1007/978-3-031-49711-7\_29

5. Grop D. Metody identifikacii sistem.// Perevod s angl. V.A. Vasil'eva, V.I. Lopatina. - Pod red. E.I. Krineckogo M.: Mir, 1979.- 302 s. [in Russian]

***Сведение об авторах***

Мазаков Т.Ж.-доктор физико-математических наук, профессор, Казахский национальный университетим. аль-Фараби, Алматы, Казахстан, e-mail: tmazakov@mail.ru;

Джомартова Ш.А. - доктор технических наук, доцент, Казахский национальный университет им. аль-Фараби, Алматы, Казахстан, e-mail: jomartova@mail.ru;

Мазакова А.Т. – докторант Казахского национального университета им.аль-Фараби, e-mail: [aigerym97@mail.ru](mailto:aigerym97@mail.ru);

Тойкенов Г.Ч. - кандидат физико-математических наук, доцент, Казахский национальный женский педагогический университет, Алматы, Казахстан, e-mail: gumyrbektoike@mail.ru;

Алиаскар М.С. - докторант Казахского национального университета им.аль-Фараби, Алматы, Казахстан, e-mail: m.alyasqar@gmail.ru;

Тойкенова У.Г.- докторант Казахского национального женского педагогического университета, Алматы, Казахстан, e-mail: ulpantoikenova@gmail.com

***Information about the authors***

Mazakov T.Zh. – Doctor of Physical and mathematical sciences, professor, Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: tmazakov@mail.ru;

Jomartova Sh.A. - Doctor of Technical Sciences, Associate Professor, Al-Farabi Kazakh National University,Almaty, Kazakhstan, e-mail: jomartova@mail.ru;

Mazakova A.T. - PhD student of the Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [aigerym97@mail.ru](mailto:aigerym97@mail.ru);

Tokenov G.Ch. - Candidate of Physical and Mathematical Sciences, Associate Professor, Kazakh National Women's Pedagogical University, Almaty, Kazakhstan, e-mail: gumyrbektoike@mail.ru;

M.S. Aliaskar - PhD student of the Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [m.alyasqar@gmail.ru](mailto:m.alyasqar@gmail.ru);

Toikenova U.G. - PhD student of the Kazakh National Women's Pedagogical University, Almaty, Kazakhstan, e-mail: ulpantoikenova@gmail.com

IRSTI 28.23.25

**Identifying effective machine learning algorithms for sentimental analysis of comments in the Kazakh language**

**1N.K. Mukazhanov🖂,  2L.Sh.** **Cherikbayeva, 1A.M. Kassenkhan, 1Zh.M. Alibieva,**

**1,3M. Turdalyuly**

1Satbayev University, Almaty, Kazakhstan,

2Al-Farabi Kazakh National University, Almaty, Kazakhstan,

3International Engineering and Technological University

**🖂**Corresponding author: [n.mukazhanov@satbayev.university](mailto:n.mukazhanov@satbayev.university)

This article presents the results of an analysis of machine learning algorithms for sentimental data analysis in the Kazakh language, and as a result of the analysis, effective algorithms are determined. With the increasing volume of Kazakh-language content on social networks, news and online stores, the need for tools and methods for processing data in the Kazakh language has also increased in order to obtain valuable information about people's opinions and views. Therefore, the dataset used in the study was collected from real online stores and news sites. The volume of the collected data set is 1500 records, 80% of which were used for training the algorithms, and 20% for testing. For sentimental data analysis, machine learning algorithms such as logistic regression, multinomial naive Bayes, support vector machine (SVM), XGBoost and long short-term memory (LSTM) deep learning are considered. The study tested algorithms by increasing the dataset from 500 records to 1500 records, and various algorithm methods such as individual, ensemble, and augmented were implemented and tested. The results obtained during testing were presented in terms of algorithm accuracy.

**Keywords:**sentiment analysis, machine learning, deep learning, NLP, comments, dataset

**Қазақ тіліндегі пікірлерді сентименталды талдау үшін машиналық оқытудың тиімді алгоритмдерін анықтау**

**1Н.К. Мукажанов🖂, 2Л.Ш.Черикбаева, 1А.М.Касенхан, 1Ж.М.Алибиева,   
1,3М. Тұрдалыұлы**

1Satbayev University, Алматы, Қазақстан,

2Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

3Халықаралық инженерлік-технологиялық университет,

e-mail: [n.mukazhanov@satbayev.university](mailto:n.mukazhanov@satbayev.university)

Ұсынылып отырған мақалада қазақ тілді деректерді сентименталды талдау үшін машиналық оқыту алгоритмдеріне талдау жасалынды, талдау нәтижесінде тиімді алгоритмдерді анықтау қарастырылды. Әлеуметтік желілерде, жаңалықтар жəне интернет дүкендердегі қолданушылардың пікірлері сияқты қазақ тіліндегі контенттің көлемі артуына байланысты, қазақ тілді деректерді өңдеу, адамдардың пікірі мен көзқарастары туралы құнды ақпаратты алу құралдары мен əдістеріне де қажеттілік артқан. Сондықтан, зерттеуде қолданылған деректер жинағы нақты интернет дүкендер мен жаңалықтар сайтынан жинақталды. Жинақталған деректердің көлемі 1500 жазба, оның 80% алгоритмдері жаттықтыру үшін, ал 20% тестілеу үшін пайдаланылды. Жинақталған деректерді сентименталды талдау үшін машиналық оқытудың Логистикалық регрессия, Multinomial Naive Bayes, Liner SVM, XGBoost және тереңдете оқытудың Long short-term memory (LSTM) қарастырылды. Зерттеу барысында деректер жинағы 500 жазбадан 1500 жазбаға дейін арттыру арқылы сынақ жасалынды, ал алгоритмдердің жеке, ансамбльдік және LSTM алгоритмінің толтырылған тізбектер әдісі сияқты түрлі әдістері жүзеге асырылып тестіленді. Тестілеу барысында алынған нәтижелер алгоритмдердің дәлдік көрсеткіштері бойынша ұсынылды.

**Түйін сөздер:** сентименталды талдау, машиналық оқыту, тереңдете оқыту, NLP, пікірлер, деректер жинағы.

**ОПРЕДЕЛЕНИЕ ЭФФЕКТИВНЫХ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ СЕНТИМЕНТАЛЬНОГО АНАЛИЗА КОММЕНТАРИЕВ НА КАЗАХСКОМ ЯЗЫКЕ**

**1Н.К. Мукажанов🖂, 2Л.Ш. Черикбаева, 1А.М. Касенхан, 1Ж.М. Алибиева,**

**1,3М. Тұрдалыұлы**

1Satbayev University, Алматы, Казахстан,

2Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан,

3Международный инженерно-технологический университет,

e-mail: [n.mukazhanov@satbayev.university](mailto:n.mukazhanov@satbayev.university)

В данной статье представлены результаты анализа алгоритмов машинного обучения для сентиментального анализа данных на казахском языке, и в результате анализа определены эффективные алгоритмы. В связи с увеличением объема казахскоязычного контента в социальных сетях, новостях и интернет-магазинах также возросла потребность в инструментах и методах обработки данных на казахском языке для получения ценной информации о мнениях и взглядах людей. Поэтому набор данных, использованный в исследовании, был собран из реальных интернет-магазинов и новостных сайтов. Объем собранного набора данных составляет 1500 записей, 80% из которых использовались для обучения алгоритмов, а 20% — для тестирования. Для сентиментального анализа данных рассмотрены алгоритмы машинного обучения, такие как логистическая регрессия, мультиномиальный наивный байесовский метод, метод опорных векторов (SVM), XGBoost и длинная краткосрочная память (LSTM) глубокого обучения. В ходе исследования тестировались алгоритмы с увеличением набора данных с 500 до 1500 записей, а также были реализованы и протестированы различные методы, такие как индивидуальный, ансамблевый и расширенный. Результаты, полученные в ходе тестирования, были представлены по показателям точности алгоритмов.

**Ключевые слова*:*** сентиментальный анализ, машинное обучение, глубокое обучение, NLP, комментарий, набор данных.

**Introduction.** Sentimental analysis is a subfield of Natural Language Processing (NLP) that aims to identify and analyze the moods (feelings) expressed in a given text. It is a powerful tool that we can use to understand people's opinions, attitudes, and emotions about a particular topic, brand, product, or service. Sentimental analysis can be applied to many different areas, such as market research, social media monitoring, customer service, and public opinion analysis [1]. Sentiment analysis is a powerful tool for companies to gain insight into customer sentiment and opinions, track their brand reputation, conduct market research, conduct competitive analysis, manage risk, make data-driven decisions, and improve their products, services, and customer interactions.

# The study of the model and methods of analyzing sentimental data in the Kazakh language is very important for a number of reasons:

- due to the increase in the amount of data available on the Internet, many business organizations can gain valuable information from large volumes of data, such as understanding customer preferences, market trends, and developing effective marketing strategies [2];

- sentimental analysis can be used for research purposes in areas such as sociology, psychology, and political science. By analyzing the data mentioned in social media posts, news, articles or other textual data, researchers can obtain information about the views, emotions and beliefs of various groups or individuals in Kazakh society;

- sentimental data analysis can also be applied in a variety of fields, including health, finance, and education. For example, it can be used to analyze patient feedback and improve health care services, detect fraudulent financial transactions, and evaluate the effectiveness of educational programs.

With the increasing use of digital content in the Kazakh language, such as social media posts, news and customer reviews, there is a growing need for automated sentiment analysis techniques which can process large amounts of data and provide valuable information about people's opinions and attitudes. This work is an experimental contribution to the analysis of Kazakh language texts and sentimental research. Different methods and their accuracies are compared using texts based on opinion data.

The purpose of the presented article is to compare machine learning and deep learning methods in the study of sentiment analysis and to provide specific models and methods **that** yield good results in the sentiment analysis of opinions in the Kazakh language.

In general, we believe that the study of models and methods of sentimental data analysis in the Kazakh language is relevant from both a theoretical and a practical point of view. It is necessary to better understand and use the huge amount of data available in Kazakhstan.

**Materials and methods.** Extensive research has been conducted on sentiment analysis of data. This is evidenced by our search on the topic of sentimental analysis in the sciencedirect (<https://www.sciencedirect.com/>) database of scientific works, which revealed a total of 77,525 scientific works, including 7171 scientific articles and textbooks published in 2022 and 8194 in 2023.

The vast majority of research on sentiment data analysis is devoted to the English language [3, 4], also there are also many studies on the widely used Arabic [5], Chinese [6], and Russian [7] languages. These languages differ from the Kazakh language in terms of word formation, language syntax, and morphology. In addition, each natural language needs its own dataset in machine learning and deep learning methods. Therefore, in order to analyze the Kazakh language data, first of all, a data set should be collected and the analysis methods should be tested with the collected data. Several numbers of works [8, 9, 10] have been published on the sentiment analysis of Kazakh texts. In [8] considered two ways of implementing transfer learning at work: null learning and fine-tuning. Experiments were designed to compare these two methods and report the results obtained. In the work, the Bidirectional Encoder Representations (BERT) model obtained from transformers suggested better results can be achieved in the Kazakh language with less resurgence. [9] this work is devoted to the development and application of an information system called the OMS system for the analysis of user comments on news portals, blogs and social networks in Kazakhstan. The system used sentimental dictionaries of the Russian and Kazakh languages and machine learning algorithms to determine the mood of texts in social networks. The article focused more on building a system than on studying methods of sentimental data analysis. The authors of the article [10] provided for an immediate analysis of Kazakh and Russian-speaking opinions. The work describes modern approaches to solving the problem of analyzing the opinions of news articles in Kazakh and Russian languages using neural networks. Thus, studies have shown that it is possible to achieve good results without knowing the linguistic characteristics of a particular language. The accuracy of the methods proposed in this work is 73%.

*Methods of sentimental data analysis* can be divided into several groups: Lexicon-based techniques, machine learning methods, deep learning methods, and hybrid methods. The main methods we consider are machine learning and deep learning.

*Lexicon-based techniques.* Lexicon-based methods use pre-built dictionaries or lexicons containing words and their corresponding sentiment scores [11]. For example, the word "good" has a positive meaning, and the word "hate" has a negative meaning. The sentiment score of the text is then calculated based on the sum or average of the scores of its component words. Lexical-based methods are relatively simple and fast, but may not be suitable for capturing the specific features and complexities of natural language. Dictionary-based methods are a popular approach to sentiment analysis, and one of the most commonly used lexicons is AFINN (Affective Norms for English Words). AFINN is a list of pre-computed sentiment scores for English words ranging from -5 (negative) to +5 (positive) [12].

*Methods based on machine learning.* Machine learning-based methods rely on algorithms that learn to identify sentiment patterns from large datasets of annotated texts. Algorithms are trained on labeled datasets, where each text is annotated with a corresponding sentiment label (positive, negative, or neutral). Machine learning-based methods are more flexible and accurate than dictionary-based methods, but they require more labeled data and may not generalize well to unseen data. Logistic regression, Naive Bayes, Decision Tree [10], and support vector machine (SVM) methods are often used in sentiment analysis of data [13, 14].

*Methods based on deep learning.* Deep learning techniques are one of the best performing methods in sentiment analysis. They can analyze the complex structures of texts and describe nuances of mood and sensitive parts of context. Deep learning-based methods require more data than machine learning-based methods, but they can achieve superior results in sentiment analysis tasks [15]. Among deep learning approaches, circuit models and recurrent neural networks are best suited for analyzing textual data. Sequence modeling plays a key role in sentiment analysis, as it allows for sentiment analysis in text, word order, and context. Sentiment analysis aims to identify the underlying sentiment or emotion expressed in a piece of text, such as a sentence, review, or social media post. Using sequense modeling techniques, sentiment analysis can capture the complex relationships between words and accurately explain the sentiment conveyed by text.

Recurrent Neural Network (RNN) is a type of neural network architecture specially developed for systematic data processing. This is useful for sequence processing tasks such as time series analysis, natural language processing, speech recognition, and handwriting recognition. Unlike feedforward neural networks, which process data in a single pass from input to output, RNNs have a feedback mechanism which allows them to store information about previous inputs and use it to influence the processing of future inputs. This feedback mechanism allows RNNs to capture temporal dependencies and patterns present in serial data [16].

# Results and discussion. Sentimental analysis is carried out by processing the data in several stages. These are: collecting current opinions, pre-processing and preparing data sets, analyzing (reviewing) processed data sets, extracting labels from processed data, analyzing and obtaining results using machine and deep learning methods, and evaluating results.

# *Gathering relevant data.* Creating a data set consisting of different opinions is the main and initial step in the research of sentiment analysis model and methods. One way to build an opinion database is to collect opinions from various online sources, such as e-commerce websites, social media platforms, and other websites. To ensure the quality of the data set, the text must be accurate and consistent. Reviews used in this work were collected from e-commerce websites, including Kaspi magazin, nur.kz and 2GIS. All non-Kazakh comments were removed from the dataset to ensure comments were in the Kazakh language. Once the comments are collected, they should be classified into positive and negative sentiment (sentiment) based on the general opinion expressed in the text. In our case, a total of 1500 comments in the dataset were manually marked as positive, negative or neutral. The final data set was divided into two parts: the training set and the test set. The training set consists of 1200 comments, and the test set consists of 300 comments. The training set was used to train the sentiment analysis model, and the test set was used to evaluate the accuracy of the model.

# *Data set preprocessing and preparation.* Text preprocessing and preparation involves cleaning, transforming, and organizing textual data to facilitate analysis. The main purpose of text preprocessing is to convert raw text into a format suitable for analysis by removing unnecessary information, converting text to common case, and reducing data sizes. Libraries such as NLTK (Natural Language Toolkit), spaCy and Scikit-learn are proposed for processing the collected data. In our work, using the NLTK library, tokenization, erasure of stop words, transformations were carried out.

After removing duplicate data, correcting misspelled words, and performing the transformation mentioned above, we can have a visual overview of the dataset. Different types of visualization were used to review the sentiment data beforehand. By visualizing sentiment data, quantitative and visual analyses were conducted to clearly see the positive and negative words that are often used. Figures 1, 2, 3 show overviews of the preprocessed data. Figure 1 shows a cloud of positive comments, while Figure 2 shows a cloud of negative comments.

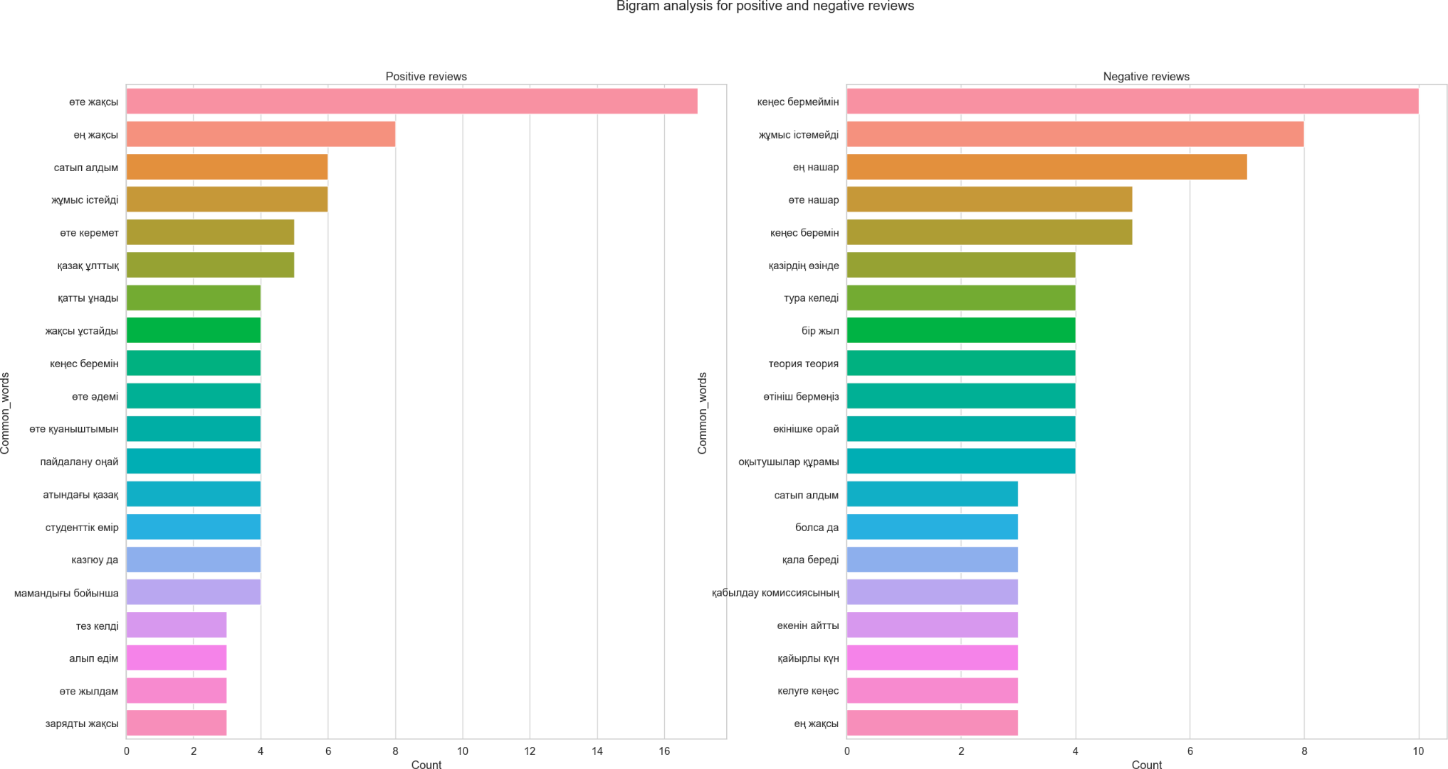


**Figure 1 - Word cloud of positive comments**



**Figure 2 - Word cloud of negative comments**

Figure 3 shows the bigram analysis of the comments. From this analysis, we can see exactly which phrases generate positive and negative feedback. In our case, "excellent", "best", "bought", "liked" and so on. bigrams were most often used in positive comments, while bigrams such as "do not recommend", "doesn't work", "worst" were most often used in describing negative comments.



**Figure 3 - Bigram analysis for negative and positive reviews**

*Feature extraction from the processed data.* Feature extraction is an important step involving the conversion of text data into vectors that can be understood by machine learning algorithms.The following methods are used to extract symbols from text data: one-hot encoding, bag of words, term frequency - inverse document frequency (TF-IDF) and word embeddings. In the work , term frequency - inverse document frequency (TF-IDF) vectorizer was used to extract features. The TF-IDF vectorizer is the most commonly used method for converting texts to digital values. TF-IDF is a numerical statistic which indicates the importance of a word for a document in a collection or corpus. It is calculated as a product of two components: Term Frequency (TF) and Inverse Document Frequency (IDF) [17].

A matrix of numeric vectors was created by applying the TF-IDF formula on the corpus of documents. Each row of the matrix represents a document, and each column represents a unique word in the corpus. The resulting matrix was fed as input to various machine learning algorithms for sentiment analysis. The TF-IDF vectorizer is a powerful tool for feature extraction from textual data in sentiment analysis because it shows both the importance of words in a document and their rarity in the corpus. This helps improve the accuracy and efficiency of machine learning models for sentiment analysis.

*Implementation of sentimental data analysis with machine learning algorithms.* Conducting sentimental analysis of reviews consists of several steps. The sequence of steps is shown in Figure 4. The first analyzed data is pre-processed. Then, the textual data in the training and test sets were transformed into numerical feature vectors based on the TF-IDF representation. This transformation allows machine learning algorithms to work efficiently with textual data, as they typically require numerical input.

Text, reviews

Feature extraction

TF-IDF

ML models

Logistic regression

SVM

XGBoost

Multi.Naive Bayes

LSTM

Preprocessing

token

Stop words

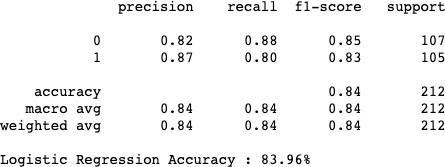
pos.

neg.

Sentiments

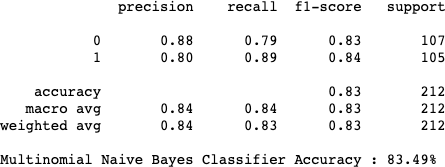
**Figure 4 - Steps of processing text data using machine learning**

The data set was divided into training and test sets using the train\_test\_split function of the scikit-learn library. The Test\_size parameter was set to 0.2, that is, 20% of this data was used for testing, and 80% was used for training. Using the converted TF-IDF data, a version of the LogisticRegression class was developed, which was displayed by the logistic regression algorithm. As the logistic regression algorithm and target labels are trained through the training data, the identification method adjusts the sample parameters to find the best fit for the given data. The accuracy, classification calculation, and confusion matrix for estimating sample performance are as follows (Figure 5):



**Figure 5 - Accuracy of logistic regression algorithm**

And to use the Multinomial Naive Bayes model, the sklearn.naive\_bayes function was used. A Multinomial Naive Bayes classifier was instantiated and trained on the training data represented by the TF-IDF vectorized feature matrix and the corresponding feature matrix. The features of the test data were predicted using the trained MNB classifier. The accuracy score was calculated by comparing the predicted signs with the real signs (accuracy indicators are shown in Figure 6).



**Figure 6-Accuracy of Multinomial Naive Bayes model**

In addition, Linear SVM and XGboost models of machine learning are also tested. Their results are lower than logistic regression (Linear SVM-79.25%, Xgboost-80.1%).

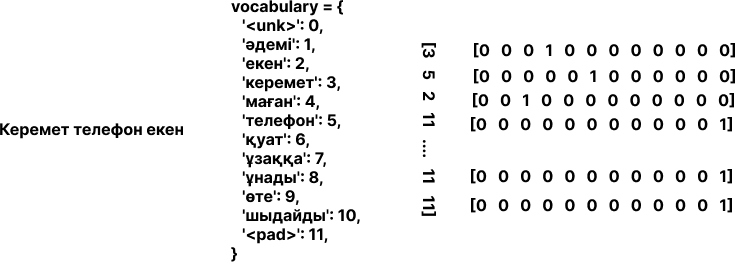
*Implementation of sentimental data analysis with deep learning algorithms.* In order to use the Recurrent Neural Network (RNN) method, it is necessary to build a vocabulary. This is an efficient lookup table where each unique word in the data set has a corresponding index (integer). Based on the reviews, the following dictionary was compiled:

*[«Керемет телефон екен», «Маған ұнады әдемі», «Қуаты өте ұзаққа шыдайды»] → vocabulary {«<unk>»:0, «әдемі»:1, «екен»:2, «керемет»:3, «маған»:4, «телефон»:5, «қуат»: 6, «ұзаққа»:7, «ұнады»:8, «өте»:9, «шыдайды»:10, «pad»:11}*

Then, using a dictionary, we assigned indexes to the text data in the test set, replacing each word in the data with the corresponding index in the dictionary. In this step, the textual data is converted into numerical representations that can be used as input to the RNN. The following conversion is obtained:

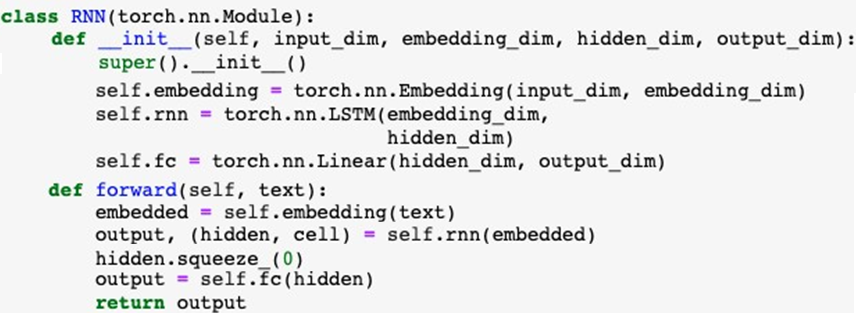
*[«Керемет телефон екен», «Маған ұнады әдемі», «Қуаты өте ұзаққа шыдайды»] → vocabulary {«<unk>»:0, «әдемі»:1, «екен»:2, «керемет»:3, «маған»:4, «телефон»:5, «қуат»: 6, «ұзаққа»:7, «ұнады»:8, «өте»:9, «шыдайды»:10, «pad»:11} → {{«Керемет телефон екен»: [3 5 2 11]}, {«Маған ұнады әдімі»: [4 8 1 11]}, {«Қуат өте ұзаққа шыдайды»: [6 9 7 10]}}*

Each index is used to create a vector for each word. One-hot vector is a vector whose size is the total number of unique words in the dictionary, where only one element is 1, and all other elements are 0 (Figure 7). Tourch libbrary was used for this task.



**Figure 7-One-hot vector**

An embedding layer was used to transform a one-hot vector into a dense embedding vector. The model constructor consists of the size of the input dictionary, the number of unique tokens in the textual data (input\_dim), the dimensionality of the embedding vectors representing the input tokens (embedding\_dim), the number of units in the hidden state of the LSTM (hidden\_dim), the dimensionality of the output representing the sentiment prediction (output\_dim). The implementation of the model is presented in Figure 8.



**Figure 8 - RNN model creation**

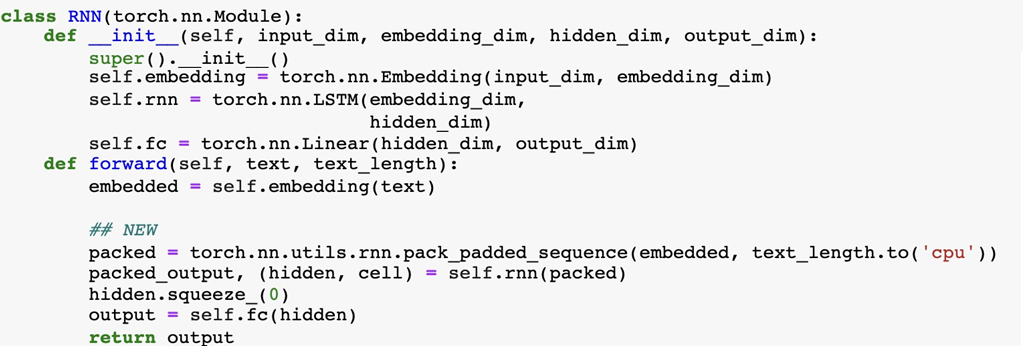
In addition, the torch.nn.Embedding module was created, which is responsible for learning and matching input tokens with dense embedding vectors. And the torch.nn.LSTM module was built to implement the long LSTM layer. This module takes embedded input as input and processes the serial information to generate hidden states. Next, a torch.nn.Linear module was created which maps the last hidden state to the output dimension responsible for sentiment prediction.

The Forward method is responsible for the forward movement of the model. It accepts text as input text data. An embedded tensor is obtained by passing the input text through an embedding layer, performing a search to transform the input tokens into dense vector representations. The embedded tensor is then passed through the RNN layer and returns the (hidden, cell) values containing the output features of the LSTM for each time step and the final hidden state of the LSTM. The latent tensor is filtered along the first dimension to remove the extra dimension added by the LSTM.

Next, the created model was trained. This process consisted of 10 stages. Logs the training progress, updating model parameters based on calculated gradients. It also evaluates the performance of the model on the training and validation sets after each phase. Finally, 60% accuracy was obtained from model training. 60% is not considered a good result, so ways to improve its accuracy rate were considered.

To increase the accuracy of the model, various modifications were made to the LSTM algorithm, including the embedded use of the packed padded sequences method, which gave good results. Important aspects of packed padded sequences are defined (example implementation code is shown in Figure 9):

* Efficient computation: Stacked filled sequences allow for more efficient computation during training and inference. In RNNs, input sequences are usually processed in parallel in a small batch. Filled circuits avoid unnecessary computations on filled elements, resulting in faster training and reduced inference time.
* Memory efficiency: padded circuits increase memory requirements because they introduce a significant number of padded elements. By wrapping strings, filled elements are effectively masked, reducing the memory footprint and optimizing memory usage.
* Improved model performance: Overloaded circuits can negatively impact model performance by introducing noise and unnecessary calculations.



**Figure 9 - Improved LSTM algorithm code**

This function creates a boxed sequence object which represents sequences without padding tokens. Using this method, 84% accuracy was achieved.

*Discussion of results.*Let's look at which method gives better results by comparing the results of sentimental analysis made using Logistic Regression, Multinomial Naive Bayes, Liner SVM, XGBoost, Long short-term memory (LSTM), LSTM-pack sequence learning methods to improve LSTM on collected Kazakh comments. Table 1 shows the results of the above-mentioned models and algorithms evaluation experiment

**Table 1 - Results of algorithm**

|  |  |  |
| --- | --- | --- |
| № | Algorithm name | Accuracy |
| 1 | Logistic regression | 83.96% |
| 2 | Multinomial Naive Bayes | 83.49% |
| 3 | Linear SVM | 79,25% |
| 4 | XGboost | 80,1% |
| 5 | LSTM | 60.00% |
| 6 | Improved LSTM | 84.12% |

Due to the lack of a large collection of data, machine learning methods have shown that they do not give more accurate results. A logistic regression algorithm can be effective if there is low-dimensional data and their capability is linearly distributed, but a large data set is required to achieve good results. Although the LSTM model showed only 60 percent accuracy, it increased its accuracy using the packed packed sequences method and achieved better results than everyone else. The LSTM model has shown poor performance on very little data. When we increased the number of reviews from 500 to 1500, we noticed a improvement in the result.

**Conclusion.** This paper is devoted to the study of ways of sentimental analysis of Kazakh language reviews. The introduction of the article explains sentimental analysis and gives citations for its importance. The paper considered lexicon (rule) based methods, machine learning based methods and deep learning methods as methods of sentimental data analysis. Among them, we differentiated the possibilities of determining the mood in Kazakh-language opinions using methods based on machine learning and deep learning methods, which are currently being extensively researched and are well used in practical developments.

Kazakh-language reviews used in sentimental analysis were collected from Kaspi online-market, nur.kz and 2GIS. Collected opinions were subjected to initial processing such as preliminary processing and preparation of the data set, analysis (review) of the processed data set, extraction of features from the processed data. Further, the prepared data were sentimentally analyzed using Logistic regression, Multinomial Naive Bayes, Linear SVM, Xgboost and deep learning LSTM models of machine learning and the accuracy indicators of the obtained results were determined. After all samples were analyzed, a comparative analysis of their accuracy indicators was performed. Comparative analysis showed that the best result was the deep learning improved LSTM model (accuracy-84.12%), followed by the machine learning Logistic regression model (accuracy-83.96%).

**References**

1. Bolshakova E.I., Voroncov K.V., Efremova N.E., Klyshinskij E.S., Lukashevich N.V., Sapin A.S. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh. - M.: Izd-vo NIU VShE, 2017. - 269 s. [in Russian]

2. Pasolini, Roberto. Learning Methods and Algorithms for Semantic Text Classification across Multiple Domains, [Dissertation thesis]. Alma Mater Studiorum Università di Bologna. Dottorato di ricerca in Ingegneria elettronica, informatica delle telecomunicazioni, 27 Ciclo. DOI 10.6092/unibo/amsdottorato/7058.

3. Jingli Shi, Weihua Li, Quan Bai, Yi Yang, Jianhua Jiang. Syntax-Enhanced Aspect-Based Sentiment Analysis with Multi-Layer Attention // Neurocomputing. -2023.

DOI [10.1016/j.neucom.2023.126730](https://doi.org/10.1016/j.neucom.2023.126730)

4. Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, G. Manikandan. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models // Alexandria Engineering Journal. – 2023. –Vol. 80(2). - P. 110-121. DOI[10.1016/j.aej.2023.08.038](http://dx.doi.org/10.1016/j.aej.2023.08.038)

5. Arwa A. Al Shamsi, Sherief Abdallah. Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects // Journal of King Saud University. -2023. -Vol. 35(8).

[DOI 10.1016/j.jksuci.2023.101691](https://doi.org/10.1016/j.jksuci.2023.101691)

6. Meng Li, Yucheng Shi. Sentiment analysis and prediction model based on Chinese government affairs microblogs // Heliyon. - 2023. DOI [10.1016/j.heliyon.2023.e19091](https://doi.org/10.1016/j.heliyon.2023.e19091)

7. Chetviorkin, I.; Loukachevitch, N. Evaluating Sentiment Analysis Systems in Russian // In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8–9 August 2013; Association for Computational Linguistics . - Sofia, Bulgaria, 2013. -P. 12–17.

8. Nugumanova A., Baiburin Ye., Alimzhanov Ye.. Sentiment Analysis of Reviews in Kazakh With Transfer Learning Techniques // SIST 2022 - 2022 International Conference on Smart Information Systems and Technologies, Proceedings, ISBN 978-166546790. - 2022.

DOI 10.1109/SIST54437.2022.9945811

9. Karyukin Vladislav, Mutanov Galimkair, Mamykova Zhanl, Nassimova Gulnar, Torekul Saule, Sundetova Zhanerke, Negri Matteo. On the development of an information system for monitoring user opinion and its role for the public // Journal of Big Data. -2022. - Vol. 9.

DOI 10.1186/s40537-022-00660-w

10. Narynov S.S., Zharmagambetov A.S.: On One Approach of Solving Sentiment Analyis Task for Kazakh and Russian Languages Using Deep Learning // International Conference on Computational Collective Intelligence. - Springer, 2016. - Vol. 9876.

DOI 10.1007/978-3-319-45246-3\_51

11. Saraswathi N., Sasirooba T., Chakaravarthi S. Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews // Measurement: Sensors. - 2023. - Vol. 29. - DOI [10.1016/j.measen.2023.100888](https://doi.org/10.1016/j.measen.2023.100888)

12. Gulmira Bekmanova, Gaziza Yelibayeva, Saltanat Aubakirova, Nurgul Dyussupova, Altynbek Sharipbay, Rozamgul Nyazova. Methods for analyzing polarity of the Kazakh texts related to the terrorist threats // Computational Science and Its Applications - ICCSA 2019. - P. 717-730.

DOI 10.1007/978-3-030-24289-3\_53

13. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R // Springer Science and Business Media. - New York, 2013.

DOI 10.1007/978-1-4614-7138-7

14. Dinara Gimadi, Richard Evans, Kiril Simov. Web-sentiment Analysis Of Public Comments (Public Reviews) For Languages With Limited Resources Such As The Kazakh Language. - 2021. DOI 10.26615/issn.2603-2821.2021\_010

15. Huseyn Hasanli , Burak Ordin , Samir Rustamov. Sentiment analysis on twitter data for azerbaijani language // Journal of Modern Technology and Engineering.-2019. - Vol.4. - No.2.   
- P.109-121. 2019.

16. Sebastian Raschka, Vahid Mirjalili. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. - 2019. P. 770.

17. Ugo Erra, Sabrina Senatore, Fernando Minnella, Giuseppe Caggianese. Approximate TF–IDF based on topic extraction from massive message stream using the GPU // Information Sciences, January. - 2015. –Vol. 292. -P. 143-161. DOI10.1016/j.ins.2014.08.062

***Information about the authors***

Mukazhanov N. K. - doctor Ph.D., associate professor, Satbayev University, Almaty, Kazakhstan, е-mail: [n.mukazhanov@satbayev.university](mailto:n.mukazhanov@satbayev.university);

Cherikbayeva L.Sh. - doctor Ph.D., Al-Farabi Kazakh National University, Almaty, Kazakhstan, е-mail: [cherikbayeva.lyailya@gmail.com](mailto:cherikbayeva.lyailya@gmail.com);

Kasenkhan A.M. – doctor Ph.D., Satbayev University, Almaty, Kazakhstan, e-mail: [a.kassenkhan@satbayev.university](mailto:a.kassenkhan@satbayev.university);

Alibieva Zh.M. - doctor Ph.D., associate professor, Satbayev University, Almaty, Kazakhstan, е-mail: [zh.alibiyeva@satbayev.university](mailto:zh.alibiyeva@satbayev.university);

Turdalyuly M. - doctor Ph.D., associate professor, International Engineering and Technological University, Almaty, Kazakhstan, е-mail: m.turdalyuly@gmail.com

***Сведения об авторах***

Мукажанов Н.К. - доктор Ph.D, ассоциированный профессор, Satbayev University, Алматы, Казахстан, е-mail: [n.mukazhanov@satbayev.university](mailto:n.mukazhanov@satbayev.university);

Черикбаева Л.Ш. - доктор Ph.D, Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан, е-mail: [cherikbayeva.lyailya@gmail.com](mailto:cherikbayeva.lyailya@gmail.com);

Қасенхан А.М - доктор Ph.D, Satbayev University, Алматы, Казахстан, e-mail: [a.kassenkhan@satbayev.university](mailto:a.kassenkhan@satbayev.university);

Алибиева Ж.М. - доктор Ph.D, ассоциированный профессор, Satbayev University, Алматы, Казахстан, е-mail: [zh.alibiyeva@satbayev.university](mailto:zh.alibiyeva@satbayev.university);

Тұрдалыұлы М. - доктор Ph.D, ассоциированный профессор, Международный инженерно-технологический университет, Алматы, Казахстан, е-mail: m.turdalyuly@gmail.com

IRSTI 28.23.29

**THERMAL COMFORT PREDICTION USING SVM AND RANDOM FOREST MODEL**

**N.B Assymkhan🖂, A.Z.Kartbaev**

Kazakh-British Technical University, Almaty, Kazakhstan

**🖂**Correspondent-author anb.asymhan@gmail.com

Predicting thermal comfort is crucial for optimizing built environments for human habitation, as it impacts health, productivity, and overall well-being. To address this imperative, interdisciplinary collaboration among architects, engineers, psychologists, and data scientists is needed to develop reliable predictive models that anticipate occupants’ thermal comfort preferences across diverse environmental conditions and architectural designs. Traditional methods rely on human comfort models, which can be subjective and time-consuming. Machine learning algorithms, such as Support Vector Machines (SVM) and Random Forest (RF), have been utilized to predict thermal comfort with high accuracy and efficiency. The Internet of Things (IoT) is revolutionizing the building management systems industry, with adaptive control algorithms and modular architectures exploring the IoT paradigm. This paper discusses the use of SVM and Random Forest algorithms for predicting thermal comfort in buildings, exploring their strengths and weaknesses and comparing their performance in different scenarios. The study analyzed a dataset of thermal comfort data, filtering by quantity and removing outliers. The data was split into 80% for training and 20% for testing. The study used SVM and Random Forest models to capture complex relationships between environmental parameters and thermal comfort responses. The results showed that the IQR method provided 3-4% accuracy, while the reducing label values method offered 20-23% accuracy. The study also tested the parameters of the models, resulting in a 2-4% difference between the two models. The study concluded that Random Forest appears more stable than SVM and plans to add new features to improve accuracy.

**Keywords:** Heating, ventilation, and air conditioning, temperature, thermal comfort, support vector machine (SVM), random forest (RF).

**ПРОГНОЗИРОВАНИЕ ТЕПЛОВОГО КОМФОРТА С ПОМОЩЬЮ МОДЕЛИ SVM И RANDOM FOREST**

**Н.Б.Асымхан🖂, А.З.Картбаев**

Казахстанско-Британский Технический Университет,Алматы, Казахстан,

e-mail: anb.asymhan@gmail.com

Прогнозирование теплового комфорта имеет решающее значение для оптимизации строительных сред для человеческого проживания, так как оно влияет на здоровье, производительность и общее благополучие. Для решения этой задачи требуется междисциплинарное сотрудничество архитекторов, инженеров, психологов и специалистов по обработке данных для разработки надежных прогностических моделей, предвидящих предпочтения жильцов к тепловому комфорту в различных климатических условиях и архитектурных решениях. Традиционные методы основаны на моделях комфорта человека, которые могут быть субъективными и затратными по времени. Алгоритмы машинного обучения, такие как Support Vector Machine (SVM) и Random Forest (RF), использовались для прогнозирования теплового комфорта с высокой точностью и эффективностью. Интернет вещей (IoT) революционизирует отрасль систем управления зданиями, с адаптивными управляющими алгоритмами и модульными архитектурами, исследующими парадигму IoT. В данной статье обсуждается использование алгоритмов SVM и Random Forest для прогнозирования теплового комфорта в зданиях, исследуются их преимущества и недостатки, а также сравнивается их производительность в различных сценариях. В рамках исследования был проанализирован набор данных по тепловому комфорту, произведена фильтрация по количеству и удаление выбросов. Данные были разделены на 80% для обучения и 20% для тестирования. В исследовании использовались модели SVM и Random Forest для выявления сложных взаимосвязей между параметрами окружающей среды и реакциями на тепловой комфорт. Результаты показали, что метод IQR обеспечил точность в 3-4%, в то время как метод уменьшения значений меток предоставил точность в 20-23%. Также были проверены параметры моделей, что привело к различию в 2-4% между двумя моделями. Исследование заключает, что Random Forest оказался более устойчивым, чем SVM, и планирует добавить новые функции для повышения точности.

**Ключевые слова:** отопление, вентиляция и кондиционирование воздуха, температура, тепловой комфорт, Support Vector Machine (SVM), Random Forest (RF).

**SVM ЖӘНЕ RANDOM FOREST МОДЕЛДЕРІН ПАЙДАЛАНУ АРҚЫЛЫ**

**ЖЫЛУЛЫҚ ЖАЙЛЫЛЫҚТЫ БОЛЖАУ**

**Н.Б. Асымхан🖂, А.З.Картбаев**

Қазақстан-Британ Техникалық Университеті, Алматы, Қазақстан,

e-mail: anb.asymhan@gmail.com

Жылулық жайлылықты болжау адам тұруы үшін салынған ортаны оңтайландыру үшін өте маңызды, өйткені ол денсаулыққа, өнімділікке және жалпы әл-ауқатқа әсер етеді. Бұл мәселені шешу сәулетшілер, инженерлер, психологтар және деректер ғалымдары арасындағы әртүрлі климаттық және сәулеттік дизайндағы тұрғынның жылулық жайлылық қалауларын болжайтын сенімді болжамды модельдерді әзірлеу үшін пәнаралық ынтымақтастықты талап етеді. Дәстүрлі әдістер субъективті және уақытты қажет ететін адам жайлылық үлгілеріне сүйенеді. Support Vector Machine (SVM) және Random Forest (RF) сияқты машиналық оқыту алгоритмдері жоғары дәлдік пен тиімділікпен термиялық жайлылықты болжау үшін пайдаланылды. Заттардың интернеті (IoT) парадигмасын зерттейтін адаптивті басқару алгоритмдері мен модульдік архитектуралары арқылы ғимараттарды басқару жүйелерінің индустриясында төңкеріс жасайды. Бұл мақалада ғимараттардағы жылулық жайлылықты болжау үшін SVM және Random Forest алгоритмдерін пайдалану талқыланады, олардың артықшылықтары мен кемшіліктері зерттеледі және әртүрлі сценарийлердегі олардың өнімділігі салыстырылады. Зерттеудің бір бөлігі ретінде термиялық жайлылық деректерінің жиынтығы талданды, саны бойынша сүзілді және шектен тыс мәндер жойылды. Деректер оқу үшін 80% және тестілеу үшін 20% бөлінді. Зерттеу қоршаған орта параметрлері мен жылулық жайлылық реакциялары арасындағы күрделі қатынастарды анықтау үшін SVM және Random Forest үлгілерін пайдаланды. Нәтижелер көрсеткендей, IQR әдісі 3-4% дәлдік береді, ал жапсырманы азайту әдісі 20-23% дәлдік береді. Модельдердің параметрлері де тексерілді, нәтижесінде екі модель арасында 2-4% айырмашылық болды. Зерттеу Random Forest SVM-ге қарағанда сенімдірек екенін дәлелдеді және дәлдікті жақсарту үшін жаңа мүмкіндіктерді қосуды жоспарлап отырмыз.

**Tүйін сөздер:** жылыту, желдету және ауаны баптау, температура, термиялық жайлылық, Support Vector Machine (SVM), Random Forest (RF).

**Introduction.** In the pursuit of optimizing built environments for human habitation, predicting thermal comfort emerges as a pivotal challenge. With climate change intensifying, the frequency and severity of extreme weather events are increasing, amplifying the significance of understanding and managing indoor thermal conditions. The need to predict thermal comfort stems from its profound impact on human health, productivity, and overall well-being. Inadequate thermal conditions, whether excessive heat or cold, can lead to discomfort, fatigue, and even health complications, thereby compromising individuals’ quality of life and impeding productivity in various settings, including workplaces, educational institutions, and residential spaces.

Furthermore, the economic implications of disregarding thermal comfort cannot be overlooked. Suboptimal indoor climates contribute to increased energy consumption as occupants resort to heating or cooling systems to mitigate discomfort, resulting in inflated utility bills and environmental repercussions. Hence, there is a pressing need to develop reliable predictive models that anticipate occupants’ thermal comfort preferences across diverse environmental conditions and architectural designs. These models should consider factors such as ambient temperature, humidity levels, clothing insulation, metabolic rates, and individual preferences to furnish accurate assessments of thermal comfort levels. Addressing this imperative requires interdisciplinary collaboration among architects, engineers, psychologists, and data scientists to integrate knowledge from environmental science, human physiology, and behavioral psychology. By leveraging advancements in sensor technology, data analytics, and machine learning algorithms, predictive models can be refined to offer real-time insights into thermal comfort dynamics, empowering building managers and occupants to optimize indoor environments for enhanced well-being and sustainable resource utilization. Let’s look at how this all affects in more detail and with an example. Many people know that temperature is a very important factor for a person, when you start to get sick, the temperature of your blood rises and this gives you a signal that you have been poisoned or caught a cold. In a word, it signals that something has gone wrong in your body. Now how does the room temperature affect and why do we need a comfortable temperature? For example, consider a summer day when you start preparing for lessons or studying something, you close the door of your room so that the noise does not interfere with your studies and close the window because it is hot outside. But, here the opposite effect occurs, since you closed the door, you reduced the area of the room and the speed of airflow into your room. Further, carbon dioxide will be released, which will fill the room, thereby reducing the oxygen in the room and increasing the temperature of the room. Thus, you become a little distracted and lethargic. You can correct the situation by opening the door of the room. Also, when you are late for a lesson or a meeting or an exam, you will release a stress hormone that will increase your body temperature and your heart rate, given that not only you are sitting in the exam, but about 40 people and everyone has an increased level of stress and this affects the fact that oxygen is quickly absorbed and replaced by carbon dioxide. This will heat up the temperature of the classrooms and reduce the efficiency level of the students inside. And therefore, usually at the beginning of the exam, some questions are not clear, then as the stress level decreases, then clarity of mind opens up. Usually, the door is opened for this because it has become hot, but there are HVAC or NV systems for this, which sometimes turn off or they do not work correctly. Thus, if the comfortable temperature recognition system works correctly, then by choosing the temperature, you can reduce the level of stress that will be at the beginning of the exam, thereby increasing efficiency. If it’s cold in the classrooms, usually people fall asleep, you can notice it when you arrive early at 8 in the morning for the first lessons, this is because the human body feels cold and goes into an energy-saving mode like bears in hibernation. That’s why thermal comfort prediction is a crucial aspect of building design and management as it determines the satisfaction level of occupants in a given space. Predicting thermal comfort involves analyzing various factors such as temperature, humidity, air velocity, and clothing insulation. Traditional methods of predicting thermal comfort rely on human comfortmodels, which can be subjective and time-consuming. In recent years, machine learning algorithms, such as Support Vector Machines (SVM) and Random Forest (RF), have been utilized to predict thermal comfort with high accuracy and efficiency. SVM and RF are both supervised learning algorithms that can be trained on a dataset of thermal comfort parameters and their corresponding human feedback to accurately predict thermal comfort in new environments.

**Literature review.**The Internet of Things (IoT) is revolutionizing the building management systems industry, with the number of connected devices expected to reach 125 billion by 2030. However, the current BMS solutions are limited in flexibility, particularly in feedback control options. To fully harness the IoT paradigm, adaptive control algorithms and modular architectures have been explored.

The authors propose the ”Semantically-Enhanced IoT-enabled Intelligent Control System” (SEMIoTICS) architecture, which exploits redundancy in control system capabilities and automatically implements alternative configurations based on quality-of-service criteria [1]. A study introduces a novel model that excludes gender and age factors in thermal comfort assessment. The model considers six thermal factors: air temperature, mean radiant temperature, relative humidity, air speed, clothing insulation, and metabolic rate. The model is designed using Supervised Machine Learning in a commercial building [2]. A study in Bilbao, Spain, analyzes human thermal perception in response to external temperatures using KUBIK, an energy efficiency research facility, to improve indoor comfort and reduce energy consumption [3]. This study evaluates indoor thermal comfort using Fanger method and ASHRAE Standard 55, focusing on real-world conditions to maintain well-being, productivity, and energy conservation in buildings [4]. This study introduces a multiple preferencesbased model for predicting group thermal comfort in shared spaces, integrating individual preferences and environmental parameters. It segments occupants based on BMI, predicts individual comfort zones, and adjusts for group satisfaction [5]. Thermal comfort optimization in buildings is crucial for occupant well-being, productivity, and energy efficiency. Assessment involves models considering air temperature, humidity, radiant temperature, and speed. ASHRAE 55 standards define acceptable conditions. Alternative models like Artificial Neural Networks, hybrid ANN-fuzzy models, SVM, decision trees, fuzzy logic, and Bayes networks offer flexibility and accuracy [6]. Thermal comfort is a crucial aspect of indoor environmental quality, categorized into static, adaptive, and data-driven models. Static models like PMV, which integrate environmental and personal factors, have limitations. Adaptive models consider psychological and behavioral factors, while data-driven models use sensor technology for realtime assessments [7]. The authors develop a building thermal model using low-resolution data from smart thermostats, enhancing accuracy and applicability across seasons. They adapt traditional empirical models into a data-driven approach, using surrogate features to approximate heat gains. The model can be implemented on edge devices or cloud infrastructure, offering advantages in data collection, model learning, and deployment [8]. Research on indoor thermal comfort has focused on innovative cooling systems like Thermoelectric Air Duct. Neural network models have shown accuracy in predicting comfort parameters, especially in dynamic environments. The relationship between climatic variables, occupant comfort, and system performance is crucial [9]. Thermal comfort prediction and energy optimization in buildings are crucial for occupant satisfaction and energy efficiency. Factors influencing comfort include metabolic rate, clothing insulation, and air temperature. Deep feedforward neural networks and reinforcement learning models help predict comfort levels. Monitoring and optimizing HVAC energy consumption is essential for building operation [10]. The authors present a novel methodology using machine learning, data mining, and statistics to develop predictive models for Combined Heat, Cooling, and Power (CHCP) systems. The methodology includes four stages: data preparation, data engineering, model building, and model evaluation. Data preparation involves retrieving failure events, labeling instances, and creating a comprehensive dataset. Data engineering enhances data representation through feature extraction and feature selection. The model building uses machine learning algorithms for classification and regression tasks. Model evaluation considers time to failure (TTF) and performance metrics for suitable selection [11]. The study explores thermal comfort in indoor environments using a novel approach called Relative Thermal Sensation (RTS). The RTS considers thermal sensation as a continuous function of time, providing a more nuanced understanding of human thermal sensation. The authors propose a 3-point RTSS to gather real-time data on relative thermal sensation, capturing subtle changes in thermal perception that traditional discrete scales may not capture. The study also integrates RTS data with Absolute Thermal Sensation data from modified versions of the ASHRAE 7-point thermal sensation scale to develop a more comprehensive understanding of thermal comfort [12]. Interpretable thermal comfort systems are being explored to improve energy efficiency and occupant satisfaction in smart building environments. Traditional models like the Predicted Mean Vote (PMV) are often uninterpretable, making it difficult for building operators to understand the underlying mechanisms driving thermal comfort. Researchers have proposed interpretable thermal comfort systems using machine learning techniques like Partial Dependence Plots (PDP) and SHAP values. These techniques help operators understand the impact of environmental conditions on human comfort and the importance of different features under varying conditions. Additionally, interpretable ML algorithms can be used to develop surrogate models of existing comfort models [13].

In this paper, we will discuss the use of SVM and Random Forest algorithms for predicting thermal comfort in buildings. We will explore their strengths and weaknesses and compare their performance in different scenarios. The aim of this study is to provide a comprehensive understanding of the potential of these machine learning algorithms in predicting thermal comfort, which can help to build designers and facility managers optimize the indoor environment and improve the comfort of building occupants.

**Materials and Methods.** The primary objective of this study is to offer a comprehensive understanding of the potential of machine learning algorithms in predicting thermal comfort. This knowledge can be instrumental in assisting designers and facility managers in optimizing the indoor environment, ultimately enhancing the comfort of building occupants.

*Hypotheses:*

Before commencing our experiments, we have formulated the following hypotheses:

1) Data Preprocessing:

- It is essential to remove NaN values and set boundaries on the number of values in each column to ensure the selection of appropriate features.

- Utilizing the IQR (Interquartile Range) method for label value reduction to handle outliers effectively.

2) Encoder Selection:

- The choice of encoder, whether it be OneHotEncoder, LabelEncoder, or Word2Vec, will be critical in transforming categorical variables into a format suitable for machine learning algorithms.

3) Feature Selection with SelectKBest:

- Utilizing the SelectKBest model will assist us in identifying a list of features that are most relevant to the thermal comfort prediction.

4) Feature Filtering:

- After initial filtering, we will choose variants of the features that closely correlate with temperature predictions.



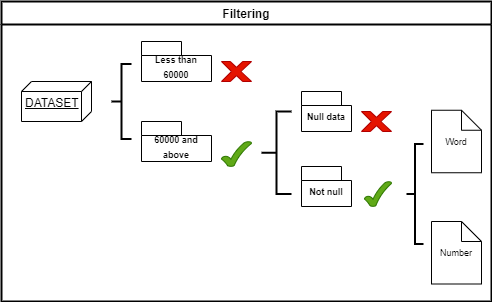
**Figure 1 – Steps**

1. *Dataset*

The data was taken from the Kaggle dataset which was taken from the ASHRAE dataset [14]. The data has 70 columns and 107583 rows.

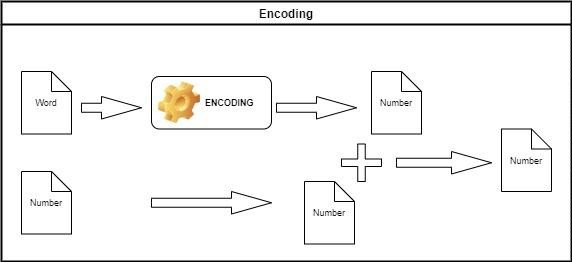
1. *Filtering data*

In the beginning, after looking at the description of the data, we do filtering, when viewing it, it turned out that some columns have little data. Because of this, filtering by quantity went on, and 60.000 lines were taken by the border. Below this boundary, all data was deleted, then it was necessary to remove the Nan value, some rows could remain empty because we had 107583 rows from the beginning. One more hypothesis to test, the idea is to use *IQR* (Inter Quartile Range) method to remove outliers if it is exist.



**Figure 2 - Filtering scheme**

1. *Encoding*



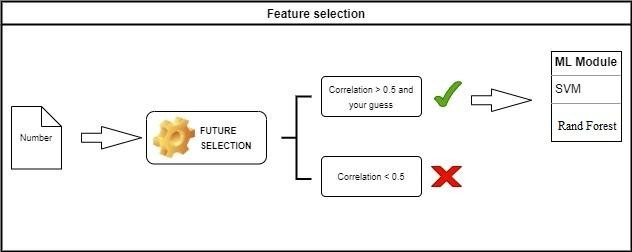
**Figure 3 - Encoding scheme**

When converting text to a number, there were two choices LabelEncoder or OneHotEncoder, the choice stopped at One- HotEncoder as it showed good results.

1. *Feature selection*

When choosing a feature, there were two ways to select using the SelectBest library or a correlation with some kind of restriction and with the hypothesis. The choice settled on correlations using a boundary above 50% of correlations. For features, was used (Age, Clo, Sex, Met, Thermal preference, Year, Season, Koppen climate classification, Cooling strat- egy building level, City, PPD, Air temperature (C), Outdoor monthly air temperature (C), Relative humidity (%), Air velocity (m/s)) columns. This showed features are the final result, before that we tested a lot of feature combinations. All combinations and variations will be presented in the Experi- ment section. As you can see, new features were added that helped improve the accuracy. The dataset was split into 80% for training and 20% for testing. Thermal comfort columns usually contained

values from 1 to 6. The next hypothesis, convert label values to integers. We will have unique 6 values, from 6 unique digits, we reduce thermal comfort values to 3 digits, which significantly improves accuracy.



**Figure 4 - Feature selection**

1. *Inter Quartile Range (IQR)*

The Interquartile Range (*IQR*) is a statistical measure that represents the spread or dispersion of a dataset. The Interquartile Range (*IQR*) is a measure of statistical dispersion that is calculated as the difference between the third quartile (*Q*3) and the first quartile (*Q*1) of a dataset. Mathematically, it is defined as:

*IQR* = *Q*3 *− Q*1

where *Q*1 is the median of the lower half of the dataset and *Q*3 is the median of the upper half of the dataset.

The Interquartile Range (*IQR*) is a statistical measure used to assess the spread or dispersion of a dataset. It is particularly useful in identifying and dealing with outliers, which are data points that significantly differ from the rest of the dataset.

Here’s how the *IQR* is calculated and how it can be used to remove outliers:

# *Calculation of IQR*:

* + - Firstly, you need to arrange your dataset in ascend- ing order.
    - Then, find the median of the dataset, which is the middle value when the data is sorted. If the dataset has an odd number of observations, the median is the middle value. If it has an even number of observations, the median is the average of the two middle values.
    - Divide the dataset into two halves at the median. The lower half contains all the values less than or equal to the median, and the upper half contains all the values greater than or equal to the median.
    - Find the median of each half. This gives you the first quartile (*Q*1) and the third quartile (*Q*3) of the dataset, respectively.
    - The Interquartile Range (*IQR*) is then calculated as the difference between *Q*3 and *Q*1: *IQR* = *Q*3 *- Q*1.

*−*

# *Identifying outliers using IQR*:

* + - Outliers can be detected using the *IQR* method by considering values that lie below *Q*1 *−* 1*.*5 *×* *IQR* or above *Q*3 + 1*.*5 *× IQR*. These values are considered to be significantly different from the rest of the dataset.
    - Values below *Q*1 *−* 1*.*5 *× IQR* or above *Q*3 + 1*.*5 *×* *IQR* are commonly referred to as lower and upper bounds, respectively.
    - Any data points falling outside these bounds can be considered outliers.

# *Removing outliers using IQR*:

* + - Once outliers are identified using the *IQR* method, you can choose to remove them from the dataset to improve the robustness of your analysis or model.
    - Outliers can be removed by filtering the dataset to exclude any observations that fall outside the lower and upper bounds defined by *Q*1 *−* 1*.*5 *× IQR* and

*Q*3 + 1*.*5 *× IQR*, respectively.

* + - After removing outliers, the dataset may be more representative of the underlying distribution and less influenced by extreme values.

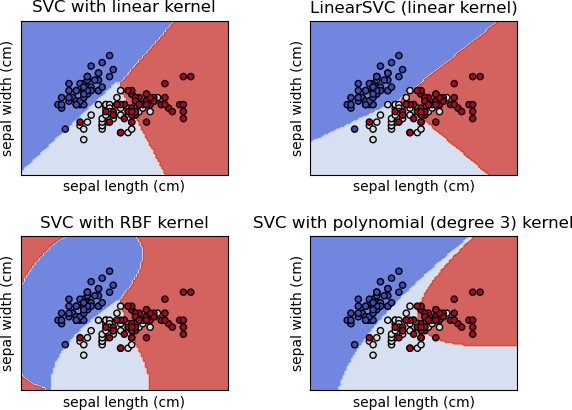
# *Considerations*:

* + - While the *IQR* method is effective in identifying and removing outliers, it’s important to exercise caution and consider the context of the data.
    - Outliers may sometimes carry valuable informa- tion or be indicative of rare but important events. Therefore, the decision to remove outliers should be made judiciously based on the specific goals of the analysis or model.
    - Additionally, the choice of the multiplier (1.5 in the conventional method) used to define the bounds can be adjusted depending on the desired level of sensitivity to outliers.

In summary, the Interquartile Range (*IQR*) is a useful statistical measure for assessing the spread of a dataset and identifying outliers. By calculating the *IQR* and defining bounds based on it, outliers can be effectively detected and removed, leading to a more robust analysis or model.

1. *Support Vector Machine (SVM)*

Support Vector Machine is a powerful supervised machine learning algorithm used for classification and regression tasks. SVM works by finding the optimal hyperplane that separates different classes or, in the case of regression, predicts continuous outcomes. The key concept behind SVM is to maximize the margin between different classes or, in regression, to minimize the error between predicted and actual values while controlling for overfitting.

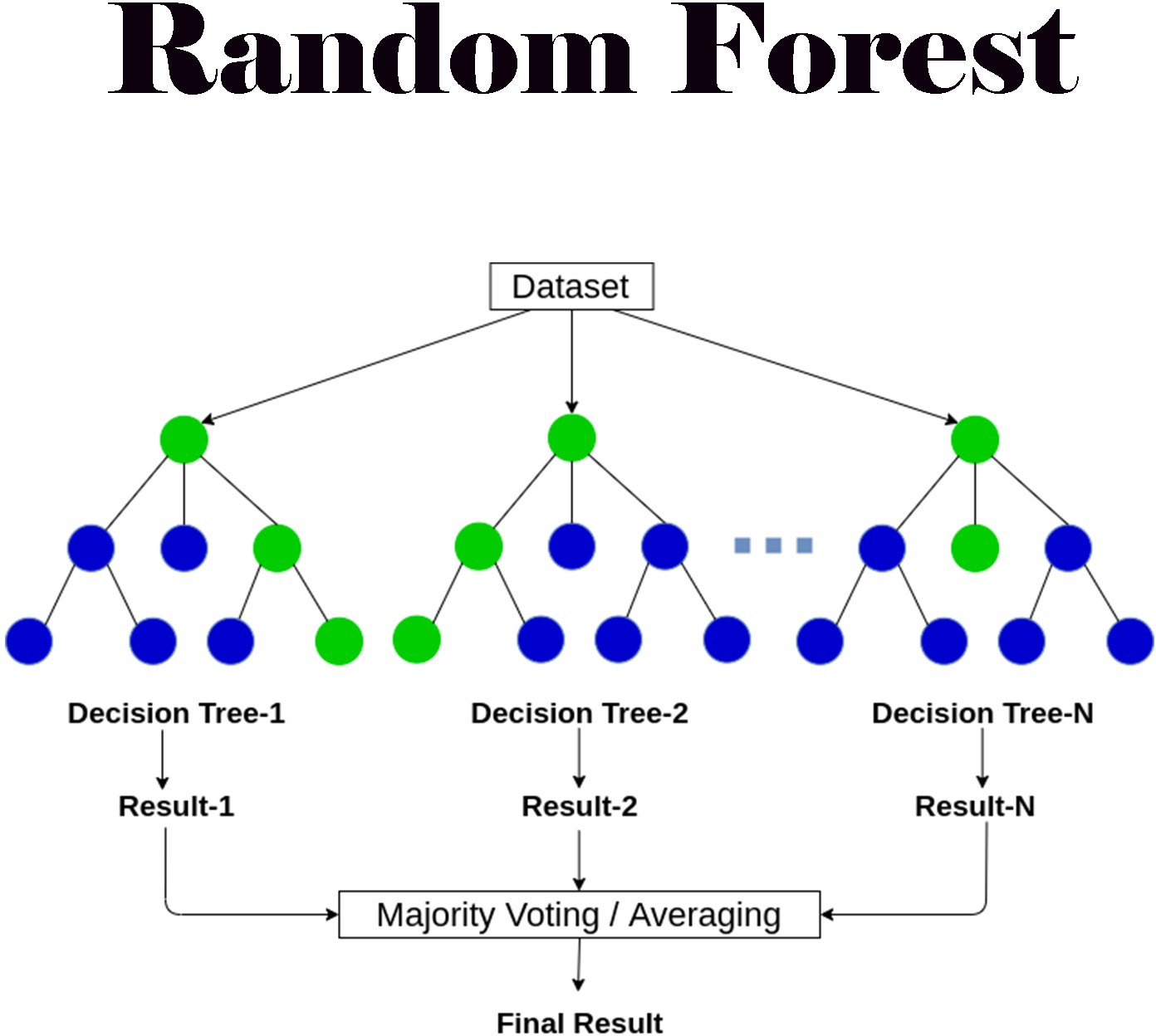


**Figure 5 - Support Vector Machine**

In the context of thermal comfort prediction, SVM can be utilized to analyze complex relationships between various environmental factors such as temperature, humidity, and air velocity, and the corresponding human thermal comfort responses. By training the SVM model on labeled datasets containing information about environmental conditions and associated thermal comfort ratings, the algorithm can learn to predict the level of thermal comfort for a given set of environmental parameters.

1. *Random Forest (RF)*

Random Forest is a popular machine-learning algorithm that can be used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and stable model.



**Figure 6 - Random Forest**

Data preparation involves cleaning the data, dealing with missing values, and transforming it to ensure it is suitable for the algorithm. Random sampling is used to randomly select a subset of the data to use for training each decision tree. Decision tree creation is created using recursive partitioning and feature selection. Voting is used to combine the predictions of all the trees to make the final prediction. Evaluation is done using a validation set. Overall, the Random Forest algorithm is a powerful machine-learning method that can be used for a wide range of tasks. It is easy to use and can produce accurate and stable predictions even with noisy or incomplete data. When applied to thermal comfort prediction, Random Forest models excel in capturing nonlinear relationships and inter- actions among various environmental factors. By aggregating predictions from multiple decision trees, Random Forest can provide accurate estimates of thermal comfort levels across different environmental conditions.

1. *Integration with IoT*

The IoT component of the system involves deploying a network of sensors within the building. These sensors collect real-time data on various environmental conditions, such as temperature, humidity, CO2 levels, and occupancy. Data from IoT sensors are transmitted to a central server for storage and analysis. Wireless communication protocols like Wi-Fi, Bluetooth, or LoRaWAN can be used for efficient data transfer. The AI models receive real-time data from the IoT sensors, enabling them to continuously update predictions and make immediate adjustments to the HVAC system for optimal ther- mal comfort. An important aspect of the system is its ability to create a feedback loop that maintains thermal comfort. The AI algorithms analyze the real-time data from IoT sensors and make recommendations or control the HVAC system to ensure that thermal comfort is maintained. For instance, if the system detects a deviation from the desired comfort level, it can adjust the temperature, humidity, or airflow accordingly.

1. *Alternative prediction value*

So, an alternative way for predicting value we use the Thermal preference column instead Thermal comfort. If we do not switch from 6 digits to 3 as before.

**Results and Discussion.** After filtering, we have 21 columns out of 70. And we make feature selections using correlation. Moreover, we avoid choosing Fanger’s features. After one more filtering by cor- relation and with model SelectKbest which will help us to get the list of features. We get more than 3 variations, but we stopped in those variants:

1. First set of 17 features: (Age, Sex, Met, Thermal pref- erence, Thermal sensation, Clo, Subjects height (cm), Subjects weight (kg), Year, Season, Koppen climate clas- sification, Building type, Cooling strategy building level, Air temperature (C), Outdoor monthly air temperature (C), Relative humidity (%), Air velocity (m/s)).
2. Second set of 9 features: (Age, Sex, Met, Clo, Year, Season, Air temperature (C), Relative humidity (%), Air velocity (m/s)).
3. Third set of 15 features: (Age, Clo, Sex, Met, Thermal preference, Year, Season, Koppen climate classification, Cooling strategy building level, City, PPD, Air tempera- ture (C), Outdoor monthly air temperature (C), Relative humidity (%), Air velocity (m/s))

In the end, we have 17 columns and 6765 rows. Starting work, we first take 17 out of 17 columns, we get not good results. Second iteration we take 9 out of 17 columns they also give results around the first iteration. In the last iteration, we take 15 out of 17 columns results are not good either. For that situation, we tested our hypothesis and IQR method gives approximately 3-4% accuracy, and the reducing label values method gives 20-23% accuracy. By changing the parameters of the models we define good parameters for our case then for the SVM model, the parameters were taken as *kernel* = ”rbf”, *gamma* = 0.001, and *c* = 3. And for the Random Forest, the parameters were taken as *estimators* = 300, *max depth* = 15. These parameters gave the maximum accuracy values. The results of comparing the use of LabelEncoder and OneHotEncoder in the dataset give a 2-4% percent difference between them. Regardless of the features, and what parameters have been entered. This influenced the fact to take OneHotEncoder. If you do data standardization, the accuracy results will not change much and remain practically the same. For standardization, we used StandardScaler and MinMaxScaler models.

Below are presented 1, 2, and 3 tables the beginning results of our prediction:

**Table 1 - Iteration of 17 features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.509 | 0.451 | 0.509 | 0.436 |
| RF | 0.543 | 0.505 | 0.543 | 0.5 |

**Table 2 - Iteration of 9 features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.507 | 0.461 | 0.507 | 0.438 |
| RF | 0.526 | 0.513 | 0.526 | 0.49 |

**Table 3 - Iteration of 15 features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.533 | 0.448 | 0.533 | 0.433 |
| RF | 0.54 | 0.475 | 0.539 | 0.482 |

According above results, we tried to improve accuracy using our hypothesis. Below presented 4, 5, and 6 tables show the results of *IQR* method:

**Table 4 - Iteration of 17 features with *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.522 | 0.44 | 0.522 | 0.441 |
| RF | 0.548 | 0.517 | 0.548 | 0.504 |

**Table 5 - Iteration of 9 features with *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.507 | 0.44 | 0.383 | 0.424 |
| RF | 0.52 | 0.501 | 0.52 | 0.479 |

**Table 6 - Iteration of 15 features with *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.563 | 0.539 | 0.563 | 0.425 |
| RF | 0.57 | 0.494 | 0.57 | 0.5 |

From previous results, *IQR* method upgrades accuracy approximately to 2-5%. Next, we work with the reduction of label values to increase the accuracy:

**Table 7 - Iteration of 17 features with reducing labels**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.715 | 0.644 | 0.715 | 0.614 |
| RF | 0.744 | 0.708 | 0.744 | 0.704 |

**Table 8 - Iteration of 9 features with reducing labels**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.688 | 0.598 | 0.688 | 0.569 |
| RF | 0.699 | 0.657 | 0.699 | 0.645 |

**Table 9 - Iteration of 15 features with reducing labels**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.78 | 0.608 | 0.78 | 0.683 |
| RF | 0.78 | 0.719 | 0.78 | 0.727 |

Adding *IQR* method to the reduced features and get such results:

**Table 10 - Iteration of 17 features with reducing labels and *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.726 | 0.598 | 0.726 | 0.621 |
| RF | 0.733 | 0.678 | 0.733 | 0.688 |

**Table 11 - Iteration of 9 features with reducing labels and *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.706 | 0.498 | 0.706 | 0.584 |
| RF | 0.717 | 0.668 | 0.717 | 0.653 |

**Table 12 - Iteration of 15 features with reducing labels and *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.835 | 0.697 | 0.835 | 0.76 |
| RF | 0.821 | 0.738 | 0.821 | 0.766 |

There was also work on the alternative method. Here Thermal comfort and Thermal preference will change places. And now we will predict Thermal preference.

**Table 13 - Alternative iteration of 15 features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.68 | 0.714 | 0.68 | 0.605 |
| RF | 0.714 | 0.712 | 0.714 | 0.685 |

**Table 14 - Alternative iteration of 15 features with reducing labels**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.67 | 0.698 | 0.67 | 0.584 |
| RF | 0.705 | 0.702 | 0.705 | 0.673 |

**Table 15 - Alternative iteration of 15 features with reducing labels and *IQR***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 0.694 | 0.55 | 0.694 | 0.577 |
| RF | 0.709 | 0.676 | 0.709 | 0.651 |

**Conclusion.** As a result, we pass a verdict that added *IQR* method, we worked with new 8 features and 7 features were already in other articles, considering the option with guessing Thermal comfort and Thermal preference, the difference between the two algorithms is 1-3%. Basically Random Forest looks more stable than SVM. I would also like to note that the alternative version was in the lead in 9 and 15 features, Table 3 and Table 13. But, when we started converting from 6 to 3 Thermal comfort values and added *IQR* method, our main option immediately won. In the future, I will add new features to improve accuracy. For example, one of them is Heart Rate Variability (HRV) [11]. On top of that, I want to test neural networks and deep learning as I have seen good results with these algorithms. I also considered these [15, 16] papers for the basis of a new work. Some commonly used algorithms for this purpose include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) Networks, Autoencoders, and Deep Belief Networks (DBNs).

**References**

1. Christos G. Polycarpou, Marios M. Milis, George M. Panayiotou. Iot-enabled automatic synthesis of distributed feedback control schemes in smart buildings.// IEEE Internet of Things Journa[l](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6488907).- 2021.-Vol.8, Iss.4.- P.2615-2626.  [DOI 10.1109/JIOT.2020.3019662](https://doi.org/10.1109/JIOT.2020.3019662)

2. Mulyana Binti Bin Omar, Ridha Mohamed Salleh and Faridah Hani Saripuddin. Predicting thermal comfort of hvac building using 6 thermal factors.// [8th International Conference on Information Technology and Multimedia (ICIMU)](https://ieeexplore.ieee.org/xpl/conhome/9243083/proceeding).-2020. [DOI 10.1109/ICIMU49871.2020.9243466](https://doi.org/10.1109/ICIMU49871.2020.9243466)

3. Sara Sorcinelli, Matteo Arnesano, Marco Uriarte, Amaia Torrens-Galdiz, J. Ignacio Revel, Gian Marco Morresi and Nicole Casaccia. Sensing physiological and environmental quantities to measure human thermal comfort through machine learning techniques//[IEEE Sensors Journal](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=7361). -2021.-Vol.21. - Iss.10.-P. 12322-12337. [DOI 10.1109/JSEN.2021.3064707](https://doi.org/10.1109/JSEN.2021.3064707)

4. Juliana Caesarendra, Wahyu Shona Laila, Dina Candra Kurnia, Jundika Widiastuti and Ratih Zaini.Prediction on the indoor thermal comfort of occupied room based on iot climate measurement open datasets // International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS).- 2020. [DOI 10.1109/ICIMCIS51567.2020.9354277](https://doi.org/10.1109/ICIMCIS51567.2020.9354277)

5. Yadong Xu, Zhanbo Wu, Jiang Liu, Yaping Guo, Yuntao Guan, Xiaohong Su and Ying Zhou. Group comfort models: Predicting indoor group thermal comfort by learning preferences of multiple occupants.//  [IEEE 16th International Conference on Automation Science and Engineering (CASE)](%20IEEE%2016th%20International%20Conference%20on%20Automation%20Science%20and%20Engineering%20(CASE)).- 2020. DOI [10.1109/CASE48305.2020.9216834](https://doi.org/10.1109/CASE48305.2020.9216834)

6. M. Hani Mohamed Salleh and Faridah Binti Saripuddin. Monitoring thermal comfort level of commercial buildings’ occupants in a hot-humid climate country using k-nearest neighbors model.// [5th International Conference on Power and Renewable Energy (ICPRE)](https://ieeexplore.ieee.org/xpl/conhome/9233041/proceeding).- 2020. DOI 10.1109/ICPRE51194.2020.9233145

7. Moez Merghem-Boulahia, Leila Khalil and Maysaa Esseghir. An iot environment for estimating occupants’ thermal comfort// [IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications](https://ieeexplore.ieee.org/xpl/conhome/9210501/proceeding).- 2020. https://doi.org/10.1109/PIMRC48278.2020.9217157

8. Manisa Chen, Tao Rahman, Saifur Zhang and Xiangyu Pipattanasomporn. An iot-based thermal model learning framework for smart buildings.//[IEEE Internet of Things Journal](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6488907).- Vol. 7, [Iss. 1](https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8955685&punumber=6488907).-2020.-P. 518-52. [DOI 10.1109/JIOT.2019.2951106](https://doi.org/10.1109/JIOT.2019.2951106)

9. Asif Irshad Irfan, Sayed Ameenuddin Alam, Md Mottahir Almalawi, Abdulmohsen Zahir, Md Hasan Irshad and Kashif Khan. Utilizing artificial neural network for prediction of occupants thermal comfort: A case study of a test room fitted with a thermoelectric air-conditioning system.// [IEEE Access](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6287639).-Vol. 8-2020.-P. 99709 – 99728. DOI 10.1109/ACCESS.2020.2985036

10. Jie Wen, Yonggang Gao and Guanyu Li.Deepcomfort: Energy-efficient thermal comfort control in buildings via reinforcement learning.// IEEE Internet of Things Journal. – 2020. - Vol. 7(9). –P. 8472-8484 DOI 10.1109/JIOT.2020.2992117

11. Burak Shi, Zixiao Shen, Weiming Yang and Chunsheng Gunay. Machine learning-based prognostics for central heating and cooling plant equipment health monitoring.// IEEE Transactions on Automation Science and Engineering/. – 2021.- Vol.18(1).- P. 346–355. DOI 10.1109/TASE.2020.2998586

12. Hiroshi Matsuhashi, Ryuji Wang and Ziyang Onodera,. Proposal of relative thermal sensation: Another dimension of thermal comfort and its investigation.// IEEE Access. – 2021.- P.(99):1-1 DOI 10.1109/ACCESS.2021.3062393

13. Yonggang Tseng, King Jet Jin, Guangyu Zhang and Wei Wen. Demystifying thermal comfort in smart buildings: An interpretable machine learning approach.// [IEEE Internet of Things Journal](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6488907).-2021.-Vol.8,Iss.10.-P.8021-8031. [DOI 10.1109/JIOT.2020.3042783](https://doi.org/10.1109/JIOT.2020.3042783)

14.SHRAE dataset, “ASHRAE Global Thermal Comfort Database II’’, 2020,<https://www.kaggle.com/datasets/claytonmiller/ashrae-global-thermal-comfort-database-ii>

15. Nivethitha Somu, Anirudh Sriram, Anupama Kowli, Krithi Ramamritham. A hybrid deep transfer learning strategy for thermal comfort prediction in buildings//[Building and Environment](https://www.sciencedirect.com/journal/building-and-environment).-2021.- Vol.204: 108133. [DOI 10.1016/j.buildenv.2021.108133](https://doi.org/10.1016/j.buildenv.2021.108133)

16. Hansaem Park, Dong Yoon Park Prediction of individual thermal comfort based on ensemble transfer learning method using wearable and environmental sensors// [Building and Environment](https://www.sciencedirect.com/journal/building-and-environment).-2022.- Vol.207: 108492. DOI 10.1016/j.buildenv.2021.108492

***Information about the authors***

Assymkhan N.B. – master of Kazakh-British Technical University, e-mail: [anb.asymhan@gmail.com](mailto:anb.asymhan@gmail.com);

Kartbaev A.Z. – PhD of Kazakh-British Technical University, e-mail: a.kartbaev@kbtu.kz;

***Сведения об авторах***

Асымхан Н.Б. – магистрант Казахстанско-Британский Технического Университета, e-mail: [anb.asymhan@gmail.com](mailto:anb.asymhan@gmail.com);

Картбаев А.З. – PhD? Казахстанско-Британский Технического Университета, e-mail: [a.kartbaev@kbtu.kz](mailto:a.kartbaev@kbtu.kz);

**ҒТАМР 20.15.13**

**МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ АРҚЫЛЫ СҮТ БЕЗДЕРІНІҢ АУРУЛАРЫН ТИІМДІ ДИАГНОСТИКАЛАУ**

**А.Р. Оразаева, Д.А. Тусупов🖂**

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

**🖂**Корреспондент-автор: [tussupov@mail.ru](mailto:oaris.83@gmail.com)

Бұл мақала заманауи машиналық оқыту технологияларын, атап айтқанда, You Only Look Once (YOLOv8) және Faster Region-based Convolutional Neural Network (R-CNN) пайдалана отырып, сүт бездерінің патологиясын тиімді анықтау әдістерін зерттеуге және дамытуға бағытталған. Мақалада сүт безі ауруларын диагностикалаудың қолданыстағы тәсілдеріне талдау жасалып, олардың тиімділігі бағаланады. YOLOv8 және Faster R-CNN архитектуралары маммографиялық кескіндердегі патологияларды анықтау үлгілерін жасау үшін қолданылады. Зерттеу әртүрлі ауырлық дәрежесін және аурудың сипаттамаларын ескере отырып, алты түрлі деңгейде анықталған сүт безі патологияларын жіктейді және талдайды. Бұл әдіс аурудың дамуын дәлірек бағалауға мүмкіндік береді және емдеуді жекелендірілген жоспарлау үшін қосымша ақпаратты ұсынады. Осы деңгейлердегі жіктеу нәтижелері медициналық шешімдер қабылдау сапасын жақсарта алады және дәрігерлерге дәлірек ақпарат береді, дәлірек айтсақ, сүт безі ауруларын диагностикалау мен емдеудің жалпы тиімділігін арттырады. Эксперименттік нәтижелер сүт бездерінің ықтимал патологияларын жылдам және сенімді анықтауға мүмкіндік беретін жоғары дәлдік пен кескінді жылдам өңдеуді көрсетеді. Зерттеу нәтижелері медициналық диагностикада машиналық оқыту алгоритмдерінің тиімділігін растайды, ерте диагностика мен емдеу нәтижелерін жақсарту үшін сүт безі ауруларын анықтаудың автоматтандырылған жүйелерін одан әрі дамыту әлеуетін көрсетеді.

**Түйін сөздер.** Терең оқыту, Faster Region-based Convolutional Neural Network (R-CNN), You Only Look Once (YOLOv8), деректер базасы, модель.

**ЭФФЕКТИВНАЯ ДИАГНОСТИКА ЗАБОЛЕВАНИЙ МОЛОЧНОЙ ЖЕЛЕЗЫ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ**

**А.Р. Оразаева, Д.А. Тусупов 🖂**

Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан,

e-mail: [tussupov@mail.ru](mailto:oaris.83@gmail.com)

Данная статья сосредоточено на исследовании и разработке методов эффективного обнаружения патологий молочной железы с использованием современных технологий машинного обучения, в частности You Only Look Once (YOLOv8) и Faster Region-based Convolutional Neural Network (R-CNN). В статье представлен анализ существующих подходов к диагностике заболеваний молочной железы и дана оценка их эффективности. Архитектуры YOLOv8 и Faster R-CNN используются для разработки моделей обнаружения патологий на маммографических снимках. Исследование классифицирует и анализирует выявленные патологии молочной железы на шести различных уровнях с учетом различной степени тяжести и характеристик заболеваний. Этот метод позволяет более точно оценить прогрессирование заболевания и предлагает дополнительные сведения для более персонализированного планирования лечения. Результаты классификации на этих уровнях могут повысить качество принятия медицинских решений и предоставить врачам более точную информацию, в конечном итоге повышая общую эффективность диагностики и лечения заболеваний молочной железы. Экспериментальные результаты показывают высокую точность и быструю обработку изображений, что позволяет быстро и надежно обнаруживать потенциальные патологии молочной железы. Результаты исследования подтверждают эффективность алгоритмов машинного обучения в медицинской диагностике, подчеркивая потенциал дальнейшего развития автоматизированных систем обнаружения заболеваний молочной железы для улучшения ранней диагностики и результатов лечения.

**Ключевые слова.** Глубокое обучение, Faster Region-based Convolutional Neural Network (R-CNN), You Only Look Once (YOLOv8), база данных, модель.

**EFFECTIVE DIAGNOSTICS OF BREAST DISEASES USING MACHINE LEARNING METHODS**

**A.R. Orazayeva, J.A. Tussupov🖂**

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,

e-mail: [tussupov@mail.ru](mailto:oaris.83@gmail.com)

This study focuses on researching and developing methods for the efficient detection of breast pathologies using modern machine learning technologies, specifically You Only Look Once (YOLOv8) and Faster Region-based Convolutional Neural Network (R-CNN). The paper provides an analysis of existing approaches to diagnosing breast diseases and evaluates their effectiveness. The YOLOv8 and Faster R-CNN architectures are employed to develop models for detecting pathologies in mammography images. The research classifies and analyzes identified breast pathologies at six different levels, considering varying degrees of severity and characteristics of the diseases. This method enables a more accurate assessment of disease progression and offers additional insights for more personalized treatment planning. Classification results across these levels can enhance medical decision-making quality and provide doctors with more precise information, ultimately improving the overall efficiency of breast disease diagnosis and treatment. Experimental results show high accuracy and rapid image processing, enabling fast and reliable detection of potential breast pathologies. The findings confirm the effectiveness of machine learning algorithms in medical diagnostics, highlighting the potential for further advancements in automated breast disease detection systems to enhance early diagnosis and treatment outcomes.

**Keywords.** Deep learning, Faster Region-based Convolutional Neural Network (R-CNN), You Only Look Once (YOLOv8), database, model.

**Кіріспе.** Қазіргі заманғы медициналық диагностикалық технологиялар [1-4] машиналық оқыту мүмкіндіктері мен қағидаларын ауруларды анықтаудың дәлдігі мен тиімділігін арттыру мақсатында қарқынды түрде даму үстінде [5, 6]. Әйелдер денсаулығын қорғаудағы маңызды мәселелердің бірі – сүт безі ауруларын [7-10], соның ішінде қатерлі ісіктің әртүрлі түрлерін [11-12] және басқа да бұзылыстарды анықтау болып табылады. Бұл мақалада сүт безі патологиясын анықтаудың тиімді әдістері қарастырылып, оның өзектілігі ерекше атап өтіледі. Машиналық оқыту, атап айтқанда, You Only Look Once (YOLOv8) [13-16] және Faster Region-based Convolutional Neural Network (R-CNN) [17-18] алгоритмдері талдау процесін автоматтандыруда келешегі зор құралдар ретінде ұсынылады. Аталған әдістер маммографиялық суреттерден ықтимал патологияларды анықтап қана қоймай, оларды әртүрлі ауырлық деңгейлеріне классификациялауға мүмкіндік береді. Бұл тәсіл сүт безі ауруларын ерте анықтау және жеке емдеу әдістерін қолдануға жаңа мүмкіндіктер ашады [19-20].

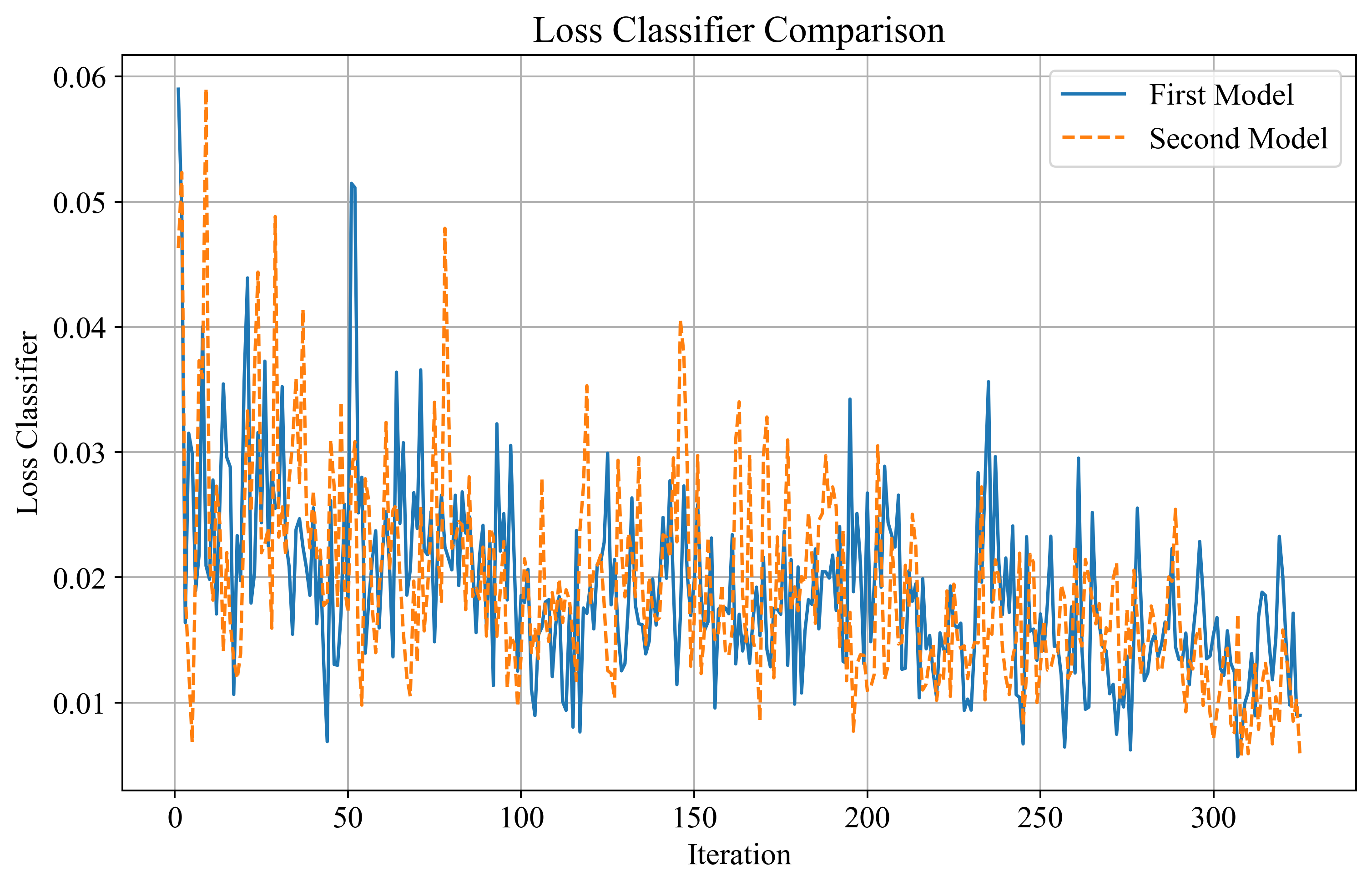
Бұл зерттеуде біз YOLOv8 және Faster R-CNN алгоритмдерін пайдалана отырып, сүт безі патологияларын анықтаудың тиімді әдістерін әзірлеуге және тестілеуге назар аударамыз. Аталған әдістерді талдау, салыстыру және олардың медициналық тәжірибедегі қолданылуы сүт безі ауруларын диагностикалау мен емдеу саласын елеулі түрде жетілдіріп, диагностикалық процедуралардың дәлдігі мен тиімділігін арттыруға мүмкіндік береді. Сүт безі патологияларын диагностикалаудың заманауи әрі тиімді әдістеріне қолжетімділікті қамтамасыз ету денсаулық сақтау стратегиясының маңызды бөлігін құрайды. Сүт безінің аурулары, соның ішінде сүт безі қатерлі ісігі, әйелдер денсаулығына ең жиі кездесетін әрі қауіпті аурулардың бірі болып табылады. Осыған орай, машиналық оқытудың алдыңғы қатарлы технологияларын қолдану скрининг пен диагностика сапасын жақсартып, ауруды анықтау мен емдеуді бастау арасындағы уақытты қысқартудың тиімді құралы бола алады. Алайда, осындай артықшылықтарына қарамастан, медициналық практикаға машиналық оқыту алгоритмдерін енгізу олардың дәлдігі мен сенімділігін, сондай-ақ денсаулық сақтау саласындағы деректер қауіпсіздігі стандарттарына сәйкестігін мұқият зерттеуді қажет етеді. Бұл мақалада машиналық оқыту технологияларын медициналық тәжірибеге қалай тиімді интеграциялауға болатынын және оларды қауіпсіз әрі тиімді пайдалану үшін қандай шаралар қабылдау керектігін қарастырамыз.

Бұл зерттеу нәтижесінде біз сүт безі патологияларын анықтаудың жаңа, тиімді әдістерін енгізуді ғана көздеп қоймай, сондай-ақ олардың медициналық тәжірибеде қолдану мүмкіндігін көрсетіп, сүт безі ауруларын диагностикалау мен емдеуді жетілдіруге елеулі үлес қосуды мақсат етеміз. Технологияның үздіксіз дамуы және денсаулық сақтау саласындағы өсіп келе жатқан қажеттіліктер аясында сүт безі патологияларын анықтау [21], [22] әлемдік денсаулық сақтаудағы маңызды мәселеге айналып отыр. Сүт безі қатерлі ісігі – әйелдер арасында ең жиі кездесетін әрі өлімге әкелетін онкологиялық аурулардың бірі болып қала береді. Заридзе Д.Г. және т.б. [23] мақаласында қатерлі ісіктерден болатын өлім-жітімді азайтуға алғашқы профилактиканы, скринингтік тексеруді және емдеуді қоса алғанда, ғылыми негізделген кешенді мақсатты бағдарламаны іске асыру нәтижесінде ғана қол жеткізуге болатынын зерттеген. Давыдов М. И. және т.б. [24] өз мақаласында симптомсыз ісіктерді ерте анықтау және оны емдеу, сонымен қатар скрининг өлімді азайтуға әкелетіндігі туралы жазған. [25] мақала сүт безі қатерлі ісігінің аурушаңдық, өлім және өмір сүрудегі негізгі факторларын және олардың рөлін анықтаған. Бұл жұмыстардың жалпы бағыты қатерлі ісікті алдын алу.

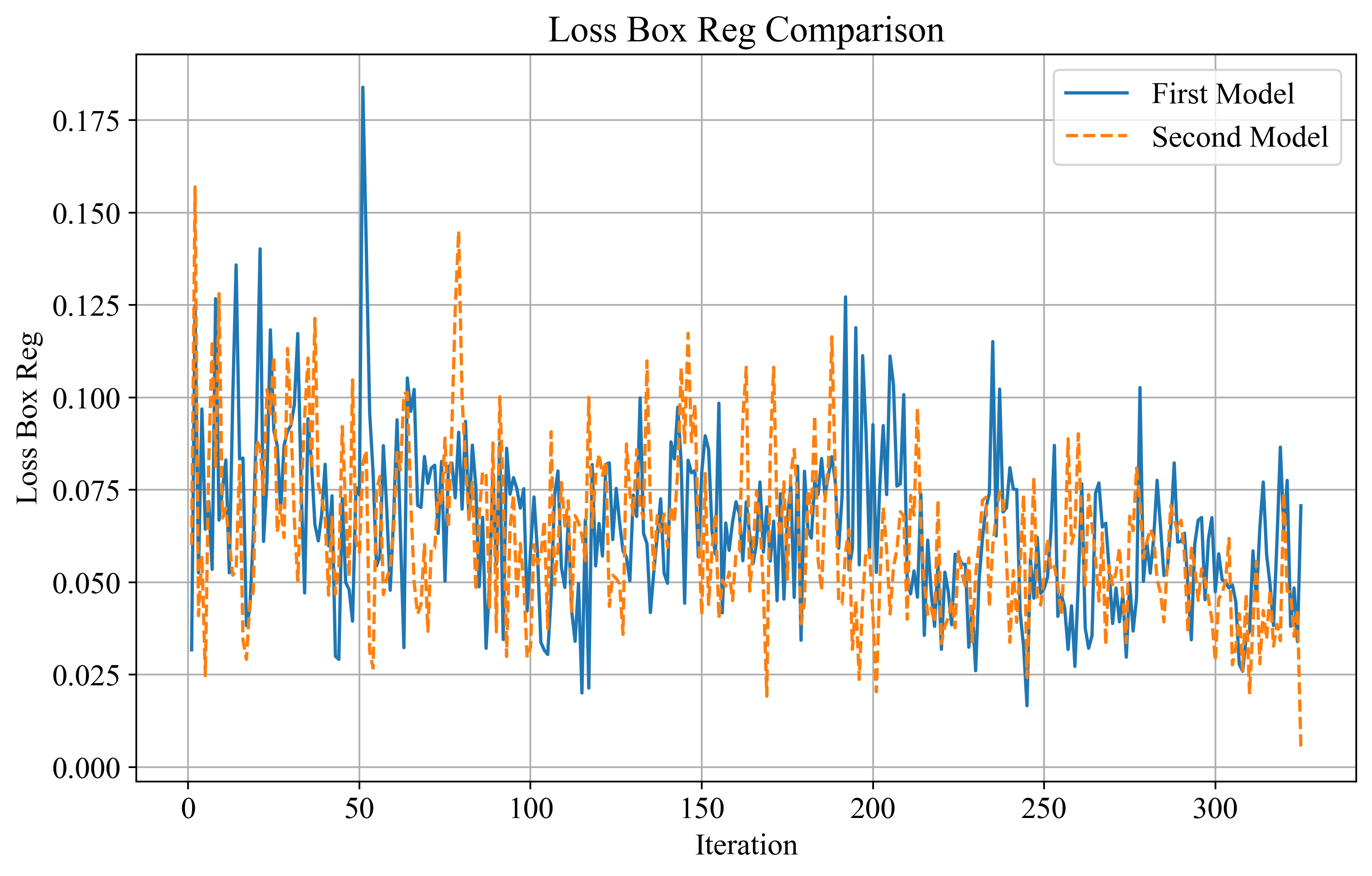
[26] мақалада конволюционды нейрондық желілерге (CNN) негізделген кескіндерді классификациялау алгоритмдеріне шолу жасалған. Авторлар қашықтықтан зондтау мақсатында CNN қолданудың негізгі архитектуралары мен әдістерін, сондай-ақ қол жеткізілген нәтижелерді талқылайды, әсіресе кескінді классификациялаудың дәлдігі мен тиімділігін арттыру мәселесіне ерекше назар аударады. Ал [27] мақалада Пап-тест кескіндерін классификациялау үшін цитопатологтарға арналған ансамбльдік терең оқыту әдісі сипатталады. Авторлар диагностикалық дәлдікті арттыру мақсатында бірнеше терең оқыту модельлерін біріктіретін жүйе әзірлеген. Бұл әдіс медициналық бейнелерді талдау процесін автоматтандыруға көмектесіп, мамандардың жұмыс жүктемесін азайтуға ықпал етеді. [26] және [27] еңбектерді талдау нәтижесінде, терең оқыту әдістері, әсіресе конволюционды нейрондық желілер мен ансамбльдік оқыту, кескін классификациясын әртүрлі салаларда жақсартуда маңызды рөл атқаратыны анықталды. Бұл технологиялар талдау дәлдігі мен тиімділігін арттырып қана қоймай, сонымен қатар процестерді автоматтандыруға, мамандардың жұмысын жеңілдетуге және диагностика мен деректерді өңдеудің сапасын жақсартуға ықпал етеді. Жасанды интеллект қолдану арқылы маммографиядан алынған кескіндер, клиникалық, генетикалық және патологиялық деректерді біріктіру тәуекел модельлерін жетілдіруге мүмкіндік береді. Жаңа бейнелеу әдістері, генетикалық тестілеу және молекулярлық профильдеу қауіп модельлерінің дәлдігін арттыруға ықпал етеді. Аурудың күрделілігі, деректердің шектеулі болуы және модельлерді енгізу параметрлері де зерттеуде талқыланады. Ерте анықтау және емдеу нәтижелерін жақсарту үшін пәнаралық тәсілдің маңыздылығы ерекше атап өтіледі. Медициналық диагностикадағы прогресті ескере отырып, сүт безі ауруларын тиімді әрі уақытылы анықтау қажеттілігі артып келеді, бұл машиналық оқыту әдістері сияқты инновациялық технологияларды енгізуді талап етеді. YOLOv8 және Faster R-CNN секілді машиналық оқыту алгоритмдерін қолдану арқылы автоматтандырылған медициналық кескін талдауында жаңа мүмкіндіктер ашылып, диагностика сапасы айтарлықтай жақсарады, бұл өз кезегінде сәтті емдеу ықтималдығын арттырады. Бұл зерттеу осы әдістердің медицинадағы маңыздылығын және сүт безі патологияларын емдеуде олардың әлеуетін көрсетуге бағытталған.

**Материалдар мен әдістер.** Бұл зерттеу сүт бездерінің маммографиялық кескіндеріндегі патологияларды анықтауға арналған екі жетілдірілген компьютерлік көру алгоритмін оқытуды талдайды: You Only Look Once (YOLOv8) және Faster Region-based Convolutional Neural Network (Faster R-CNN). Алайда, олардың тиімділігі мен сенімділігін толығымен растау үшін контрастты күшейту және тін құрылымының маңызды бөлшектерін айқындау мақсатында кескінді өңдеу әдістері, мысалы, пиксель мәндерін қалыпқа келтіру және гистограмманы теңестіру, қолданылып, әртүрлі әрі көлемді деректер базасында қосымша сынақтар қажет. Деректерді алдын ала өңдеудің бірегей әдістерінің бірі – аномальды аймақтарды маска ретінде белгілеу болды, бұл аномалиялардың шектеу ұяшықтарын (Bounding Boxes) тиімді бөлектеуге мүмкіндік берді. Бұл маскалар патологияның түрі мен деңгейі туралы ақпарат беретін модельдерге аннотация ретінде қолданылды. Модельдерді оқыту мен тексерудің сапасын қамтамасыз ету үшін деректер оқыту және валидация жиындарына бөлінді. Оқыту процесі YOLOv8 және Faster R-CNN модельдері бойынша 1-суретте көрсетілгендей екі модель бойынша өтті. Алдын ала дайындалған салмақтар қолданылып, модельдер бірнеше оқу дәуірлері бойы жаттығудан өтті. Валидация кезінде модельлердің валидациялық деректер базасындағы өнімділігі бағаланды. Нәтижелер жоғалтудың, дәлдіктің және F1 өлшемінің оқу дәуірлеріне тәуелділігін көрсететін графиктер арқылы талданды. Медициналық бейнелерден аномалияларды анықтау үшін компьютерлік модельдерді оқыту күрделі әрі көпқырлы процесс болып табылады. Осы зерттеу нәтижелеріне сүйене отырып, YOLOv8 және Faster R-CNN модельдерінің медициналық диагностикада, әсіресе маммографиялық суреттердегі сүт безі патологияларын анықтауда перспективалары бар деген қорытынды жасауға болады. Дегенмен, олардың тиімділігін және сенімділігін дәлелдеу үшін қосымша сынақтар қажет, сонымен қатар ықтимал қателіктерді талдау және түзету үшін модельлерді одан әрі зерттеу маңызды.

Модельдің дәлдігі мен жалпылау қабілеті арасында оңтайлы тепе-теңдікке қол жеткізу үшін гиперпараметрлерді, соның ішінде регуляризация параметрлері мен оқу жылдамдығын мұқият баптау маңызды қадам болып табылады. Сонымен қатар, Faster R-CNN моделінің жалпы өнімділігі мен тұрақтылығын бағалау мақсатында валидация және сынақ модельлерінде тестілеу жүргізіледі. Зерттеу Faster R-CNN-ның оқыту және тестілеу нәтижелерін дәлдік, еске түсіру және F1-өлшем сияқты көрсеткіштерді пайдаланып салыстырмалы талдаудан басталады. Бұл көрсеткіштер модельдің сүт безі патологияларын дәл анықтау және локализациялау қабілетін бағалауға мүмкіндік береді. Кейіннен Faster R-CNN моделінің жоғары дәлдікпен патологияны анықтау артықшылықтарын көрсету үшін, 2-суретте көрсетілгендей, YOLOv8 сияқты баламалы әдіспен салыстыру жүргізіледі.



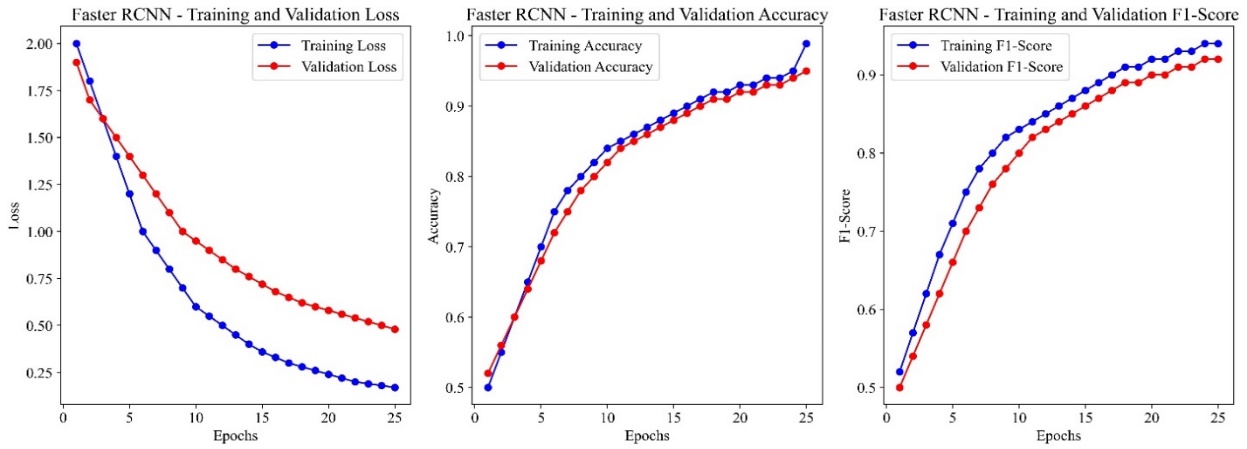
**1-сурет. Екі модель үшін классификатор жоғалтуының графикалық салыстырылуы**



**2-сурет. Екі модель арасындағы рамканың регрессия жоғалтуын салыстыру**

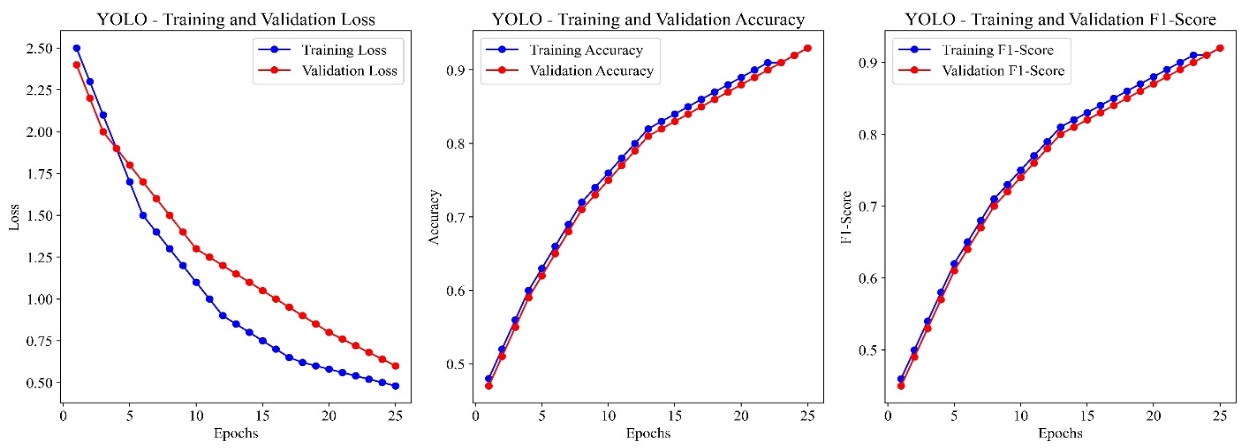
Әдістеме Faster R-CNN-ның соңғы нәтижелеріне оқу моделінің мөлшері мен түрлі оқу параметрлерінің әсерін талдауға бағытталған. Бұл сүт бездерінің патологияларын анықтау кезінде максималды дәлдікке қол жеткізу үшін оңтайлы параметрлер мен жағдайларды анықтауға мүмкіндік береді. Соңғы нәтижелер мен қорытындылар медициналық тәжірибеде сүт безі патологиясын жоғары дәлдікпен анықтаудың тиімді құралы ретінде Faster R-CNN әдісін таңдаудың негізін құрайды. Әдістеме сонымен қатар, Faster R-CNN арқылы анықталған патологияларды автоматтандырылған және қолмен талдау механизмдерімен сапаны бақылау жүйесін құруды көздейді. Бұл кезең нақты клиникалық тәжірибеде модель нәтижелерінің сенімділігі мен дәлдігін тексеруге бағытталған. Анықталған өзгерістерді медициналық мамандардың қолмен тексеруі патологияларды дәл анықтауға қосымша сенімділік деңгейін қамтамасыз етеді, әсіресе жоғары дәлдік аса маңызды болған жағдайда. Бұдан бөлек, әдістеме Faster R-CNN моделін нақты уақытта немесе шын уақыт режимінде қолдану үшін қажетті есептеу ресурстарын талдауды қамтиды. Бұл әзірленген әдісті диагностикалық тиімділігі жоғары медициналық мекемелерге интеграциялау мүмкіндігін қарастырған кезде маңызды аспект болып табылады. Faster R-CNN нәтижелерін басқа машиналық оқыту әдістерімен және маммографиядағы дәстүрлі тәсілдермен салыстыру сүт безі патологиясын анықтаудағы осы әдістің артықшылықтарын айқындайды. Ұсынылған тәсіл диагностика сапасын айтарлықтай жақсартуға және сүт безі ауруларын ерте анықтау арқылы тиімділігін арттыруға мүмкіндік береді.

**Нәтижелер және талқылау.** Бұл жұмыста біз сүт безі аномалияларын анықтау үшін маммография деректер базасында оқытылған екі негізгі модельді: Faster R-CNN және YOLOv8-ді қарастырамыз. Осы модельлердің жаттығу нәтижелері жаттығу және валидацияның жоғалуы, дәлдік және F1-өлшемі сияқты көрсеткіштер түрінде тіркелді. Faster R-CNN-ның оқыту және валидация жоғалтуының графигі оқыту дәуірі барысында деректер базасындағы жоғалтудың қалай төмендейтінін көрсетеді. Бұл модельдің қателерді азайта отырып, жалпылау қабілетін сәтті жақсарып жатқанын білдіреді. Қисықтардың конвергенциясы, сондай-ақ, оқу және валидация деректер базасында жоғары дәлдік пен F1 көрсеткішінің байқалатындығы артық сәйкестіктің жоқтығын көрсетеді. Faster R-CNN үшін оқыту және тестілеу дәлдіктері әр дәуір сайын тұрақты түрде өсіп, жаттығудың соңында шамамен 0,99 мәніне жетеді. Бұл модель деректердегі модельлерді сәтті танып, болжамдарының барған сайын дәлірек болып келе жатқанын көрсетеді. Осыған ұқсас жетістіктер 3-суретте көрсетілген F1 ұпайында байқалады, ол дәлдік пен еске түсіру арасындағы тепе-теңдікті көрсетіп, 0,97-ге жақындайды.

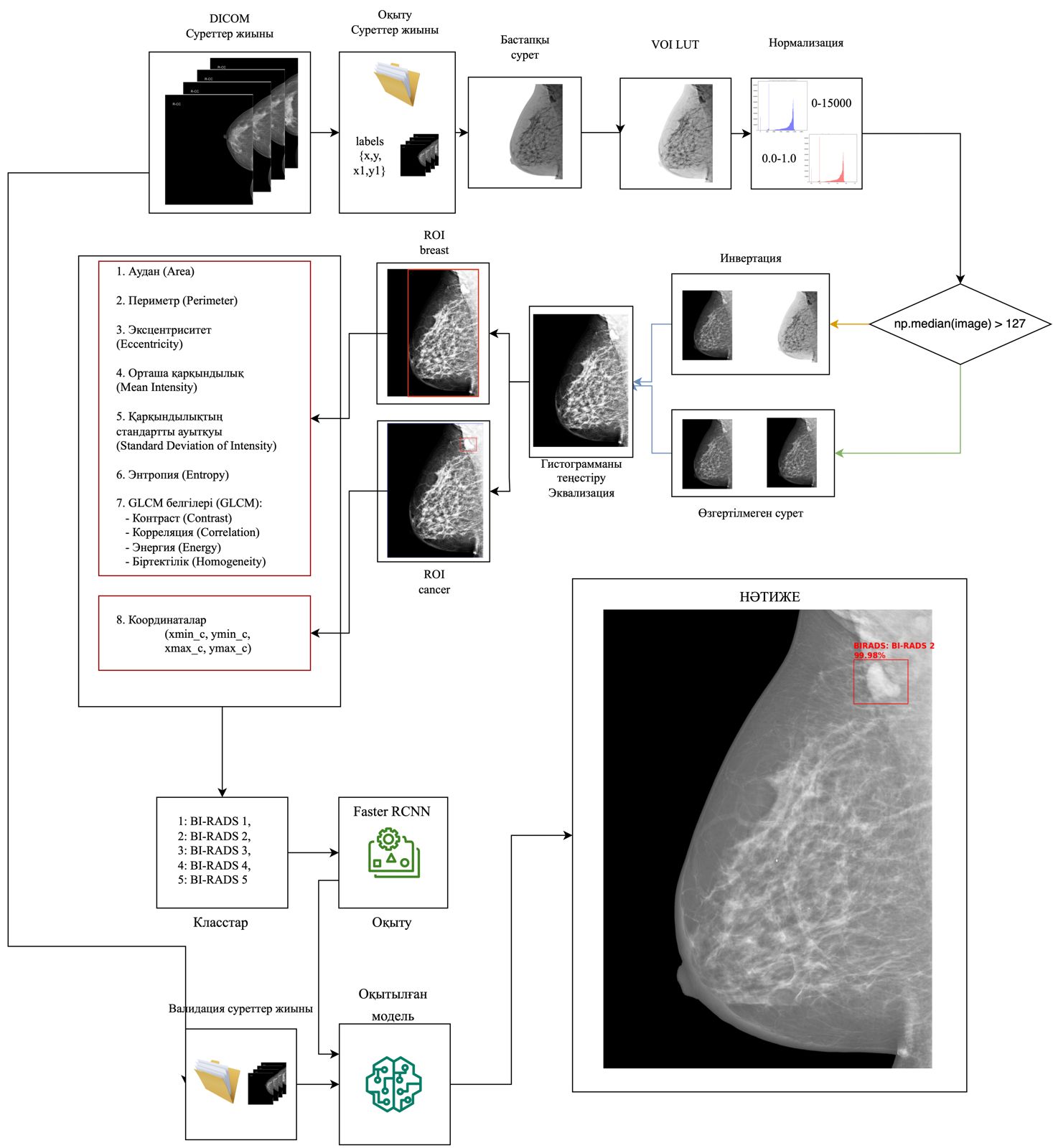


**3-сурет. Faster R-CNN моделін оқыту нәтижелері**

Faster R-CNN сияқты, YOLOv8 де оқу процесінің табысты көрсеткіштерін көрсетеді. Жаттығу және валидация деректеріндегі жоғалту графигі бастапқы кезеңдерде жоғалтудың жылдам төмендеуін көрсетіп, модельдің жоғары конвергенция жылдамдығын білдіреді. Жаттығу және валидация деректер базасының дәлдігі мен F1-өлшемі, 4-суретте көрсетілгендей, сәйкесінше шамамен 0,97 және 0,96 мәндеріне жетіп, дәйекті түрде жақсарады. Екі модельні салыстыра отырып, олардың екеуі де жоғары өнімділікті көрсететінін атап өту маңызды. Алайда, Faster R-CNN YOLOv8-мен салыстырғанда, әсіресе оқудың соңғы кезеңдерінде дәлдік пен F1-өлшеу мәндерінің аздап жоғары екенін айту керек. 5-суретте көрсетілген модель патологияның әртүрлі деңгейлері бар 1318 кескінді пайдалана отырып, сүт безі қатерлі ісігін жіктеуге арналған. Кескінді өңдеу процесі патологиялық өзгерістерді анықтау үшін қалыпқа келтіру, туралау және бүркемелеу қадамдарын қамтиды. YOLOv8 және Faster R-CNN әдістерінің тиімділігін бағалау үшін патологияның әртүрлі формаларын қамтитын 1318 кескіннен тұратын ауқымды деректер жинағы қолданылды. Деректер жинағы қатерлі және қатерсіз кескіндер арасында теңгерілген және сүт безі қатерлі ісігінің бес деңгейлі жіктелуін ұсынды. Бұл тәсіл сүт безінің медициналық кескіндеріндегі патологияларды анықтау үшін екі модельдің тиімділігін неғұрлым толық және объективті бағалауға мүмкіндік береді.



**4-сурет. YOLOv8 модельсі бойынша оқыту нәтижелері**



**5-сурет. Сүт безі қатерлі ісігі патологиясын анықтау моделі**

Сүт безінің кескінін өңдеуге және жіктеуге арналған бұл әдіс тіндердің денсаулығын жүйелі және жан-жақты бағалауды қамтамасыз етеді, бұл медициналық қолданбалар үшін ерекше маңызды. Қалыпқа келтіру және туралау әдістерін қолдану, сондай-ақ YOLOv8 және Faster R-CNN алгоритмдерінің пайдаланылуы модельдің әртүрлі патологияларды анықтауда жоғары сезімталдық пен дәлдікті көрсетуіне мүмкіндік береді. Әртүрлі сыныптардан тұратын теңгерілген деректер базасы сүт безі патологиясының әртүрлі формалары мен деңгейлерін ескере отырып, модельдің өнімділігін неғұрлым өкілді бағалауға көмектеседі. Бұл тәсіл жіктеу нәтижелерінің сенімділігін арттырады және сүт безі қатерлі ісігінің диагностикасын айтарлықтай жақсартуға ықпал етеді.

**Қорытынды.** Ұсынылған зерттеу маммографиялық кескіндердегі сүт бездерінің патологиясын анықтау үшін YOLOv8 және Faster R-CNN терең оқыту модельдерін тиімді пайдалану мүмкіндігін көрсетеді. Екі модель де дәлдік, F1 ұпайы және өңдеу жылдамдығы бойынша жоғары нәтижелер көрсетті, бұл олардың медициналық салада практикалық қолдану әлеуетін айқындайды. Faster R-CNN моделі, әсіресе, жаттығудың соңғы кезеңдерінде YOLOv8-ден дәлдік пен F1 ұпайы бойынша тұрақты түрде жоғары болды, шамамен 0,99 дәлдік және 0,97 F1 ұпайына қол жеткізіп, сүт безі ауытқуларын анықтаудағы сенімділігін көрсетті. YOLOv8, сәл төменірек дәлдікке ие болғанымен, жоғары өңдеу жылдамдығын көрсетті, бұл оны жылдам талдау талап етілетін қолданбалар үшін қолайлы етеді, ал оның дәлдігі шамамен 0,97, F1 ұпайы 0,96 болды, бұл оның сүт безі патологиясын анықтаудағы тиімділігін растайды. Деректерді алдын ала өңдеу, оның ішінде қалыпқа келтіру, гистограмманы теңестіру және аномалия аймақтарын маскалау, оқу процесінің тұрақтылығына және сүт бездерінің патологияларын анықтауда модельлердің сенімділігіне елеулі әсер етті. Екі модель де патологияның әртүрлі деңгейлерін қамтитын 1318 маммографиялық кескіннің теңгерілген деректер базасында оқытылды және тексерілді, бұл олардың сүт безі ауруларының қатерсіз және қатерлі формаларымен күресу қабілетін нығайтады. Бұл модельлерді клиникалық тәжірибеге енгізу сүт безі ауруларын ерте анықтау мен диагностикалауды жақсартуға, сондай-ақ пациенттердің нәтижелерін жақсартуға мүмкіндік береді. Оқыту параметрлері мен модель архитектурасын оңтайландыруға, деректер базасын кеңейтуге және модельдерді одан әрі жетілдіруге, олардың шынайы клиникалық параметрлерде сенімділігі мен дәлдігін қамтамасыз ету үшін қосымша зерттеулер жүргізу ұсынылады.

**Әдебиеттер**

1. A. Orazayeva et al. Biomedical image segmentation method based on contour preparation // in Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments. – 2022. DOI: 10.1117/12.2657929

2. Z. He, P. Wang, and X. Ye, Novel endoscopic optical diagnostic technologies in medical trial research: recent advancements and future prospects // BioMedical Engineering OnLine. -2021. -Vol. 20(1). DOI: 10.1186/s12938-020-00845-5

3. O. Yu. Bogaevskaya, A. V Yumashev, A. L. Zolkin, O. A. Smirnova, and M. S. Chistyakov, Application of progressive information technologies in medicine: computer diagnostics and 3D technologies // Journal of Physics: Conference Series. -2021. - Vol. 1889(5). DOI: 10.1088/1742-6596/1889/5/052001

4. R. Boina et al. Enhancing intelligence diagnostic accuracy based on machine learning disease classification // International Journal of Intelligent Systems and Applications in Engineering. -2023. -Vol. 11. -P. 765–774, 2023.

5. C.-H. Hsu et al. Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning // Measurement. -2021. -Vol. 175. DOI: 10.1016/j.measurement.2021.109145

6. P. Singh, N. Singh, K. K. Singh, and A. Singh Diagnosing of disease using machine learning // in Machine Learning and the Internet of Medical Things in Healthcare. - Elsevier, 2021. - P. 89–111.

7. A. Orazayeva et al. Imaging fuzzy expert system for assessing dynamic changes in biomedical tumor images in breast cancer // in Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2022. -Dec. 2022. DOI: 10.1117/12.2657923.

8. A. Orynbayeva, N. Shyndaliyev, and A. Aripbayeva Improving statistical methods of data processing in medical universities using machine learning //World Transactions on Engineering and Technology Education. -2023. -Vol. 21(1). -P. 58–63.

9. Y. Liu, D. Han, A. V Parwani, and Z. Li Applications of artificial intelligence in breast pathology // Archives of Pathology & Laboratory Medicine. -2023. -Vol. 147(9). -P. 1003–1013. DOI: 10.5858/arpa.2022-0457-ra

10. Y. Li, Y. Zhang, Q. Yu, C. He, and X. Yuan Intelligent scoring system based on dynamic optical breast imaging for early detection of breast cancer // Biomedical Optics Express. -2024. -Vol. 15(3). -P. 1515–1527. DOI: 10.1364/BOE.515135

11. E. A. Rakha, G. M. Tse, and C. M. Quinn An update on the pathological classification of breast cancer // Histopathology. -2022. -Vol. 82(1). -P. 5–16. DOI: 10.1111/his.14786

12. C. Quinn, A. Maguire, and E. Rakha Pitfalls in breast pathology // Histopathology. -Vol. 82(1). -P. 140–161. DOI: 10.1111/his.14799

13. M. Ensenyat-Mendez et al. Current triple-negative breast cancer subtypes: dissecting the most aggressive form of breast cancer // Frontiers in Oncology. -2021. -Vol. 11. DOI: 10.3389/fonc.2021.681476

14. S. Iranmakani et al. A review of various modalities in breast imaging: technical aspects and clinical outcomes // Egyptian Journal of Radiology and Nuclear Medicine. -2020. -Vol. 51(1). DOI: 10.1186/s43055-020-00175-5

15. S. M. H. Hashemi, L. Safari, and A. D. Taromi Realism in action: anomaly-aware diagnosis of brain tumors from medical images using YOLOv8 and DeiT. -2024. DOI: 10.48550/arXiv.2401.03302

16. R.-Y. Ju and W. Cai Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm // Scientific Reports. -2023.-Vol. 13(1). DOI: 10.1038/s41598-023-47460-7

17. S. Pandey, K.-F. Chen, and E. B. Dam Comprehensive multimodal segmentation in medical imaging: combining YOLOv8 with SAM and HQ-SAM models // 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). - Paris, France, 2023. -P. 2584-2590. DOI: 10.1109/iccvw60793.2023.00273

18. B. W. Choi, S. Kang, H. W. Kim, O. D. Kwon, H. D. Vu, and S. W. Youn Faster region-based convolutional neural network in the classification of different parkinsonism patterns of the striatum on maximum intensity projection images of [18F] FP-CIT positron emission tomography // Diagnostics. -2021. -Vol. 11(9). DOI: 10.3390/diagnostics11091557

19. Z. Jiao et al. Deep learning for automatic detection of cephalometric landmarks on lateral cephalometric radiographs using the Mask Region-based Convolutional Neural Network: a pilot study // Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. -2024. -Vol. 137(5). -P. 554–562. DOI: 10.1016/j.oooo.2024.02.003

20. A. Orazayeva, J. Tussupov, W. Wójcik, S. Pavlov, G. Abdikerimova, and L. Savytska Methods for detecting and selecting areas on texture biomedical images of breast cancer // Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Srodowiska. -2022. -Vol. 12(2). -P. 69–72. DOI: 10.35784/iapgos.2951

21. G. Abdikerimova et al. Detection of lung pathology using the fractal method // International Journal of Electrical and Computer Engineering. -2023. -Vol. 13(6). -P. 6778–6786. DOI: 10.11591/ijece.v13i6.pp6778-6786

22. G. Abdikerimova et al. Detection of chest pathologies using autocorrelation functions // International Journal of Electrical and Computer Engineering. -2023. -Vol. 13(4). -P. 4526–4534. DOI: 10.11591/ijece.v13i4.pp4526-4534

23. Zaridze D.G., Maksimovich D.M. Prevention of malignant neoplasms. // V-Petersburg International Oncology Forum. Collection. - 2017. - No. 4 (2). DOI: 10.17650/2313805X-2017-4-2-8-25

24. Davydov M.I., Zaridze D.G. Skrining zlokachestvennyh opuholej // Vestnik FGBNU «RONC im N.N. Blohina». -2014. –T. 25. -№3-4. –S. 5-16. URL:https://cyberleninka.ru/article/n/skrining-zlokachestvennyh-opuholey/viewer [in Russian]

25. Nurmanova A., Sultanova Z.I., Y.A. Annaorazov Y.A. Faktory i ih rol' v zabolevaemosti, smertnosti, vyzhivaemosti pri rake molochnoj zhelezy // Vestnik KazNMU. -2018. -№ 1. URL: <https://cyberleninka.ru/article/n/faktory-i-ih-rol-v-zabolevaemosti-smertnosti-vyzhivaemosti-pri-rake-molochnoy-zhelezy/viewer> [in Russian]

26. Chen L. et al. Review of image classification algorithms based on convolutional neural networks //Remote Sensing. – 2021. –Vol.13(22). DOI:10.3390/rs13224712

27. N. Diniz D. et al. A deep learning ensemble method to assist cytopathologists in pap test image classification //Journal of Imaging. -2021. -Vol. 7(7). DOI:10.3390/jimaging7070111

**Авторлар туралы мәліметтер**

Тусупов Д.А. – физика-математика ғылымдарының докторы, профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, e-mail: [tussupov@mail.ru](mailto:oaris.83@gmail.com);

Оразаева А.Р. - Л.Н. Гумилев атындағы Еуразия ұлттық университетінің докторанты, Ақпараттық жүйелер мамандығы, Астана, Қазақстан, e-mail: oar\_is@mail.ru.

**Information about the authors**

Tussupov J.A. - doctor of physical and mathematical sciences, professor, L.N. Gumilev Eurasian National University, Astana, Kazakhstan, e-mail: tussupov@mail.ru;

Orazayeva A.R. - L.N. Ph.D. student of Gumilev Eurasian National University, Information Systems specialty, Astana, Kazakhstan, e-mail: oar\_is@mail.ru.

МРНТИ 81.93.29

**ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ИДЕНТИФИКАЦИИ ГОЛОСА С УЧЕТОМ КАЗАХСКОГО АКЦЕНТА**

**1,2 Т.С. Шорманов, 2А.Т. Мазакова, 1 М.С. Алиаскар, 2Ш.А. Джомартова,**

**1,2Т.Ж. Мазаков🖂**

1Международный инженерно-технологический университет, Алматы, Казахстан,

2Казахский национальный университет имени аль-Фараби, Алматы, Казахстан

**🖂**Корреспондент-автор: tmazakov@mail.ru

Статья посвящена использованию нейронных сетей для идентификации человека по голосу.

Идентификация и распознавание речи с учетом акцента — это одна из сложных задач в области обработки естественного языка (NLP) и автоматического распознавания речи (ASR). Казахский акцент, как и другие региональные особенности речи, представляет собой уникальные лингвистические и акустические характеристики, которые могут влиять на эффективность традиционных моделей распознавания речи. В последние годы нейронные сети, особенно глубокие нейронные сети (DNN), становятся основными инструментами для решения таких задач, включая идентификацию голоса с учетом акцентов.

Несмотря на достижения в области нейронных сетей, задачи распознавания речи с учетом акцента остаются достаточно сложными. Вот несколько основных проблем: 1) Многообразие акцентов. В Казахстане существует несколько региональных акцентов, что требует учета разнообразных вариантов произношения. 2) Проблемы с трансляцией специфических казахских звуков. Некоторые звуки казахского языка могут быть трудны для стандартных моделей распознавания речи. 3) Отсутствие данных. Как уже было сказано, для качественного обучения нейронных сетей для казахского языка часто не хватает доступных и разнообразных наборов данных с акцентами.

В работе рассматриваются возможные подходы к биометрической идентификации: мел-частотные кепстральные коэффициенты (MFCC), языковая модель, а также анализируются преимущества и недостатки использования нейронных сетей. В ходе исследования выявлены основные проблемы, которые могут возникнуть в ходе ее решения. Дано общее описание проблемы и формулировка целей исследования.

**Ключевые слова**: биометрия, нейронная сеть, мел-частотные кепстральные коэффициенты, идентификация голоса, казахский акцент.

**ҚАЗАҚ ТІЛІНІҢ АКЦЕНТІН ЕСКЕРТПЕ ДАУЫСТЫ АНЫҚТАУ ҮШІН НЕЙРЛІК ЖЕЛІЛЕРДІ ҚОЛДАНУ**

**1,2 Т.С. Шорманов, 2Ә.Т. Мазақова, 1 М.С. Әлиасқар, 2Ш.А. Джомартова,**

**1,2Т.Ж. Мазақов🖂**

1Халықаралық инженерлік және технология университеті,

**2** Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

e-mail: tmazakov@mail.ru

Мақала адамды дауыс арқылы анықтау үшін нейрондық желілерді пайдалануға арналған.

Екпінді ескеретін сөйлеуді анықтау және тану табиғи тілді өңдеу (NLP) және автоматты түрде сөйлеуді тану (ASR) саласындағы күрделі тапсырмалардың бірі болып табылады. Қазақ екпіні басқа аймақтық сөйлеу ерекшеліктері сияқты дәстүрлі сөйлеуді тану үлгілерінің орындалуына әсер ете алатын бірегей тілдік және акустикалық сипаттарды ұсынады. Соңғы жылдары нейрондық желілер, әсіресе терең нейрондық желілер (DNN), акцентті білетін дауысты сәйкестендіруді қоса алғанда, мұндай тапсырмалар үшін негізгі құралдарға айналды. Нейрондық желілердегі жетістіктерге қарамастан, акцентті білетін сөйлеуді тану тапсырмалары өте қиын болып қала береді. Міне, бірнеше негізгі мәселелер: 1) Акценттердің әртүрлілігі. Қазақстанда бірнеше аймақтық екпіндер бар, олар айтылудың әртүрлі нұсқаларын қарастыруды талап етеді. 2) Нақты қазақ дыбыстарын таратудағы мәселелер. Кейбір қазақ дыбыстарын стандартты сөйлеуді тану үлгілері үшін түсіну қиын болуы мүмкін. 3) Мәліметтердің жетіспеушілігі. Жоғарыда айтылғандай, қазақ тіліне арналған нейрондық желілерді сапалы дайындау үшін акценттері бар қолжетімді және алуан түрлі деректер жинақтары жиі жеткіліксіз.

Мақалада биометриялық сәйкестендірудің ықтимал тәсілдері қарастырылады: мел-жиілік цестральды коэффициенттер (MFCC), тілдік модель, сондай-ақ нейрондық желілерді пайдаланудың артықшылықтары мен кемшіліктері талданады. Зерттеу барысында оны шешу барысында туындауы мүмкін негізгі проблемалар анықталды. Проблеманың жалпы сипаттамасы және зерттеу мақсаттарының тұжырымы берілген.

**Түйін сөздер:** биометрия, нейрондық желі, мел-жиілік цестральды коэффициенттер, дауысты анықтау, қазақша акцент.

**APPLICATION OF NEURAL NETWORKS FOR VOICE IDENTIFICATION TAKING INTO ACCOUNT THE KAZAKH ACCENT**

**1,2 T.S. Shormanov, 2A.T. Mazakova, 1M.S. Aliaskar, ²Sh.A. Jomartova,**

**1,2T.Zh. Mazakov🖂**

1International Engineering and Technology University, Almaty, Kazakhstan,

2Kazakh National University named after Al-Farabi, Almaty, Kazakhstan,

tmazakov@mail.ru

The article is devoted to the use of neural networks for human voice identification.

Accent-aware speech identification and recognition is one of the challenging tasks in the field of natural language processing (NLP) and automatic speech recognition (ASR). The Kazakh accent, like other regional speech features, presents unique linguistic and acoustic characteristics that can affect the performance of traditional speech recognition models. In recent years, neural networks, especially deep neural networks (DNNs), have become the main tools for solving such problems, including accent-aware voice identification.

Despite the advances in neural networks, accent-aware speech recognition tasks remain quite challenging. Here are some of the main challenges: 1) Diversity of accents. There are several regional accents in Kazakhstan, which requires taking into account a variety of pronunciation options. 2) Problems with translating specific Kazakh sounds. Some sounds of the Kazakh language can be difficult for standard speech recognition models. 3) Lack of data. As already mentioned, there is often a lack of accessible and diverse datasets with accents for high-quality training of neural networks for the Kazakh language.

The paper considers possible approaches to biometric identification: mel-frequency cepstral coefficients (MFCC), language model, and analyzes the advantages and disadvantages of using neural networks. The study identified the main problems that may arise in the course of its solution. A general description of the problem and formulation of the research objectives are given..

**Keywords:** biometrics, neural network, mel-frequency cepstral coefficients, voice identification, Kazakh accent.

**Введение**. Информационная безопасность является одной из важных задач в мировом сообществе, и это связано с резко возросшим влиянием глобальных информационных технологий на большинство сфер деятельности современного общества. Современные интернета угрозы требуют эффективных методов обнаружения и защиты данных. Учитывая современную тенденцию к более активному использованию удаленных сервисов при проведении финансовых транзакций, обеспечение кибербезопасности при проведении удаленных финансовых транзакций стало обязательно, так как при проведении удаленных финансовых транзакций появляется возможность компрометации данных, что привлекает киберпреступников. Предлагаемое в данной статье решение позволяет повысить эффективность биометрической голосовой аутентификации. Человек может быть аутентифицирован тремя способами: через знание, владение или наследование. Фактор знания относится к тому, что человек должен иметь для получения доступа. Аутентификация на основе предопределенного значения (пароля) является самым распространенным методом аутентификации. Право собственности относится к учетным данным аутентификации пользователя на основе предметов, которыми владеет пользователь, как правило, аппаратных устройств, таких как мобильный телефон пользователя или токен безопасности. С точки зрения внутренних факторов наиболее распространенным методом является аутентификация на основе биометрии, которая использует отпечатки пальцев, голос или распознавание лиц для аутентификации пользователей. Среди этих биометрических методов речь является основным способом общения между людьми. Пример возросшего использования голоса в мае 2021 года HSBC сообщил, что HSBC Voice ID защитил своих клиентов от телефонных мошенников в 2020 году. По данным банка, благодаря новой биометрической системе количество случаев мошенничества сократилось более чем на 50%. В целом HSBC считает, что его система голосовой биометрии предотвратила кражу примерно 249 миллионов фунтов стерлингов (346,5 миллиона долларов США) денег клиентов [1]. Индивидуальные характеристики голоса являются результатом влияния нескольких характеристик: анатомо-физиологических особенностей (особенности строения рта говорящего, например, наличие или отсутствие зубов и т. д.); артикуляционных особенностей говорящего (индивидуальные особенности произношения, например, картавость или шепелявость). Сейчас существуют много сервисов по созданию поддельных голосовых копий, таких как Elevenlabs.io и другие. Исходя из вышеперечисленных фактов, для повышения качества биометрической идентификации по голосу необходимо анализировать не только акустические параметры голоса, но и анализировать и распознавать индивидуальные характеристики речи, такие как пол, возраст, акцент; такой подход позволит повысить устойчивость системы распознавания к попыткам фальсификации голоса, что особенно важно в сфере финансовых услуг.

**Обзор литературы**. В статье [2] авторы рассматривают четыре схемы взвешивания терминов и используют десять существующих схем в сочетании с вариантами векторов tf и tf-idf для исследования их производительности на коротких текстовых наборах данных. Этот подход дает хорошие результаты для классификации небольших текстов, таких как сообщения Twitter. Текст в пространстве терминов можно смоделировать как *d*=(*w*1,. . . , *wn*), где *n* обозначает пространство терминов или размер признаков Эффективность модели часто повышается с использованием N-грамм или схем взвешивания терминов, таких как частота терминов-обратная частота документа. Недостатком метода является, что в коротких текстах таких как твиты или заголовки новостей, содержится ограниченное количество терминов, что делает их представление и взвешивание в модели векторного пространства неэффективными для классификации текста.

В статье [3] представлен подход к использованию модели векторного пространства (VSM), представление текста – это задача преобразования содержимого текстового документа в вектор в пространстве терминов, чтобы документ можно было распознать и классифицировать. Представленный подход к использованию контролируемого метода взвешивания терминов, tf.rf, показывает лучшую производительность, чем другие методы взвешивания терминов. Хотя этот подход имеет некоторые недостатки, он не предназначен для работы с большими наборами данных и методами взвешивания терминов для неструктурированных текстовых данных или использования внешних знаний для улучшения оценки важности терминов.

Статья [4] представляет альтернативный подход, основанный на векторном представлении слов. Преимуществом метода «встраивания слов»(word embeddings) заключается в способности улавливать сходства в классификации значений слов. Точность процедуры оценивается с помощью кодированных обучающих предложений; валидация проводилась через изучение негативности в речах австрийского парламента. Результаты показали потенциал подхода «встраивания слов» для анализа настроений для небольших текстов, но в больших текстах метод улавливает семантические сходства между словами, но он не может полностью учитывать контекст и порядок слов в предложении. Это ограничивает возможности моделей для анализа настроений и нюансов, особенно для больших текстов.

В статье [5] представлен алгоритм кодирования пар байтов (BPE алгоритм), это способ получения токенов, поскольку ранее языковая модель использовала только слова из заранее представленного словаря, но в представленной статье использовался другой подход к методам сегментации слов, который заключается в разделении слов на более мелкие подслова или слоги. Представленный подход помогает решить проблемы с редко используемыми словами а также с ограничением по лексике, чем использование языковой моделью обученной только на основании слов, которые были в обучающем корпусе. Однако этот подход имеет трудности со словами, которые не встречались в обучающей выборке данных. Это может ограничить практическую применимость моделей машинного перевода, особенно для языков с богатым словарным запасом или с новыми словами и текстами.

В статье [6] представлен алгоритм WordPiece для обучения токенов на разбиение слов на обучающими подслова или слоги с добавлением специальных символов в начале каждого токена, чтобы показать, что это часть большого слова. Этот процесс помогает анализировать отдельные символы и контекст больших слов, данный подход позволяет лучше обобщать новые слова и уменьшить размер обучающей выборки слов. Алгоритм может быть применен к различным языкам, что делает его универсальным для многоязычных приложений. Алгоритм WordPiece имеет недостаток связанный со значительными вычислительными требованиями, это может быть ограничивающим фактором для практического использования, особенно для задач реального времени или очень больших текстов.

В статье [7] BERT представлен как многослойный двунаправленный кодер Transformer. Эта языковая модель, основанная на архитектуре трансформер, использует двунаправленный внутренний анализ. Модель используется совместно с классификатором, входные данные которого являются результатом работы BERT – векторным представлением входных данных. В основе обучения модели лежат две идеи. Первая – заменить 15 % слов масками и обучить сеть предсказывать эти слова. Вторая – дополнительно обучить BERT определять, может ли одно предложение следовать за другим. Архитектура Transformer состоит из различных компонентов, которые были описаны в [5]: Токенизаторы, которые преобразуют текст в токены. – один слой внедрения, который преобразует токены и позиции токенов в векторные представления. – слои преобразователя, которые выполняют повторяющиеся преобразования векторных представлений, извлекая все больше и больше лингвистической информации. Они состоят из чередующихся слоев внимания и прямой связи. – (необязательно) слой деинсталляции, который преобразует окончательные векторные представления обратно в распределение вероятностей по токенам. Предварительно обученные модели, такие как BERT, имеют ограничения, связанные с обучением на огромных корпусах данных, который могут содержать смещения и ограничения, отражающие данные, на которых они обучались. Это может привести к тому, что модели примут те же смещения и ограничения. Пример таких смещений и ограничений представлен в статье [8] как результат обучения многоязычной языковой модели BERT, анализ для греческого и испанского языков и обнаружено, что в некоторых случаях языковая модель предпочитает использовать более англоязычную настройку (явные местоимения и порядок субъект-глагол-объект) по сравнению с одноязычной моделью языка управления. Это исследование показывает, что многоязычная модель BERT имеет недостатки, связанные с доминированием английского языка, поскольку модель изучалась на данных из нескольких языков, но данные из английского языка доминировали, и это повлияло на результаты на других языках.

Для решения вопросов, связанных с голосовой идентификацией, необходимо использовать не только акустический голос для идентификации, а проводить распознавание речи с учетом особенностей конкретного человека. Помимо физиологических особенностей голоса существует фонетическая особенность речи, которая указывает на региональную, этническую или иноязычную принадлежность человека. Каждый человек обладает определенным словарным запасом, этот словарный запас определяется его социальной и ментальной средой. Особенности речи, голоса, интонации, а также манера речи, сформированные в юности, сохраняются на протяжении всей жизни и имеют набор определенных характеристик, присущих только им [9]. Анализируя отдельные элементы речи, можно определить индивидуальный стиль речи человека. Однако такой подход имеет ряд недостатков, для анализа индивидуальных голосовых характеристик речи необходимо собрать достаточно большую базу данных речи конкретного человека.

Поскольку в Республике Казахстан два основных языка общения – казахский и русский, то при разработке необходимо учитывать особенности и акценты, используемые в Республике Казахстан. Акцент – фонетическая особенность речи, связанная с региональной, этнической или иноязычной принадлежностью человека. Человек с акцентом может неправильно произносить некоторые звуки или изменять их непосредственно в словах. Особенности казахского акцента в русском языке. Акцент – это изменение произношения звуков, распространенное искажение. Акцент может привести к замене или искажению определенных звуков. К таким звукам относятся, например, «ы» и «и» в русском языке. Еще одной распространенной проблемой является использование звука «у», так как его произношение в казахском языке отличается. Интонация – еще один признак акцента. Казахский язык имеет свою собственную интонацию, которая меняется в зависимости от культурного контекста, и эти интонации появляются в других языках. Ударение в русском языке – это выделение гласного звука в слове; оно может быть как в середине слова, так и в конце; в казахском ударении чаще всего используется ударение на конце слова, по аналогии с казахским языком. Фонетические особенности ударения: измененное произношение звуков; часто изменения в произношении звуков; в первую очередь, это безударные гласные после твердых шипящих «ш», «ж» и «ц». Использование собственных казахских звуков. В казахском языке есть специфические звуки, которых нет в русском языке, это 9 звуков, которые часто используются в качестве замен русских звуков и существенно влияют на произношение [10]. Таким образом, целью данного исследования является совершенствование способов анализа коротких аудиозаписей говорящего с целью выявления наличия казахского акцента для повышения качества биометрической аутентификации.

**Материалы и методы**. При работе использованы аудиозаписи. При проведении исследования в статье использованы методы машинного обучения, основанный на применении нейронных сетей.

**Результаты и обсуждение.** Для решения данной задачи предлагается использовать двухэтапный процесс анализа аудиозаписей с целью выявления наличия акцентов. Первый этап получение акустической модели говорящей персоны, второй этап получение языковой модели говорящей персоны.

Первый этап получения голосовой модели – получение голосовой модели, но основании анализа вышеперечисленной литературы наиболее оптимальный метод для получения голосовой информации является Мел-частотные кепстральные коэффициенты (MFCC). Это метод, который позволяет анализировать частотный контент аудиозаписи путем разложения звука на мел-частотные компоненты.

Таким образом, можно анализировать как частотные, так и временные характеристики звука. Акустическая модель – преобразование звука в удобный для использования формат Мел-частотные кепстральные коэффициенты (MFCC) являются одним из наиболее распространенных и эффективных способов представления речевых данных. Этот подход с использованием мел-кепстральных коэффициентов был представлен в работе [11]. Поскольку разница в тембрах разных голосов описывается различными частотными спектрами. Математический аппарат для анализа частотного спектра – это преобразование Фурье, как способ описания сложной звуковой волны с помощью спектрограммы. Обработка данных, представляющих числовые амплитудно-временные зависимости. Учитывая особенности человеческого слуха, а именно его нелинейную природу по отношению к восприятию звуковых частот, иными словами, люди гораздо лучше различают небольшие изменения высоты звука на низких частотах, чем на высоких.

Для этой задачи используется преобразование из шкалы Герца в шкалу мел (мел-психофизическая единица высоты тона). Для нашей задачи [12] идентификации человека по голосу очень важен охват всех частот, поскольку, как было сказано во введении, для задач идентификации очень важно выявить все характеристики человека, связанные с вокальностью и высотой тона речи конкретного человека. Также преимуществом алгоритма MFCC стало то, что он уже имеет реализацию во многих языках программирования.

Второй этап: Получение языковой модели. Для обучения модели языка целесообразно использовать модель языка BERT. Модель языка BERT была представлена [13] на основе архитектуры transformer и предназначена для предобучения языковых представлений (для задач NPL). Главной особенностью этой модели языка является ее двунаправленность; эта модель языка также была доработана и может послужить основой для других алгоритмов. BERT использует четыре основных типа специальных токенов для обозначения начала и конца предложения или продолжения.

1) [CLS]: обозначает начало предложения.

2) [SEP]: обозначает конец предложения.

3) [UNK]: обозначает неизвестную часть/слово в предложении.

4) [PAD]: используется для заполнения.

Ввод данных состоит из трех этапов:

1) Реализации токенов

2) Реализация сегментов

3) Реализация позиционирования.

Токенизатор подслов BERT: BERT использует токенизатор подслов, который разбивает слова на более мелкие единицы, чем слова (слоги). Токенизатор, используемый BERT, – это токенизатор WordPiece, который представляет собой алгоритм, процесс создания окончательного набора слов (словаря) путем слияния слогов (подслов) из букв.

Обучение BERT проводится в два этапа: этапы предварительной подготовки и тонкой настройки Этап предварительной подготовки выполняется одновременно по двум задачам: Next Sentence Piece (NSP) и Masked Language Model (MLM). Masked Language Model (MLM) Для предварительной подготовки BERT случайным образом маскирует 15 % слов из входного текста с помощью токена [MASK]. Этот подход используется только для предварительной подготовки, а не для тонкой настройки. Выполняя задачу сопоставления токенов [MASK], BERT развивает способность понимать контекст.

Next Sentence Piece (NSP) После отправки двух предложений BERT обучается на них, сравнивая, является ли данное предложение одним непрерывным предложением или нет. Для этого BERT использует два фактически связанных предложения и два случайно связанных предложения в соотношении 50:50. Эта маркировка использует специальный токен [CLS], который помещается в начало предложения, чтобы определить, являются ли два предложения фактически последовательными предложениями.

Тонкая настройка BERT – это этап тестирования уже предварительно обученного BERT путем его дополнительного обучения для заданных задач, которые мы хотим решить. Для наших задач, тонкая настройка языковой модели BERT включает использование готовых уже обученных моделей BERT на основе русского языка с расширением словаря с учетом казахских слов и казахского акцента. Для анализа мы сравниваем две языковые модели BERTmultilingual и kazakhBERTmulti [14] для целей распознания речи на 15 коротких аудиозаписей. По результатам анализа для определения казахского акцента для более точной оценки необходимо исключить проверку орфографии, так как казахский акцент в основном проявляется в виде неправильно произнесенных слов. Для определения акцента для наших задач мы убрали проверку на орфографические ошибки, а именно неправильно произнесенные слова, чем больше таких слов было выявлено, тем более выраженным был акцент в аудиозаписи. Результаты, полученные с многоязычной моделью BERT и моделью kazakhBERTmulti, приведены в таблице 1.

**Таблица 1- Сравнение результатов**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Recall | Precision | F1-score |
| BERTmultilingual | 0.92 | 0.77 | 0.85 | 0.64 |
| kazakhBERTmulti | 0.92 | 0.84 | 0.92 | 0.68 |

Для multilingual BERT Russian F1 Score равен 0.64, для kazakhBERTmulti F1 Score равен 0.68 что показывает, что языковая модель работает достаточно хорошо на тестовой выборке.

**Преимущество**. Предлагаемый двухэтапный подход к извлечению голосовых признаков с помощью сбора аудиоданных MFCC представляет собой набор аудиопризнаков, извлеченных путем математической обработки частот. Это обеспечивает удобство входящего сигнала с потерей неважной информации. С их помощью создается упрощенная модель звуков речи, она также похожа на подслова или слоги, которые используются для языковых моделей BERT.

**Недостатки**. Использованный подход показывает приемлемые результаты по распознанию для коротких аудиозаписей, использованием данного подхода для анализа больших аудио записей требует использование другим алгоритмов. Ограничения голосовой модели. Алгоритм MFCC плохо работает с тоновыми языками, а особенностью казахского акцента является то, что ударения расставляются по-другому, в отличие от русского языка, и такие особенности плохо различаются в выходных файлах.

**Выводы**. Использование предложенного метода анализа позволяет получить четкую информацию о тексте речи. Все аудиозаписи были правильно распознаны для одного голоса. Применение нейронных сетей для биометрической идентификации по голосу дает значительное преимущество, поскольку позволяет проводит идентификацию не только физиологической характеристики такой как голос человека, но и дополнительно проверять индивидуальные голосовые поведенческие характеристики.

***Финансирование.*** *Работа выполнена за счет средств НИИ математики и механики при КазНУ имени аль-Фараби и грантового финансирования научных исследований на 2023–2025 годы по проекту AP19678157.*

**Литература**

1.HSBC's Voice ID prevents £249 million of attempted fraud. [Электронный ресурс] URL: <https://www.finextra.com/newsarticle/37989/hsbcs-voice-id-prevents-249-million-of-attempted-fraud>

2. Samant S.S., Bhanu Murthy N.L., Malapati A. Improving term weighting schemes for short text classification in vector space model.//IEEE Access.-2019.- Vol.7. - P.166578 -166592.

DOI [10.1109/ACCESS.2019.2953918](https://doi.org/10.1109/ACCESS.2019.2953918)

3. Lan M., Tan C.L., Su J., et al. Supervised and traditional term weighting methods for automatic text categorization. // IEEE Trans Pattern Anal Mach Intell.- 2009. - Vol.31(4 ). - P.721 - 735.

DOI [10.1109/TPAMI.2008.110](https://doi.org/10.1109/tpami.2008.110)

4. Rudkowsky E., Haselmayer M., Wastian M., Jenny M., Emrich Š., & Sedlmair M. More than Bags of Words: Sentiment Analysis with Word Embeddings. // Communication Methods and Measures. – 2018.- Vol. 12(5).- P.140 - 157. DOI [10.1080/19312458.2018.1455817](http://dx.doi.org/10.1080/19312458.2018.1455817)

5. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units. // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany. Association for Computational Linguistics.- 2016.-Vol.1.- P.1715-1725. DOI [10.18653/v1/P16-1162](https://doi.org/10.18653/v1/P16-1162)

6. Wu Y., Schuster M., Chen Z., Le Q. V., Norouzi M., Macherey W., ... & Dean J. Google's neural machine translation system: Bridging the gap between human and machine translation. // arXiv preprint arXiv:1609.08144.2016. DOI [10.48550/arXiv.1609.08144](http://dx.doi.org/10.48550/arXiv.1609.08144)

7. Jacob D., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // North American Chapter of the Association for Computational Linguistics. - 2019. - P. 4171-4186. DOI [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)

8. Papadimitriou I., Lopez K., Jurafsky D. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. // arXiv preprint arXiv:2210.05619.- 2023.- P.1194-1200. DOI [10.18653/v1/2023.findings-eacl.89](https://doi.org/10.18653/v1/2023.findings-eacl.89)

9. Россинская Е.Р. Профессия – эксперт. Введение в юридическую специальность. – М.: Юрист, 1999. - 192 с. Rossinskaya, E. (1999). Professiy – expert. Vvedenie v uridicheskuu spechialnoct. <http://lawlibrary.ru/izdanie20672.html>. [ in Russian]

10. Основные черты казахского акцента в русской речи в области вокализма и суперсегментной фонетики. Электронный ресурс] URL:

<https://rcdo.kz/publ/9555-osnovnye-cherty-kazahskogo-akcenta-v-russkoy-rechi-v-oblasti-vokalizma-i-supersegmentnoy-fonetiki.html>. [in Russian]

11. Davis S., Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences // IEEE Transactions on Acoustics, Speech, and Signal Processing. -1980. -Vol. 28, N4.- P. 357- 366. DOI 10.1109/TASSP.1980.1163420

12. Aliaskar M., Mazakov T., Mazakova A., Jomartova S., Shormanov T. Human voice identification based on the detection of fundamental harmonics // IEEE 7th International Energy Conference (ENERGYCON). - 2022.- P. 1-4. DOI 10.1109/ENERGYCON53164.2022.9830471

13. Li F., Jin Y., Liu W., et al. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study.// JMIR Med Inform. - 2019. -Vol. 7(3) - P.1-13. DOI [10.2196/14830](https://doi.org/10.2196/14830)

14. Mazakov T.Zh., Jomartova Sh. A., Shormanov T. S., Ziyatbekova G. Z., Amirkhanov B. S., Kisala P . The image processing algorithms for biometric identification // News of the national academy of sciences of the Republic of Kazakhstan. Series of Geology and Technical Sciences.- 2020.-Vol.1(439). – P.14-22. DOI 10.32014/2020.2518-170Х.2

***Сведение об авторах***

Шорманов Т.С. – докторант КазНУ имени аль-Фараби, старший преподаватель МИТУ, Алматы, Казахстан, e-mail[shormanov@gmail.com](mailto:shormanov@gmail.com);

Мазақова А.Т. – докторант КазНУ им.аль-Фараби, e-mail: [aigerym97@mail.ru](mailto:aigerym97@mail.ru);

Әлиасқар М.С. - старший преподаватель МИТУ, Алматы, Казахстан, e-mail: [m.alyasqar@gmail.ru](mailto:m.alyasqar@gmail.ru);

Джомартова Ш.А. - доктор технических наук, доцент, КазНУ им. аль-Фараби, Алматы, Казахстан, e-mail: jomartova@mail.ru;

Мазаков Т.Ж. – доктор физико-математических наук, профессор, КазНУ им. аль-Фараби, Алматы, Казахстан, e-mail: tmazakov@mail.ru.

***Information about the authors***

Shormanov T.S. - PhD student of the Al-Farabi Kazakh National University, Lecturer at the International University of Engineering and Technology, Almaty, Kazakhstan, e-mail: [shormanov@gmail.com](mailto:shormanov@gmail.com);

Mazakova A.T. - PhD student of the Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [aigerym97@mail.ru](mailto:aigerym97@mail.ru);

Aliaskar M.S. - Lecturer at the International University of Engineering and Technology, Almaty, Kazakhstan, e-mail: [m.alyasqar@gmail.ru](mailto:m.alyasqar@gmail.ru);

Jomartova Sh.A. - Doctor of Technical Sciences, Associate Professor, Al-Farabi Kazakh National University,Almaty, Kazakhstan, e-mail: jomartova@mail.ru;

Mazakov T.Zh. – Doctor of Physical and mathematical sciences, professor, Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: [tmazakov@mail.ru](mailto:tmazakov@mail.ru)

IRSTI 81.93.29

**THE STUDY OF MACHINE AND DEEP LEARNING MODELS FOR MALWARE CLASSIFICATION**

**1,2A.Zhumabekova🖂, 1,3O.Ussatova, 1M.Kalimoldayev, 1,2V.Karyukin, 1**,**3Y Begimbayeva**

1Institute of Information and Computational Technologies, Almaty, Kazakhstan,

2Al-Farabi Kazakh NationalUniversity, Almaty, Kazakhstan,

3G.Daukeev Almaty University of Energy and Communications, Almaty, Kazakhstan

**🖂**Correspondent-author: [zhumabekova2702@gmail.com](mailto:zhumabekova2702@gmail.com)

The rapid growth of cyber threats and attacks has highlighted the need for robust information security, confidentiality, and integrity measures. Malware, a significant category of cyber threats, is designed to disrupt operations, damage information environments, and gain unauthorized access to systems, networks, and data. Various types of malware, including viruses, worms, trojans, spyware, and rootkits, pose pervasive and evolving dangers, often spread through the Internet or removable devices. While effective against known threats, traditional signature-based detection methods struggle to identify new malware. Modern machine learning-based approaches offer a more flexible solution by learning from large datasets without relying on predefined signatures. This research presents a machine learning-based malware detection system using a dataset of diverse network threats. The study explores both classical machine learning algorithms and advanced deep learning models, including dense neural networks (DNN), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRU), to enhance malware detection accuracy. Moreover, a newly developed hybrid LSTM-GRU deep learning model was utilized for classifying the malware dataset. This model combines the strong specifications of both LSTM and GRU neural networks. The used machine learning and deep learning models demonstrated good classification results. The Decision Tree, Random Forest, and XGBoost machine learning models were superior to neural networks by around 0.02. The experiments showed that machine learning algorithms are still strong in the classification tasks of the cybersecurity field. Among neural networks, the simple DNN model was a little worse than LSTM and GRU by around 0.01. The recurrent LSTM and GRU models showed mostly identical scores. The proposed LSTM-GRU model outperformed other deep learning models by 0.01 and was comparable with the Random Forest model that reached the metrics score of 0.99.

**Keywords**: malware, information security, threat detection, machine learning, deep learning, Chi-square, class balancing

**ЗИЯНДЫ БАҒДАРЛАМАЛАРДЫ КЛАССИФИКАЦИЯЛАУ ҮШІН МАШИНАЛЫҚ ЖӘНЕ ТЕРЕҢ ОҚУ МОДЕЛЬДЕРІН ЗЕРТТЕУ**

**1,2А.Жұмабекова🖂, 1,3О.Усатова, 1М.Қалимолдаев, 1,2В.Карюкин, 1,3Е.Бегимбаева**

1Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан,

2әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан,

3Ғ. Дәукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан,

e-mail: zhumabekova2702@gmail.com

Киберқауіптер мен шабуылдардың жылдам өсуі сенімді ақпараттық қауіпсіздік, құпиялылық және тұтастық шараларының қажеттілігін көрсетті. Зиянды бағдарламалар, киберқауіптердің маңызды санаты ақпараттық орталарды бұзуға, зақымдауға және жүйелерге, желілерге және деректерге рұқсатсыз кіруге арналған. Вирустар, құрттар, трояндар, шпиондық бағдарламалар мен руткиттерді қоса алғанда, зиянды бағдарламалардың әртүрлі түрлері Интернет немесе алынбалы құрылғылар арқылы жиі таралатын кең таралған және дамып келе жатқан қауіп болып табылады. Қолтаңбаға негізделген анықтаудың дәстүрлі әдістері белгілі қауіптерге қарсы тиімді болғанымен, олар жаңа зиянды бағдарламаны анықтауда қиындықтарға тап болады. Заманауи машиналық оқыту модельдері алдын ала анықталған қолтаңбаларға сүйенбестен үлкен деректер жиынынан үйрену арқылы икемді шешім ұсынады. Бұл зерттеу әртүрлі онлайн қауіптердің деректер жинағын пайдалана отырып, машиналық оқытуға негізделген зиянды бағдарламаларды анықтау жүйесін ұсынады. Жұмыс зиянды бағдарламаны анықтау дәлдігін жақсарту үшін классикалық машиналық оқыту алгоритмдерін де, тереңдетіп оқытудың кеңейтілген үлгілерін де, соның ішінде толық қосылған нейрондық желілерді (DNN), ұзақ қысқа мерзімді жад желілерін (LSTM) және қақпалы қайталанатын блоктарды (GRU) қарастырады. Сонымен қатар, зиянды бағдарлама деректер жинағын жіктеу үшін жақында жасалған гибридті терең оқыту үлгісі LSTM-GRU пайдаланылды. Бұл модель LSTM және GRU нейрондық желілерінің күшті сипаттамаларын біріктіреді. Қолданылған машиналық оқыту және терең оқыту үлгілері классификацияның жақсы нәтижелерін көрсетті. Decision Tree, Random Forest және XGBoost машиналық оқыту үлгілері нейрондық желілерден шамамен 0,02-ге асып түсті. Эксперименттер машиналық оқыту алгоритмдері киберқауіпсіздік саласындағы жіктеу тапсырмаларында әлі де күшті екенін көрсетті. Нейрондық желілер арасында қарапайым DNN моделі LSTM және GRU-дан шамамен 0,01-ге нашар болды. Қайталанатын LSTM және GRU үлгілері іс жүзінде бірдей бағалауларды көрсетті. Ұсынылған LSTM-GRU моделі тереңдетіп оқытудың басқа үлгілерінен 0,01-ге асып түсті және 0,99 метрикалық ұпайға қол жеткізген Random Forest үлгісімен салыстыруға болатын. Эксперименттер машиналық оқыту алгоритмдері киберқауіпсіздік саласындағы жіктеу тапсырмаларында әлі де күшті екенін көрсетті. Нейрондық желілер арасында қарапайым DNN моделі LSTM және GRU-дан шамамен 0,01-ге нашар болды. Қайталанатын LSTM және GRU үлгілері іс жүзінде бірдей бағалауларды көрсетті. Ұсынылған LSTM-GRU үлгісі басқа тереңдетіп оқыту үлгілерінен 0,01-ге асып түсті және 0,99 метрикалық ұпайға қол жеткізген Random Forest үлгісімен салыстыруға болатын.

**Түйін сөздер:** Зиянды бағдарламалар, ақпараттық қауіпсіздік, қауіпті анықтау, машиналық оқыту, терең оқыту, Chi-square, классты теңдестіру

**ИССЛЕДОВАНИЕ МОДЕЛЕЙ МАШИННОГО И ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ВРЕДОНОСНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ**

**1,2А.Жұмабекова🖂, 1,3О.Усатова, 1М.Қалимолдаев, 1,2В.Карюкин, 1,3Е.Бегимбаева**

¹Институт информационных и вычислительных технологий, Алматы, Казахстан,  
²Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,  
³Алматинский университет энергетики и связи имени Г.Даукеева, Алматы, Казахстан,  
e-mail: [zhumabekova2702@gmail.com](mailto:zhumabekova2702@gmail.com)

Быстрый рост киберугроз и атак выявил необходимость в надежных мерах информационной безопасности, конфиденциальности и целостности. Вредоносное ПО, являясь значительной категорией киберугроз, предназначено для нарушения работы, повреждения информационных сред и получения несанкционированного доступа к системам, сетям и данным. Различные виды вредоносного ПО, такие как вирусы, черви, трояны, шпионские программы и руткиты, представляют собой все проникающие и развивающиеся угрозы, часто распространяемые через интернет или съемные устройства. Традиционные методы обнаружения, основанные на сигнатурах, эффективны против известных угроз, но сталкиваются с трудностями в идентификации нового вредоносного ПО. Современные методы машинного обучения предлагают более гибкое решение, обучаясь на больших наборах данных без необходимости использования заранее определенных сигнатур. Данное исследование представляет систему обнаружения вредоносного ПО на основе машинного обучения с использованием набора данных о различных сетевых угрозах. В работе изучены как классические алгоритмы машинного обучения, так и продвинутые модели глубокого обучения, включая полносвязные нейронные сети (DNN), сети с длинной краткосрочной памятью (LSTM) и рекуррентные блоки с управляемыми воротами (GRU), с целью повышения точности обнаружения Вредоносного ПО. Более того, для классификации датасета Вредоносного ПО использовалась недавно разработанная гибридная модель глубокого обучения LSTM-GRU. Эта модель сочетает в себе сильные характеристики нейронных сетей LSTM и GRU. Используемые модели машинного обучения и глубокого обучения продемонстрировали хорошие результаты классификации. Модели машинного обучения Decision Tree, Random Forest и XGBoost превосходили нейронные сети примерно на 0,02. Эксперименты показали, что алгоритмы машинного обучения по-прежнему сильны в задачах классификации в области кибербезопасности. Среди нейронных сетей простая модель DNN была немного хуже LSTM и GRU примерно на 0,01. Рекуррентные модели LSTM и GRU показали в основном идентичные оценки. Предложенная модель LSTM-GRU превзошла другие модели глубокого обучения на 0,01 и была сопоставима с моделью Random Forest, которая достигла оценки метрик 0,99. Эксперименты показали, что алгоритмы машинного обучения по-прежнему сильны в задачах классификации в области кибербезопасности. Среди нейронных сетей простая модель DNN была немного хуже LSTM и GRU примерно на 0,01. Рекуррентные модели LSTM и GRU показали в основном идентичные оценки. Предложенная модель LSTM-GRU превзошла другие модели глубокого обучения на 0,01 и была сопоставима с моделью Random Forest, которая достигла значений метрик в 0,99.

**Ключевые слова**: вредоносное ПО, информационная безопасность, обнаружение угроз, машинное обучение, глубокое обучение, Chi-square, балансировка классов.

**Introduction.** Nowadays, the significant growth of different types of cyber threats and attacks makes people seriously consider possessing information security [1], confidentiality [2], and integrity [3]. The specific characteristics of cyber threats belong to malware. Malware is software designed to damage the information environment, disrupt operations, and gain unauthorized system access, network, and data. Malware [4] represents pervasive and dangerous threats that evolve continuously to exploit vulnerabilities in individual systems and large-scale networks. There is a great variety of malware, including viruses (a type of malware that attaches itself to legitimate files or programs), worms (self-replicating malware that can spread across networks without needing to attach themselves to a host file or program), trojans (this malware pretends to be legitimate software and deceives users into downloading and executing them), spyware (it monitors and collects information about users, such as logins, passwords, credit card numbers, and other sensitive data), rootkits (this malware is designed to give attackers the privilege to access the system while hiding its presence), and different types of threats.

Malware [5] is usually distributed via the Internet and removable devices like flash drives. They affect systems by significantly reducing the computer’s performance, significantly reducing the free space of its HDD and SSD drives, and displaying various advertisements on the screen. This is one of the most obvious signs that the user’s computer system is infected with malware. Dangerous malware steals files containing confidential data, hides them inside the computer, and continues to perform malicious actions. In order to protect against malware, various approaches are utilized. The signature-based approach is a traditional method used by antivirus and anti-malware programs. It identifies unique patterns, or signatures, associated with the known malware. During the process of file scanning, the software compares its code to the signature stored in the database. Although this approach can detect many different types of malware, it also encounters problems with recognizing new and previously unseen malware. A modern machine learning-based method proposes a completely new way of detecting malware. It learns from large datasets and does not rely on predefined signatures, which makes it more flexible and capable of identifying previously unseen malware.

The detection of malware with the use of machine and deep learning approaches was explored in many scientific works. The paper [6] focused on classifying malware with DNN and Bi-LSTM. The performance of these two models was strong. DNN reached an accuracy score of 98%, while Bi-LSTM got 99%. In the cybersecurity field, the work [7] demonstrated a new machine-learning framework for detecting different types of malware, including Ransomware, Spyware, and Trojan Horses. The main algorithm of this framework was the Ensemble-based classifier that proved effective in handling threats and reached an accuracy score of 98%. The research paper [8] analyzes spyware detection using a decision tree machine learning algorithm. It gave an opportunity to get an accuracy score of 99%. This classifier gave an accuracy score of 97%, precision – 88.9%, recall – 88.6%, and F1-score – 88.6%. A very effective Gated Recurrent Unit (GRU) model was implemented in [9], which proved to be especially effective in detecting malicious attacks on the Internet of Things. Its use allowed to achieve a precision score of 99.5%. The work [10] is devoted to analyzing flexible and scalable methods for malware detection in Android mobile devices. The proposed machine learning approach on the DataMD dataset gave the classification accuracy of 98% and 99%.

In this paper, the malware dataset consists of various types of network threats. The machine learning approach is chosen to detect malware. Moreover, it is not restricted to only classical machine learning algorithms but also explores the classification results obtained by such deep learning models as dense neural network (DNN), Long short-term memory (LSTM) neural network, Gated recurrent unit (GRU) neural network, and proposed LSTM-GRU model.

Materials and methods. Software development for malware classification includes many significant steps crucial to this task. A suitable dataset with malicious and legitimate elements is gathered. The dataset is characterized by its appropriate specifications for analysis and classification. This repository includes the most relevant malware. When the dataset is gathered, it is necessary to do the subsequent steps where data cleaning, data normalizing, feature selection, class balancing, and classification using the machine and deep learning models are implemented. Data cleaning is significant because noisy or irrelevant data could interfere with the analysis or model performance. Data normalization refers to adjusting the values in the dataset to a common scale without distorting differences in the ranges of values. This is crucial for machine learning algorithms sensitive to input features’ scale. Feature selection leaves only the most relevant features for classification because too many irrelevant or redundant features can lead to overfitting and longer training times. Class balancing allows all classes to be equal, which is important because underrepresented classes can lead to biased models.

**Materials and methods.** The Malware dataset comprises the most relevant malware, comprising 216352 elements and 58 features. It is available by the following link (<https://github.com/saurabh48782/Malware_Classification/blob/master/MalwareData.csv>).The analysis of this dataset showed that three columns (‘ID’, ‘md5’, ‘Unnamed: 57’) are not valuable and do not carry any meaningful things. Therefore, they were removed. The other features describe more significant characteristics of the dataset. Their detailed specifications are described in the following way:

- Machine: Information about the architecture type, such as x86, x64, etc.

- SizeOfOptionalHeader: The size of the optional header in the PE (Portable Executable)

format that provides important information about the file, including entry point, stack size, etc.

- Characteristics: The bit field that indicates attributes of the file, such as if it is an executable

or a DLL.

- MajorLinkerVersion: The major version number of the linker, indicating the primary

version of the tool used for linking code.

- MinorLinkerVersion: The minor version number of the linker.

- SizeOfCode: The code section’s size that indicates the amount of space allocated for

executable code.

- SizeOfInitializedData: The size of the initialized data section that includes data already

initialized in the file.

- SizeOfUninitializedData: The size of the uninitialized data section that represents

the memory that will be allocated at runtime.

- AddressOfEntryPoint: The entry point address, marking where execution begins when the

file is loaded.

- BaseOfCode: The base address of the code section, showing where the executable code

starts in memory.

- BaseOfData: The base address of the data section, marking where initialized and uninitialized data sections start.

- ImageBase: The preferred base address of the file when loaded in memory.

- SectionAlignment: Aligning sections in memory, ensuring that sections are placed at

consistent memory boundaries.

- FileAlignment: Aligning sections in the file, ensuring consistency in the physical file

layout.

- MajorOperatingSystemVersion: The major version of the operating system required to run

the file.

- MinorOperatingSystemVersion: The minor version of the operating system required to run

the file.

- MajorImageVersion: The major version of the image or file version.

- MinorImageVersion: The minor version of the image or file version.

- MajorSubsystemVersion: The major version of the subsystem needed to run the executable,

such as the Windows GUI or console.

- MinorSubsystemVersion: The minor version of the subsystem needed to run the

executable, such as the Windows GUI or console.

- SizeOfImage: The total size of the image in memory, including headers, code, and data

sections.

- SizeOfHeaders: The size of all headers combined, providing metadata about the PE file layout.

- CheckSum: The checksum value used to verify file integrity.

- Subsystem: The specifications of the required subsystem, such as Windows GUI, console,

or device drivers.

- DllCharacteristics: The characteristics of a DLL file, including settings for security and memory handling.

- SizeOfStackReserve: The amount of memory reserved for the stack, which handles

function calls and local variables.

- SizeOfStackCommit: The amount of memory committed to the stack, ready for immediate

use.

- SizeOfHeapReserve: The amount of memory reserved for the heap, where dynamic

allocations are made.

- SizeOfHeapCommit: The amount of memory committed to the heap, allocated and ready

for use.

- LoaderFlags: The flags for the PE loader that are usually set to zero but can indicate

specific loading requirements.

- NumberOfRvaAndSizes: The number of data directory entries in the PE header.

- SectionsNb: The number of sections in the file representing a logical part of the file, such

as code or data.

- SectionsMeanEntropy: The mean entropy of all sections used to detect obfuscation or

encryption in malware.

- SectionsMinEntropy: The minimum entropy among the sections that can help to identify

highly ordered or structured data.

- SectionsMaxEntropy: The maximum entropy among the sections useful for identifying

highly randomized data.

- SectionsMeanRawsize: The average raw size of sections in the file.

- SectionsMinRawsize: The minimum raw size of any section in the file.

- SectionMaxRawsize: The maximum raw size of any section in the file.

- SectionsMeanVirtualsize: The average virtual size of sections when loaded in memory.

- SectionsMinVirtualsize: The minimum virtual size of sections when loaded.

- SectionMaxVirtualsize: The maximum virtual size of sections when loaded.

- ImportsNbDLL: The number of DLLs imported by the executable.

- ImportsNb: The total number of imports across all DLLs.

- ImportsNbOrdinal: The number of ordinal imports, which use numbers instead of names

to locate functions.

- ExportNb: The number of exports in the file, representing functions the file makes available

to other modules.

- ResourcesNb: The number of resources embedded in the file, such as icons or strings.

- ResourcesMeanEntropy: The average entropy of resources, useful for detecting

compressed or encrypted resources.

- ResourcesMinEntropy: The minimum entropy among resources.

- ResourcesMaxEntropy: The maximum entropy among resources.

- ResourcesMeanSize: The average size of resources in the file.

- ResourcesMinSize: The minimum size of resources in the file.

- ResourcesMaxSize: The maximum size of resources in the file.

- LoadConfigurationSize: The size of the load configuration table, which can specify security

and debugging settings.

- VersionInformationSize: The size of the version information, providing metadata like

version numbers and copyright.

**Data normalization.** The dataset is cleaned and normalized. In the data-cleaning phase, the elements with many missing values are removed. The duplicate records are also identified and eliminated. When the dataset elements are cleared, a more important normalization step occurs. Normalization [11] usually presents the scaling numerical data of a specific range, commonly from 0 to 1. Normalization guarantees that features with larger ranges do not dominate those with smaller ranges, making the model more stable and preventing bias in the learning process. One of the most convenient ways to normalize the values of the dataset is to utilize the Min-Max scaling technique. It is a widespread approach that efficiently normalizes the values of features. It is computed by the following formula (1):

, (1)

where  is an initial value;  is a minimum value;  is a maximum value;  is a scaled value.

**Feature selection.** The feature selection process is realized with the use of the Chi-square metric. This metric is a statistical measure used to estimate the relationship between features and the target value, providing the degree of their independence. Before calculating the statistical value of every feature, all of them are converted to the numerical form. The contingency table is created for all the features with the following formula (2):

, (2)

where  is the observed frequency, and  is the expected frequency.  is computed as (3):

 (3)

The features are ordered in a descending way by the calculated Chi-square values.

**Class balancing.** The vectorized data is checked for balancing. If the dataset is imbalanced, the special balancing technique called Random oversampling is applied to the dataset to make classes equal. This method randomly selects elements from the minority class and duplicates them to balance the dataset. The duplicated elements are added to the training set until the number of elements in the minority class equals the number in the majority class. The class balancing with the use of Random oversampling is shown in Figure 1.

Изображение выглядит как текст, снимок экрана, дизайн, линия

Автоматически созданное описание

**Figure 1-Class balancing with Random oversampling**

**Classification with machine learning and deep learning models.** When the classes are balanced, the dataset is classified with machine and deep learning models. There are several such models, among which the most popular are Decision Tree, Random Forest, XgBoost, Dense neural network (DNN), Gated recurrent unit (GRU) neural network, and Long short-term memory (LSTM) neural network.

A Decision Tree[12]is a supervised machine-learning model for the classification task. This model is structured like a tree, with internal nodes representing decisions based on feature values, branches that demonstrate the outcomes of those decisions, and leaf nodes of the final prediction and output. The topmost node of the tree presents the entire dataset, which is the starting point for decision-making and is split based on the feature that provides the best separation of the data. At the internal nodes, the dataset is split into two or more subgroups based on the chosen feature’s value. The final nodes do not further split and correspond to a class label. The structure of the Decision Tree is shown in Figure 2.

A Random Forest [13] is a machine-learning algorithm that is based on the concept of ensemble learning. It builds multiple decision trees and combines their results to improve the Accuracy of a single decision tree. The Random Forest algorithm uses bagging to create multiple training datasets by randomly sampling from the original dataset with replacement. Each decision tree is trained on a different bootstrapped dataset, ensuring the trees are diverse and uncorrelated. Generally, the Random Forest algorithm has the following specifications:

- It trains pretty quickly.

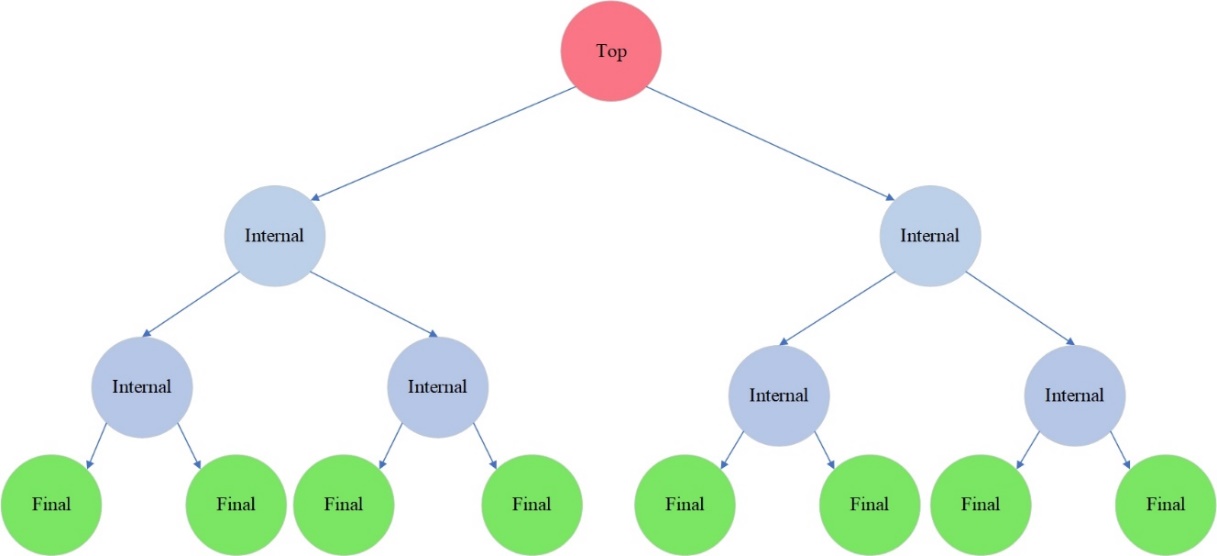
- It effectively processes datasets with a large number of features.

- It predicts data with very high Accuracy.

- It shows good efficiency even with a large number of missing data.

- It has high scalability.

The structure of Random Forest is shown in Figure 3.



**Figure 2-Decision Tree**

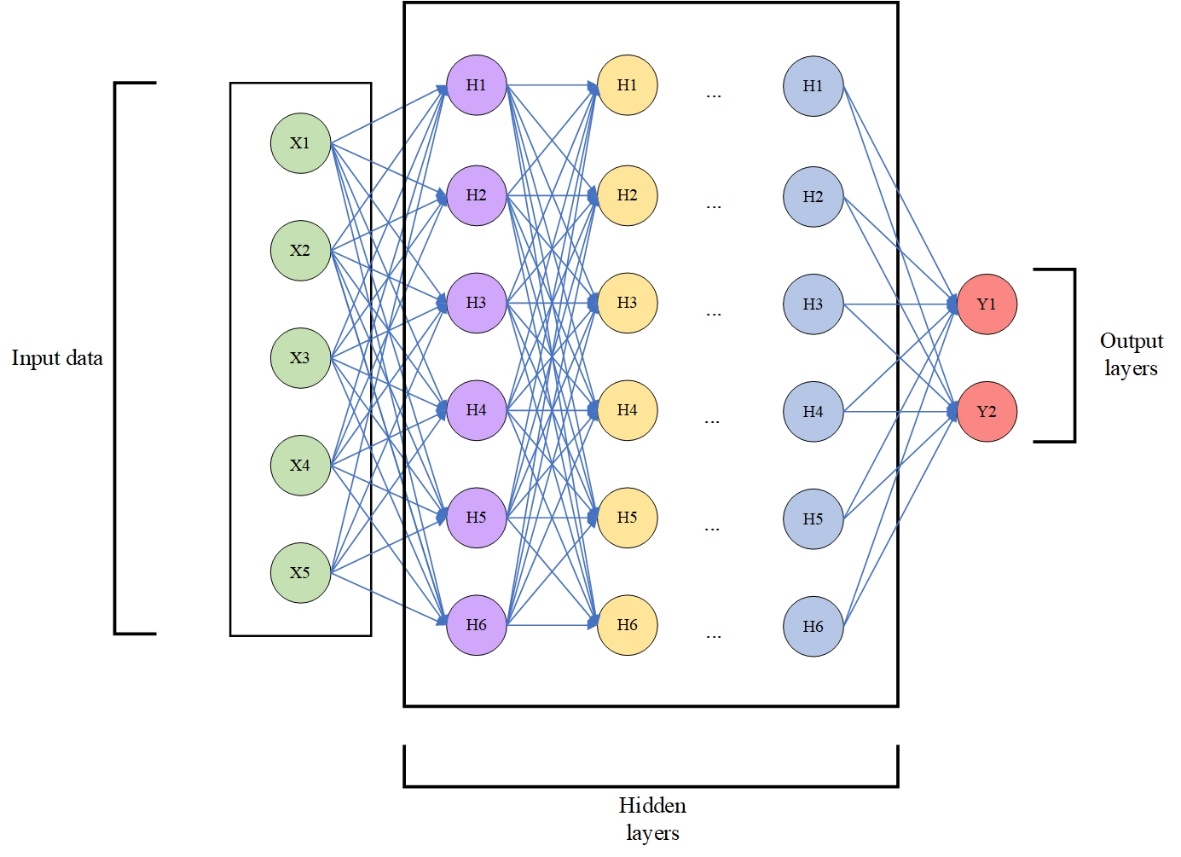
Изображение выглядит как текст, диаграмма

Автоматически созданное описание

**Figure 3 - Random Forest**

XGBoost[14]is a very strong machine-learning algorithm widely used in the classification task. It lies in a series of boosting algorithms that combine the predictions of multiple learners. XGBoost is known for its efficiency and high performance. Employing an ensemble approach corrects the errors made in previous iterations through the next model. During each step, the deviations of the ensemble’s predictions are assessed on the training data, and the model is optimized by introducing new tree forecasts to reduce the overall deviation. This process continues until the desired error threshold is met or an early stopping condition is triggered.

A Dense neural network (DNN) [15] is a fundamental type of neural network in which each neuron in one layer is connected to every neuron in the subsequent layer. It is the simplest architecture in neural networks, typically used as the basic one in deep learning. In this architecture, the input layer receives the input data , and  present weights of each level, and  show the bias of the input layer. The hidden layers consist of neurons that apply weights to the input features, followed by an activation function that produces non-linearity. There are usually many hidden layers in a neural network. The output layer shows the final prediction that usually corresponds to the number of classes: binary (two – 0 or 1) and multiclass (three or more). The scheme of DNN is presented in Figure 4.



**Figure 4 - Dense neural network**

A Long short-term memory (LSTM)[16]is a type of Recurrent neural network designed to effectively capture long-term dependencies in the sequential data and avoid problems related to the vanishing gradient. An LSTM cell contains three types of gates: input, forget, and output, which regulate the stream of information through the whole network. The forget gate decides which information to discard from the cell state, while the input gate selects new information to store. The output gate defines the information that is required to be extracted from the current cell state.

The forget gate is calculated as (4):

, (4)

where  is a forget gate output;  is a sigmoid activation function;  are weights for the forget gate;  is a hidden state from the previous time step;  is a current input;  is bias for the forget gate.

The input gate is computed as (5):

, (5)

where  is an input gate;  are weights for the input gate;  is bias for the input gate.

The cell state is calculated as (6):

, (6)

where  is a cell state;  are weights for the cell state;  is bias for the cell state.

The output gate is computed as (7):

, (7)

where  is the output gate;  is bias for the output gate.

The scheme of the LSTM model is shown in Figure 5.

**Изображение выглядит как текст, диаграмма, снимок экрана, линия

Автоматически созданное описание**

**Figure 5 - Long short-term memory neural network**

A GRU [17] is another type of Recurrent neural network designed to capture dependencies in sequential data like LSTM but with a simpler structure. GRU combines the hidden and cell states into one entity, making them computationally more efficient than LSTM [18]. GRU includes update, reset gates, candidate hidden state, and final hidden state.

The update gate is calculated as (8):

, (8)

where  is an update gate;  is a sigmoid activation function;  is a weight matrix for the update gate.

The reset gate is computed as (9):

, (9)

where  is a reset gate;  is a weight matrix for the reset gate;  is bias for the reset gate.

The hidden state is calculated as (10):

, (10)

where  is a candidate hidden state, which incorporates the current input and the reset hidden state;  is the weight matrix for the candidate hidden state;  is a bias term for the candidate hidden state;  is the element-wise multiplication.

The final hidden state is computed as (11):

, (11)

where  is a final hidden state for the current time step;  is an update gate controlling how much of the previous hidden state to keep and how much of the candidate state to add.

The scheme of the GRU neural network is shown in Figure 6.



**Figure 6 - Gated recurrent unit neural network**

The machine learning and neural network algorithms presented above have been widely used and distributed in many works due to their success in data classification tasks. In this study, they also successfully coped with the task of malware detection. However, to improve the efficiency of threat detection, a new hybrid neural network model, which includes LSTM and GRU layers, was developed. This model combines the advantages of both architectures for more efficient data processing. The first recurrent LSTM layer helps model long-term dependencies in data, revealing patterns useful for identifying threats. Since LSTM has memory, it retains information about previous steps, which is especially important in sequence analysis. The GRU layer follows LSTM and helps simplify the model training by reducing the number of parameters compared to LSTM, making it more efficient. The hybrid LSTM-GRU model uses the advantages of both architectures. The LSTM layer is used as the first layer to process the data, and the GRU can be used in the next stage to speed up the training and processing of the output data. This reduces the training time on large datasets. Overall, the hybrid LSTM-GRU model offers enough functionality to handle complex models without overloading the computational resources of the machines. The scheme of the hybrid LSTM-GRU neural network model is shown in Fig. 7.

****

**Figure 7 - LSTM-GRU neural network**

**Results and discussion.** Malware classification experiments [19] were tested on the corresponding dataset. First of all, the dataset was cleaned and preprocessed. Three unmeaning features were eliminated, while the remaining 55 features were normalized with the Min-Max scaling technique [20]. The feature selection algorithm was applied to the normalized dataset, retrieving the 20 most significant ones. In total, 3 machine learning (Decision Tree, Random Forest, XgBoost), 3 deep learning (DNN, LSTM, and GRU), and 1 proposed hybrid LSTM-GRU deep learning model were used to classify the dataset.

The results were evaluated with the use of 4 measures: Accuracy, Precision, Recall, and F1-score. Accuracy measures the portion of correctly classified elements out of the whole number of elements. The formula for calculating Accuracy is shown in (12):

, (12)

Precision is the portion of correctly predicted positive elements to the total elements that were predicted as positive. Precision is computed as (13):

, (13)

Recall is the portion of correctly predicted positive elements to the total actual positive elements. Recall is computed as (14):

, (14)

F1-score is the metric characterizing the balance between Precision and Recall. F1-score is computed as (15):

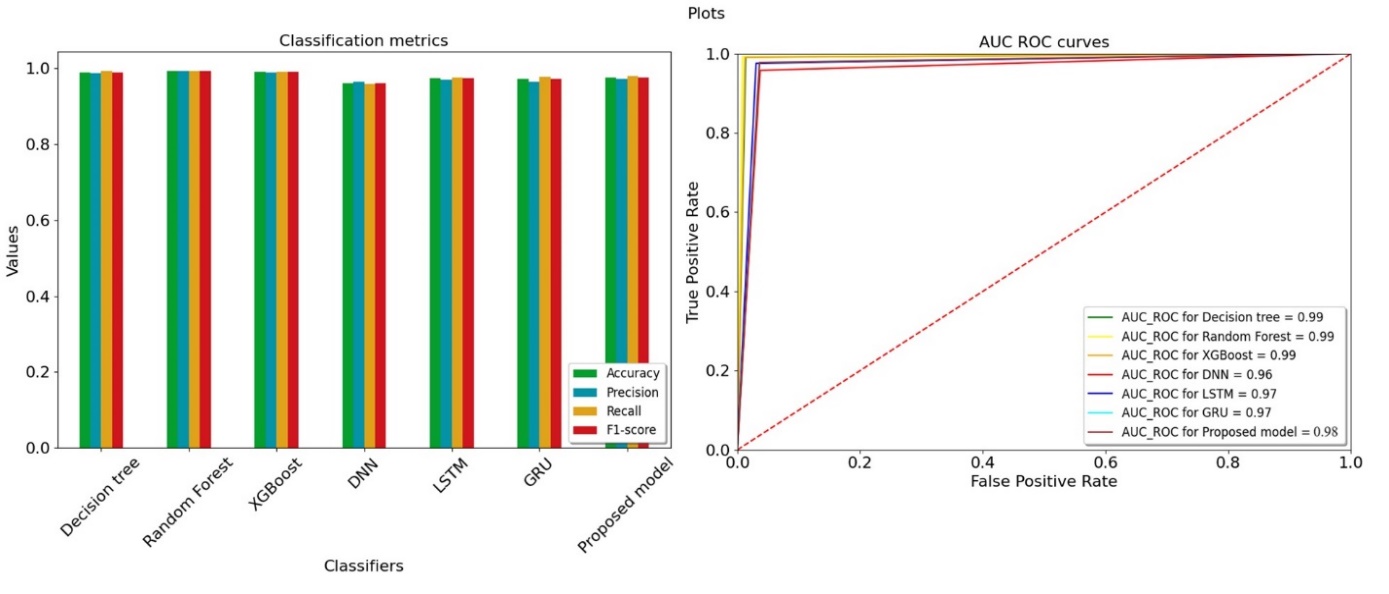
, (15)

The results of classifying the Malware dataset with machine learning and deep learning models were evaluated with the described metrics, the histograms, and the Area Under the ROC curve. The AUC-ROC value is in the range of 0 to 1.0. A value of 1.0 means an ideal classifier; a value of 0.5 describes random guessing, and a value less than 0.5 identifies a problematic case.

The results of the experiments that were conducted are shown in Table 1 and Figure 7.

**Table 1. The values of Malware classification**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Metrics | Decision Tree | Random Forest | XGBoost | DNN | LSTM | GRU | Proposed LSTM-GRU model |
| Ассurасy | 0.988 | 0.992 | 0.989 | 0.959 | 0.972 | 0.971 | 0.986 |
| Рrесisiоn | 0.985 | 0.991 | 0.987 | 0.962 | 0.972 | 0.966 | 0.981 |
| Rесаll | 0.991 | 0.993 | 0.991 | 0.957 | 0.972 | 0.976 | 0.989 |
| F1-sсоrе | 0.988 | 0.992 | 0.989 | 0.960 | 0.972 | 0.971 | 0.985 |



**Figure 7-Histograms and an AUC-ROC curve**

All seven models showed good classification results. The Decision Tree, Random Forest, and XGBoost machine learning models gave high accuracy scores. It proves that machine learning algorithms are still strong in the classification tasks of the cybersecurity field. The Random Forest model reached scores of 0.99, being the best among other machine learning models. This proves that Random Forest has always had the greatest potential in classification tasks. Among neural networks, the simple DNN model was a little worse than LSTM and GRU by around 0.01. The recurrent LSTM and GRU models showed mostly identical scores. The proposed LSTM-GRU model outperformed other deep learning models by 0.01 and demonstrated good tendencies in getting even higher scores in case of having larger datasets.

**Conclusion.** The fast development of computer technologies has led to the appearance of a large number of cyber threats that advance in breaches of security and confidentiality of information systems. The practical side of this research presents the creation and testing of malware detection models that can improve the security and resilience of information systems to modern cyber threats. Malware is one of the most dangerous threats because many of its forms exist, including viruses, worms, trojans, spyware, etc. They can rapidly spread in the information environment. Therefore, effective measures of their detection and prevention have become especially relevant in the last decade. The traditional ways of protecting against such cyber threats have become less significant in recent years. It is necessary to utilize other advanced methods. This research is based on machine learning algorithms and deep learning models, which allow the analysis of malware based on known signatures and identify new, previously unknown threats. The practical experiments of this research work covered the detection of malware threats with the use of seven machine learning and deep learning models, including Decision Tree, Random Forest, XGBoost, DNN, LSTM, GRU, and the proposed hybrid LSTM-GRU, giving accuracy scores above 0.95. The LSTM-GRU and Random Forest models reached scores of 0.98 and 0.99, being very close to the ideal result. It is also planned to extend this research, adding more experiments to classify various other types of threats in future works. Therefore, the study demonstrates that using machine learning and deep learning to analyze malware can improve the accuracy and reliability of classification and increase the flexibility of security systems, making them less vulnerable to new, unpredictable threats.

***Financing.*** *This work was carried out under project AP19675957, titled “The research and development of the system for ensuring the protection of medical data using blockchain technology and artificial intelligence methods,” implemented at the Institute of Information and Computational Technologies.*

**References**

1. Khando, K., Gao, S., Islam, S. M., & Salman, A. Enhancing employees information security awareness in private and public organisations: A systematic literature review // Computers & security. -2021.- Vol. 106. DOI [10.1016/j.cose.2021.102267](https://doi.org/10.1016/j.cose.2021.102267).
2. Nissenbaum, H. Protecting privacy in an information age: The problem of privacy in public //In The ethics of information technologies. Routledge. -2020.- P. 141-178. DOI [10.2307/3505189](https://doi.org/10.2307/3505189).
3. Shelke, M.V., Deshmukh, J.Y., Ajalkar, D.A., & Dhumal, R.B. A Robust Ensemble Learning Approach for Malware Detection and Classification // Journal of Advanced Research in Applied Sciences and Engineering Technology.-2024.-Vol. 48(1).- P. 152-167.

DOI [10.37934/araset.48.1.152167](https://doi.org/10.37934/araset.48.1.152167).

1. Zhao, Q., Chen, S., Liu, Z., Baker, T., & Zhang, Y. Blockchain-based privacy-preserving remote data integrity checking scheme for IoT information systems // Information Processing & Management. -2020. - Vol. 57(6). [DOI 10.1016/j.ipm.2020.102355](https://doi.org/10.1016/j.ipm.2020.102355).
2. Tun Li, Ya Luo, Xin Wan, Qian Li, Qilie Liu, Rong Wang, Chaolong Jia, Yunpeng Xiao. A malware detection model based on imbalanced heterogeneous graph embeddings // Expert Systems with Applications.-2024.-Vol. 246. 123109. DOI [10.1016/j.eswa.2023.123109](https://doi.org/10.1016/j.eswa.2023.123109).
3. Caixia Gao, Yao Du, Fan Ma, Qiuyan Lan, Jianying Chen, Jingjing Wu. A new adversarial malware detection method based on enhanced lightweight neural network // Computers & Security. -2024. -Vol. 147. DOI [10.1016/j.cose.2024.104078](https://doi.org/10.1016/j.cose.2024.104078).
4. Wira Zanoramy, Mohd Faizal Abdollah, Othman Abdollah, & S.M. Warusia Mohamed S.M.M. Ransomware Early Detection using Machine Learning Approach and Pre-Encryption Boundary Identification // Journal of Advanced Research in Applied Sciences and Engineering Technology. -2024. -Vol. 47(2).- P. 121 - 137. DOI [10.37934/araset.47.2.121137](https://doi.org/10.37934/araset.47.2.121137).
5. Hossain, M.A., Islam, M.S. Enhanced detection of obfuscated malware in memory dumps: a machine learning approach for advanced cybersecurity // Cybersecurity. -2024. -Vol. 7(16).

DOI [10.1186/s42400-024-00205-z](https://doi.org/10.1186/s42400-024-00205-z).

1. Mosleh M. Abualhaj, Ahmad Sami Al-Shamayleh, Alhamza Munther, Sumaya Nabil Alkhatib, Mohammad O. Hiari, Mohammed Anbar. Enhancing spyware detection by utilizing decision trees with hyperparameter optimization // Bulletin of Electrical Engineering and Informatics. -2022. -Vol. 13(5). - P. 3653-3662. DOI [10.11591/eei.v13i5.7939](https://doi.org/10.11591/eei.v13i5.7939).
2. Ahmad Heryanto, Deris Stiawan, Adi Hermansyah, Rici Firnando, Hanna Pertiwi, Mohd Yazid Bin Idris, Rahmat Budiarto The incorporation of stacked long short-term memory into intrusion detection systems for botnet attack classification // IAES International Journal of Artificial Intelligence. -2024. -Vol. 13(3).- P. 3657-3670. DOI [10.11591/ijai.v13.i3.pp3657-3670](https://doi.org/10.11591/ijai.v13.i3.pp3657-3670).
3. Aso Mafakheri, Sadegh Sulaimany. Android malware detection through centrality analysis of applications network//Applied Soft Computing. -2024.-Vol. 165.

DOI [10.1016/j.asoc.2024.112058](https://doi.org/10.1016/j.asoc.2024.112058).

1. Singh, D., & Singh, B. Investigating the impact of data normalization on classification performance // Applied Soft Computing. -2020. -Vol. 97. DOI [10.1016/j.asoc.2019.105524](https://doi.org/10.1016/j.asoc.2019.105524).
2. Andelic, N., & Baressi Šegota, S. Development of symbolic expressions ensemble for breast cancer type classification using genetic programming symbolic classifier and decision tree classifier // Cancers. -2023. - Vol. 15(13). DOI [10.3390/cancers15133411](https://doi.org/10.3390/cancers15133411).
3. Manzali, Y., & Elfar, M. Random forest pruning techniques: a recent review // In Operations research forum. -2023. -Vol. 4(43). DOI [10.1007/s43069-023-00223-6](https://doi.org/10.1007/s43069-023-00223-6).
4. Hariharan, B., Siva, R., Sadagopan, S., Mishra, V., & Raghav, Y. Malware Detection Using XGBoost based Machine Learning Models-Review // In 2023 2nd International Conference on Edge Computing and Applications (ICECAA). -2023. -P. 964-970.

DOI [10.1109/ICECAA58104.2023.10212327](https://doi.org/10.1109/ICECAA58104.2023.10212327).

1. Gupta, K., Jiwani, N., Sharif, M. H. U., Datta, R., & Afreen, N. A Neural Network Approach For Malware Classification // In 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). -2022. -P. 681-684.

DOI [10.1109/ICCCIS56430.2022.10037653](https://doi.org/10.1109/ICCCIS56430.2022.10037653).

1. Catak, F. O., Ahmed, J., Sahinbas, K., & Khand, Z. H. Data augmentation based malware detection using convolutional neural networks // Peerj computer science. - 2021.

DOI [10.7717/peerj-cs.346](https://doi.org/10.7717/peerj-cs.346).

1. Thakur, P., Kansal, V., & Rishiwal, V. Hybrid deep learning approach based on lstm and cnn for malware detection // Wireless Personal Communications. -2024. -Vol. 136(3). - P.1879-1901. DOI [10.1007/s11277-024-11366-y](https://doi.org/10.1007/s11277-024-11366-y).
2. Kolli, S., Balakesavareddy, P., & Saravanan, D. Neural Network based Obfuscated Malware detection // In 2021 International Conference on System, Computation, Automation and Networking (ICSCAN). -2021. DOI [10.1109/ICSCAN53069.2021.9526496](https://doi.org/10.1109/ICSCAN53069.2021.9526496).
3. Ussatova, O., Zhumabekova, A., Begimbayeva, Y., Matson, E. T., & Ussatov, N. Comprehensive DDoS Attack Classification Using Machine Learning Algorithms // Computers, Materials & Continua. -2022. -Vol. 73(1). - P. 577-594. DOI [10.32604/cmc.2022.026552](https://doi.org/10.32604/cmc.2022.026552).
4. Ussatova, O., Zhumabekova, A., Karyukin, V., Matson, E. T., Ussatov, N. The development of a model for the threat detection system with the use of machine learning and neural network methods //International Journal of Innovative Research and Scientific Studies. -2024.-Vol. 7(3). -P. 863- 877. DOI [10.53894/ijirss.v7i3.2957](https://doi.org/10.53894/ijirss.v7i3.2957).

***Information about the authors***

Zhumabekova A.- мaster, Institute of Information and Computational Technologies, Al-Farabi Kazakh National University,Almaty, Kazakhstan, e-mail: [zhumabekova2702@gmail.com](mailto:zhumabekova2702@gmail.com);

Ussatova O.- PhD, associate professor, Institute of Information and Computational Technologies, G.Daukeev Almaty University of Energy and Communications,Almaty, Kazakhstan*,* e-mail: [uoa\_olga@mail.ru](mailto:uoa_olga@mail.ru);

Kalimoldayev M. – doctor of Physical and Mathematical Sciences, professor, Academician of the National Academy of Sciences of the Republic of Kazakhstan, Institute of Information and Computational Technologies, Almaty, Kazakhstan, e-mail: [mnk@ipic.kz](mailto:mnk@ipic.kz);

Karyukin V.-PhD, Institute of Information and Computational Technologies, Al-Farabi Kazakh National University,Almaty, Kazakhstan, e-mail: [vladislav.karyukin@gmail.com](mailto:vladislav.karyukin@gmail.com);

Begimbayeva Y.-PhD, Associate professor, Institute of Information and Computational Technologies, G.Daukeev Almaty University of Energy and Communications,Almaty, Kazakhstan*,* Almaty, Kazakhstan,e-mail: [Enlik\_89@mail.ru](mailto:Enlik_89@mail.ru)

***Сведения об авторах***

Жұмабекова А.Т.-магистр, Институт информационных и вычислительных технологий, КазНУ им. аль-Фараби, Алматы, Казахстан, e-mail: [zhumabekova2702@gmail.com](mailto:zhumabekova2702@gmail.com);

Усатова О.А. - PhD, ассоциированный профессор (доцент), Институт информационных и вычислительных технологий, Алматинский Университет Энергетики и Связи им. Г.Даукеева,Алматы, Казахстан, e-mail: [uoa\_olga@mail.ru](mailto:uoa_olga@mail.ru);

Калимолдаев М.Н.- доктор физико-математических наук, профессор, Академик НАН РК, Институт информационных и вычислительных технологий, Алматы, Казахстан, e-mail: [mnk@ipic.kz](mailto:mnk@ipic.kz);

Карюкин В.И.- PhD, Институт информационных и вычислительных технологий, КазНУ им. аль-Фараби, Алматы, Казахстан, e-mail: [vladislav.karyukin@gmail.com](mailto:vladislav.karyukin@gmail.com);

Бегимбаева Е.Е. - PhD, ассоциированный профессор (доцент), Институт информационных и вычислительных технологий, Алматинский Университет Энергетики и Связи им. Г.Даукеева, Алматы, Казахстан, e- mail: [Enlik\_89@mail.ru](mailto:Enlik_89@mail.ru)

ҒТАМР 20.23.25

**КАТАЛИТКАЛЫҚ КРЕКИНГ ҚОНДЫРҒЫСЫНАН ӨНІМНІҢ ШЫҒУЫН PYTHON БАҒДАРЛАМАЛЫҚ ОРТАСЫН ҚОЛДАНУ АРҚЫЛЫ АНЫҚТАУ**

**1Н.Боранбаева, 1Б.Оразбаев🖂, 2Л.Рзаева, 3Ж.Карабаев, 4Б.Серимбетов**

1Л.Н.Гумилев атындағы Еуразия Ұлттық Университеті, Астана, Қазақстан,

2Astana IT University, Астана қ., Қазақстан,

3Auezov University, Шымкент қ., Қазақстан,

4 Қ.Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан

**🖂**Корреспондент-автор: [batyr\_o@mail.ru](mailto:batyr_o@mail.ru)

Мақалада Python бағдарламалық ортасында регрессия әдістерін қолдану арқылы каталитикалық крекинг қондырғысынан өнімнің шығуын анықтау тапсырмасы қарастырылады. Каталитикалық крекинг мұнай өндірудің негізгі процесі болғандықтан, өнімнің шығуын нақты болжау технологиялық параметрлерді оңтайландыру және өнімнің тиімділігін арттыру үшін өте маңызды. Сараптама үшін Шымкент мұнай өндіру зауытының статистикалық мәліметтері қолданылды, бұл ұсынылған әдісті нақты өндіріс шарттарында қолдану мүмкіндігін көрсетеді. Мақалада реактордағы және регенератордағы температура, қысым, шикізат тығыздығы және катализатор шығыны сияқты әр түрлі технологиялық параметрлердің әсер етуін талдау нәтижесінде модельдерді жасап шығару әдістемесі қарастырылған. Pandas, Scikit-learn және Matplotlib сияқты Python мәліметтерді талдау инструменттері мен кітапханаларын қолдану толық талдау жасап, нәтижелерін визуализациялауға мүмкіндік береді. Модельдің сапасын бағалау детерминациялау коэфициенті (R2) және қалдық қателерді талдау арқылы жүргізіледі. Алынған нәтижелер моделдің жоғарғы дәлдігі мен ұсынылған тәсілдің өндірістегі мұнай өндіру процестерін болжау және оңтайландыру мүмкіндігін дәлелдейді. Сондай-ақ, ұсынылған тәсілді мұнай өңдеу процестерін оңтайландыру, технологиялық шешімдерді жақсарту және мұнай-химия саласындағы өндірістердің жалпы экономикалық тиімділігін арттыру үшін сәтті қолдануға болатындығы атап өтілді.

**Түйін сөздер**: модельдеу, каталитикалық крекинг, python, детерминациялау коэффициенті, айқын емес логика, компьютерлік модельдеу, оңтайландыру, каталитикалық крекинг қондырғысы.

**ОПРЕДЕЛЕНИЕ ВЫХОДА ПРОДУКТА ИЗ УСТАНОВКИ КАТАЛИТИЧЕСКОГО КРЕКИНГА ИСПОЛЬЗОВАНИЕМ ПРОГРАМНОЙ СРЕДЫ PYTHON**

**1Н.Боранбаева, 1Б.Оразбаев🖂, 2Л.Рзаева, 3Ж.Карабаев, 4Б.Серимбетов**

1Евразийский Национальный Университет им. Л.Н.Гумилева, г.Астана, Казахстан,

2Astana IT University, г. Астана, Казахстан,

3Auezov University, г. Шымкент, Казахстан,

4Казахский университет технологии и бизнеса им.К.Кулажанова,

е-mail: [batyr\_o@mail.ru](mailto:batyr_o@mail.ru)

В данной статье рассматривается задача определение выхода продукта из установки каталитического крекинга с использованием методов регрессии в программной среде Python. Каталитический крекинг является одним из ключевых процессов в нефтепереработке, и точное прогнозирование выхода продукта имеет важное значение для оптимизации технологических параметров и повышения эффективности производства. Для анализа использованы статистические данные Шымкентского нефтеперерабатывающего завода, что позволяет применить предложенные методы к реальным производственным условиям. В работе представлены методология разработки модели на основе анализа влияния различных технологических параметров, таких как температура, давление, плотность сырья и расход катализатора. Применение инструментов анализа данных и библиотек Python, таких как Pandas, Scikit-learn и Matplotlib, позволяет провести детальный анализ и визуализацию результатов. Оценка качества модели производится с помощью коэффициента детерминации (R2) и анализа остаточных ошибок. Полученные результаты демонстрируют высокую точность модели и подтверждают возможность использования предложенного подхода для прогнозирования и оптимизации процессов нефтепереработки в промышленной практике*.* Также подчеркивается, что предложенный подход может быть успешно применен для оптимизации процессов нефтепереработки, улучшения технологических решений и повышения общей экономической эффективности производств в нефтехимической отрасли.

**Ключевые слова:** моделирование, каталитический крекинг, python, коэффициент детерминации, нечеткая логика, компьютерное моделирование, оптимизация, установка каталитического крекинга.

**DETERMINATION OF PRODUCT YIELD FROM A CATALYTIC CRACKING UNIT USING THE PYTHON PROGRAMMING ENVIRONMENT**

**1N.Boranbayeva, 1B.Orazbayev🖂, 2L.Rzayeva, 3Zh.Karabayev, 4B.Serimbetov**

1 L. N. Gumilyov Eurasian National University, Astana, Kazakhstan,

2Astana IT University, г. Астана, Kazakhstan,

3Auezov University, Shymkent, Kazakhstan,

4K. KulazhanovKazakh University of Technology and Business, Astana, Kazakhstan,

е-mail: [batyr\_o@mail.ru](mailto:batyr_o@mail.ru)

This article addresses the task of forecasting product yield from a catalytic cracking unit using regression methods within the Python programming environment. Catalytic cracking is one of the key processes in oil refining, and accurate product yield forecasting is crucial for optimizing technological parameters and improving production efficiency. Statistical data from the Shymkent Oil Refinery were used for the analysis, which allows us to apply the proposed methods to real production conditions. The paper presents a model development methodology that analyzes various technological parameters, such as temperature, pressure, feedstock density, and catalyst consumption. Using data analysis tools and Python libraries, such as Pandas, Scikit-learn, and Matplotlib, enables detailed analysis and visualization of results. The model's quality is assessed using the coefficient of determination (R2) and residual error analysis. The results obtained demonstrate the high accuracy of the model and confirm the feasibility of using the proposed approach for forecasting and optimizing oil refining processes in industrial practice. Against the background of the growing relevance of issues related to the diagnosis of brain stroke, modern research in the field of medical diagnostics is trying to use advanced deep learning methods to improve the detection of this serious disease. It is also emphasized that the proposed approach can be successfully applied to optimize refining processes, improve technological solutions and increase the overall economic efficiency of production in the petrochemical industry.

**Keywords:** modeling, catalytic cracking, Python, coefficient of determination, fuzzy logic, computer modeling, optimization, unit of catalytic cracking.

**Кіріспе.** Каталитикалық крекинг жоғары октанды бензиндер мен басқа да жеңіл мұнай өнімдерін өндіруде шешуші рөл атқаратын мұнай өңдеудегі маңызды процестердің бірі болып табылады. Бұл процестің тиімділігі температура, қысым, шикізаттың тығыздығы және катализатордың шығыны сияқты көптеген технологиялық параметрлерге тікелей байланысты. Каталитикалық крекинг қондырғыларының жұмысын оңтайландыру және олардың өнімділігін арттыру үшін технологиялық параметрлер туралы мәліметтер негізінде өнімнің шығуын дәл болжау қажет. Деректерді талдау мен машиналық оқытудың заманауи әдістері өндіріс процестерін едәуір жақсарта алатын болжамды модельдерді құрудың қуатты құралдарын ұсынады. Атап айтқанда, регрессиялық талдау машиналық оқытудың негізгі әдістерінің бірі ретінде айнымалылар арасындағы байланысты талдау және әртүрлі көрсеткіштерді болжау үшін кеңінен қолданылады.

Бұл мақалада Python бағдарламалық жасақтамасында жүзеге асырылған регрессия әдістерін қолдана отырып, каталитикалық крекинг қондырғысынан өнімнің шығуын болжау тәсілі ұсынылады. Python Pandas, Scikit-learn және Matplotlib сияқты қуатты деректерді талдау және визуализация кітапханаларының арқасында мұндай модельдерді өнеркәсіптік тәжірибеге енгізу және жасақтау үшін таңдаулы құралға айналуда.

Бұл жұмыстың мақсаты болжау моделін әзірлеу және бағалау болып табылады, ол өнімнің шығуын жоғары дәлдікпен болжап қана қоймай, сонымен қатар нақты жағдайларда технологиялық процестерді одан әрі оңтайландыру үшін осы болжамдарды қолдануға мүмкіндік береді. Мақалада параметрлерді таңдаудан бастап болжау сапасын бағалауға дейінгі модельді әзірлеудің барлық кезеңдері қарастырылады, сонымен қатар ұсынылған тәсілдің тиімділігін растайтын нәтижелер талқыланады. Өнімнің каталитикалық крекинг қондырғысынан шығуын болжау күрделі міндет болып табылады, ол шикізаттың құрамы, процесс шарттары және катализатордың сипаттамалары сияқты көптеген факторларды ескеруді талап етеді. Соңғы жылдары мұнай-химия саласындағы зерттеушілер мен тәжірибешілер болжау дәлдігін жақсарту және процесті оңтайландыру үшін деректерді талдау және машиналық оқыту әдістеріне көбірек бет бұруда.

Регрессия-химиялық процестерді болжауда ең көп қолданылатын әдістердің бірі. [1] жұмыста регрессиялық модельдер айнымалылар арасындағы тәуелділікті талдаудың негізгі мүмкіндіктерін ұсынады. Зерттеулер көрсеткендей, регрессиялық модельдер қарапайымдылығына қарамастан, әсіресе деректерді алдын ала талдау және негізгі трендтерді анықтау жағдайында күрделі тәсілдер үшін жақсы бастама бола алады, [2]. Python бағдарламалық ортасының Pandas, NumPy және scikit-learn сияқты қуатты кітапханалары мен құрылымдарының арқасында деректерді талдаудың танымал құралына айналды. Бірқатар жұмыстар [3] бұл құралдар регрессиялық модельдерді құру және бағалау процесін едәуір жеңілдететінін атап өтті. Python сонымен қатар деректерді визуализациялау және модельдеу үшін ыңғайлы құралдарды ұсынады, бұл оны мұнай-химия практиктері үшін ерекше құнды етеді. Каталитикалық крекинг саласындағы заманауи зерттеулер көбінесе нейрондық желілер сияқты күрделі машиналық оқыту әдістерін қолдануды қамтиды [4]. Бұл әдістер үлкен көлемдегі деректермен және күрделі тәуелділіктермен жұмыс істегенде жоғары тиімділікті көрсетеді. Алайда, [5] атап өткендей, мұндай әдістерді өнеркәсіптік тәжірибеге біріктіру айтарлықтай есептеу ресурстарын және процестерді терең түсінуді қажет етеді. Қосымша зерттеуді қажет ететін маңызды аспектілер - болжамдардың дәлдігі және нәтижелерді түсіндіру. [6] ғылыми жұмыста атап өткендей, қолданыстағы модельдер өзгеретін процестер жағдайында қайта оқыту және дәлдіктің жеткіліксіздігімен жиі ұшырасады. Перспективалық бағыттарға сызықтық регрессия элементтері мен машиналық оқытудың күрделі алгоритмдерін біріктіретін гибридті модельдерді дамыту, сондай-ақ деректерді жинау және алдын ала өңдеу әдістерін жақсарту кіреді [7].

Әртүрлі дереккөздерді талдау нәтижелері каталитикалық крекинг қондырғысы сияқты күрделі айқын емес сипатталған химиялық-технологиялық жүйелерді басқару, модельдеу мәселелерінің жеткіліксіз зерттелгендігін көрсетеді. Осыған байланысты модельдерді әзірлеу, оңтайландыру мәселелерін зерттеу, күрделі айқын емес химиялық-технологиялық жүйелерді басқару үшін шешім қабылдау тапсырмасын қалыптастыру заманауи ғылымның өзекті міндеті болып қала береді. Математикалық модельдерді әзірлеу принциптері мен әдістерін талдау, шешім қабылдау және өнеркәсіптік объектілерді оңтайландыру нәтижесінде ғылыми жұмыстарда лингвистикалық модельдерді әзірлеу және олардың жұмыс режимдерін оңтайландыру мәселелері аз қамтылғандығы анықталды. Жұмыстарда [8, 9] математикалық модельдерді әзірлеу және бастапқы ақпараттың айқынсыздығымен сипатталатын технологиялық объектілердің параметрлерін оңтайландыру тәсілдері зерттеліп, ұсынылды. 2000 жылдан бастап Р.А. Алиева, н. Р. Юсупбекова, Azeem M. F., Taskin H., Osofisan P. B. сияқты авторлардың каталитикалық крекингті басқару алгоритмдерінде айқын емес логика әдістерін қолдану арқылы жүзеге асырылған бірқатар еңбектері жарық көрді. Бірақ күрделі объектілерді модельдеу және оңтайландыру бойынша осы және басқа талданған жұмыстарда объектінің кіріс және шығыс параметрлері айқын емес модельдерді әзірлеу мәселелері толық зерттелмеген. Сонымен қатар, оңтайландыру мәселелерін шешудің белгілі әдістерінде есептің қойылуы кезеңінде айқын емес есеп мәселе нақты есептер жиынтығына айналады және одан әрі қолданыстағы әдістермен шешіледі. Бұл тәсілде көбінесе бастапқы жиналған айқын емес ақпараттың едәуір бөлігі жоғалады (білім, сарапшылардың тәжірибесі), нәтижесінде алынған шешімдердің шындыққа сәйкестігі төмендейді [10].

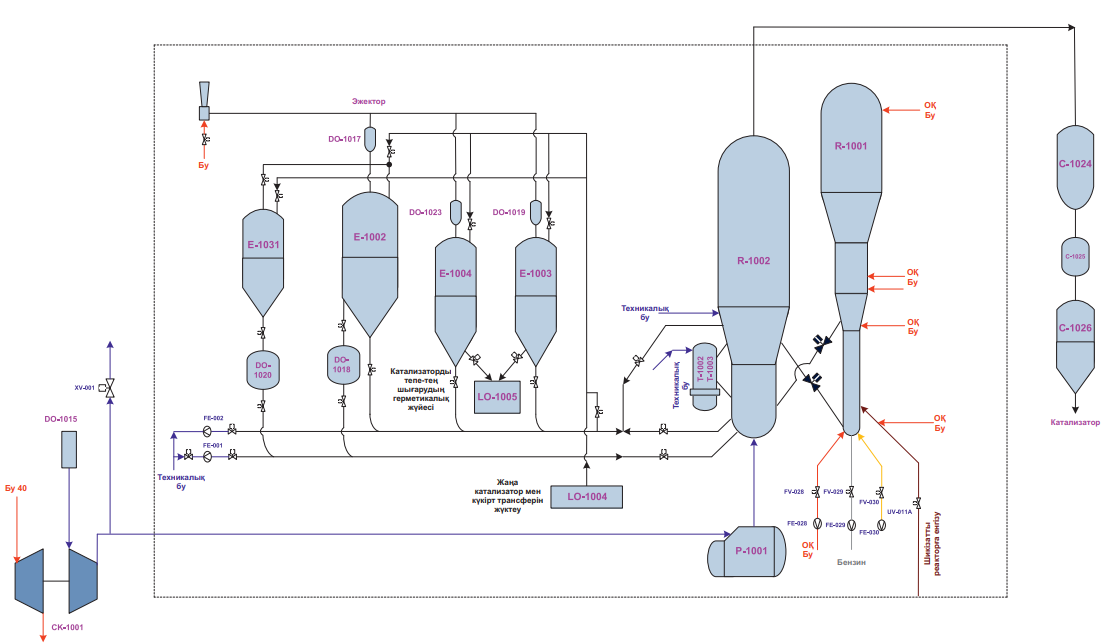
Әдебиеттерге шолу өнімнің каталитикалық крекинг қондырғысынан өнімнің шығуын анықтауда Python және регрессия әдісін қолдану процестерді талдауға және оңтайландыруға арналған қуатты құралдарды қамтамасыз ететінін көрсетеді. Дегенмен, дәлірек және сенімді нәтижелерге қол жеткізу үшін деректерді талдаудың дәстүрлі әдістерімен бірге заманауи әдістерді де ескеру қажет.

**Материалдар мен әдістер.** Бұл жұмыстың зерттеу нысаны Шымкент мұнай өңдеу зауытының титулы 1000 RFCC ауыр қалдықтардың каталитикалық крекинг қондырғысының реактор-регенератор блогы болып табылады. Каталитикалық крекинг қондырғысы тікелей мазуттан каталитикалық крекинг процесі арқылы автомобиль бензиндері мен сұйытылған көмірсутек газдарының жоғары октанды компонентін алуға арналған. Каталитикалық крекинг процесінің жүруіне әсер ететін негізгі факторлар шикізаттың физика-химиялық сипаттамалары, катализатордың қасиеттері, реактордағы қысым, реактордағы шикізат температурасы, реактордың температурасы, катализатордың шығыны болып табылады. Каталитикалық крекинг қондырғысынан бензиннің шығуы мен сапасы шикізаттың құрамына, реактордағы температура мен қысымға, шикізаттың берілу жылдамдығына, регенератордағы температураға, регенератор-реактор блогының жұмыс режиміне, крекинг катализаторларының түріне және басқа параметрлерге байланысты. Технологиялық режимдер алынатын өнімінің түріне байланысты майлы газ, бензин немесе дизель фракциясы ретінде ерекшеленеді. Басқару объектісінің техникалық регламенті негізінде процестің негізгі технологиялық параметрлері анықталды: реактордың жоғарғы бөлігіндегі қысым, реактордан шығатын реакция газдарының температурасы, регенератордан шығатын газдардың температурасы. Әрі қарай, процеске әсер ететін басқару объектісінің кіріс, шығыс, ауытқушы айнымалылары анықталады. Реактор блогындағы технологиялық параметрлер процесс режимінің оңтайлы қаттылығына қол жеткізу мақсатында реттеледі. Терең оқыту технологияларының дамуы диагностиканың дәлдігі мен жеделдігін жақсартудың жаңа перспективаларын ұсына отырып, осы процеске белсенді әсер етеді.

Анықтамалық векторлық әдіс (SVM) және кездейсоқ ормандар сияқты Машиналық оқыту әдістерінің пайда болуы диагностиканы жақсартудағы маңызды қадам болды. Бұл тәсілдер медициналық кескіндерді талдау процесін автоматтандыруға мүмкіндік береді, бірақ көбінесе күрделі үш өлшемді деректер мен үлкен көлемдегі ақпаратты өңдеуде шектеулерге тап болады.

Нейрондық желілердің архитектуралары, соның ішінде конволюциялық нейрондық желілер (CNN) және қайталанатын нейрондық желілер (RNN) деректердегі кеңістіктік және уақыттық тәуелділіктерді ескере отырып, жоғары дәлдікті қамтамасыз етеді [11].

Каталитикалық крекинг-мұнай өңдеуді тереңдетуге бағытталған негізгі процесс. Процесс келесі технологиялық агрегаттарды қамтитын каталитикалық крекинг қондырғысында жүреді: бу конденсациясы жүйесі, реактор-регенератор блогы, түтін газын дымқыл тазарту блогы, шикізатты дайындау және ректификациялау блогы, сіңіру және газ фракциялау блогы. Реактор мен регенератор арасындағы сұйық катализатордың үздіксіз айналуы каталитикалық крекинг процестерін үздіксіз режимде жүргізуге мүмкіндік береді [12].



**1- сурет. RFCC каталитикалық крекинг қондырғысының реактор және регенератор блогының технологиялық сызбасы**

1-суретте RFCC 1000 титулды каталитикалық крекинг қондырғысының реактор және регенератор блогының технологиялық схемасы көрсетілген [13]. Қондырғының шикізаты - шикізатты араластыру контейнеріне берілетін тікелей айдау мазуты немесе мазут қоспасы және тікелей айдау вакуумды газойль фракциясының қоспасы. Араластыру контейнеріне шикізат сорғымен реакторға беріледі, алдымен текше өнім ағындарының, жеңіл газойль фракциясының (қажет болған жағдайда) және текше циркуляциялық суарудың жылуы есебінен шикізат жылу алмастырғыштар блогында қызады. Реакторға кіретін шикізаттың температурасы соңғы жылу алмастырғыштың айналма жолындағы клапанмен реттеледі.

RFCC каталитикалық крекинг қондырғысында жұмыс жасайтын өндіріс қызметкерлерінің, сарапшылардың сауалнамасынан алынған ақпарат реактор блогының технологиялық жұмыс режимінің параметрлерінің өзгеру себептері реакторға кіретін температураның өзгеруі, катализаторға жиналатын кокс мөлшерінің өзгеруіне, сондай-ақ қоршаған ауа температурасының өзгеруіне әкелетінін көрсетеді. Бұл өзгерістер регенератордың шығысындағы катализаторының температурасына және реактордың кірісіндегі катализатордың температурасына әсер етеді. Бұл каталитикалық крекинг процесінің параметрлерінің өзгеруіне әкеледі [14]. Қазіргі уақытта RFCC каталитикалық крекинг қондырғысында реактор мен регенератор блогының жұмысын басқаруды өндіріс операторы параметрлерді (автоматты) тұрақтандыру жүйелерінің параметрлерін, реактордағы, шикізат пен бу шығынын, регенератордағы ауа мен бу шығынын, диспенсердегі бу мен ауа шығынын, катализатордың қалыпты айналымын қамтамасыз ету мақсатында өзгерту арқылы қолмен жүзеге асырады. Осы себепті қажетті октан саны бар сапалы бензин алу үшін каталитикалық крекинг процесін басқару үшін шешім қабылдауды автоматтандыру өзекті мәселе болып табылады [15].

Реактор-регенератор блогының жұмысын сипаттайтын негізгі көрсеткіштер реактордағы температура мен қысым болып табылады, бұл регенератордағы температураның өзгеруіне әсер етеді, нәтижесінде қондырғының тұрақты жұмысына, конверсия дәрежесіне, өнімнің тығыздығы мен фракциялық құрамына әсер етеді. Көптеген басқа технологиялық процестер сияқты RFCC каталитикалық крекинг қондырғысында жүретін зерттелетін процесс алынған бензиннің сапасы туралы айқын емес бастапқы ақпаратпен сипатталады, ол модельдерді әзірлеу және каталитикалық крекинг процесін оңтайландыру үшін қажет. Бұл жағдайда сараптамалық білімді ұсынудың айқын емес модельдерін қолдана отырып, каталитикалық крекинг қондырғысының реактор-регенератор блогындағы технологиялық процесті басқарудың интеллектуалды жүйесінің алгоритмдерін жасау қажет. Сарапшылардың тәжірибесіне, біліміне және түйсігіне сүйене отырып, толық емес ақпарат жағдайында лингвистикалық айнымалылардың мәндері үшін тиістілік функциялары автоматты түрде қалыптасады [16]. Реактор-регенератор блогының кіріс және шығыс параметрлері айқын емес лингвистикалық модельдерді синтездеу әдісі сараптамалық бағалау әдістеріне және айқын емес шартты қорытынды ережесіне негізделген.

Каталитикалық крекинг қондырғысынан өнімнің шығуын болжау үшін Шымкент мұнай өңдеу зауытында жиналған деректер пайдаланылды. Деректер жиынтығына келесі параметрлер кірді: шикізат көлемі, шикізат тығыздығы, шикізат температурасы, реактор температурасы, реактор қысымы, катализатор шығыны, бензин шығысы. Деректер үш жылдық кезеңді қамтыды. Процеске әсер ететін негізгі параметрлер 1-кестеде көрсетілген.

**1-кесте. Процестің негізгі кіріс, шығыс параметрлері**

|  |  |  |  |
| --- | --- | --- | --- |
| № | Белгіленуі | Параметр атауы | Өлшем бірлігі |
| 1 | (*x1*) | Шикізат шығыны | т/тәулік |
| 2 | (*x2*) | Шикізат тығыздығы | т/м3 |
| 3 | (*x3*) | Шикізат температурасы | С |
| 4 | (*x4*) | Реактор температурасы | С |
| 5 | (*x5*) | Реактор қысымы | кгс/см2 |
| 6 | (*x6*) | Катализатор шығыны | т/тәулік |
| 7 | *Y* | Бензин көлемі | % |

**2- кесте. Каталитикалық крекинг процесінің параметрлерінің статистикалық көрсеткіштеріне сәйкес болжау**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | *x1* | *x2* | *x3* | *x4* | *x5* | *x6* | *y*-тің өндірістегі мәні |  | *y*-тің анықталған мәні | Ауытқу |  |
| 1 | 241.8 | 0.899 | 206.3 | 517.8 | 2.26 | 1679.46 | 48.7 |  | 49.028339 | -0.328339 |
| 2 | 241.6 | 0.898 | 208.8 | 518.0 | 2.3 | 1692.4 | 48.6 |  | 48.719417 | -0.119417 |
| 3 | 242.5 | 0.898 | 209.1 | 517.6 | 2.3 | 1695.8 | 49.2 |  | 48.755016 | 0.444984 |
| 4 | 241.4 | 0.898 | 208.8 | 517.7 | 2.3 | 1684.0 | 49.1 |  | 48.665739 | 0.434261 |
| 5 | 240.9 | 0.901 | 210.4 | 517.8 | 2.2 | 1725.0 | 48.8 |  | 48.503400 | 0.296600 |
| 6 | 241.6 | 0.901 | 213.4 | 517.9 | 2.3 | 1763.0 | 48.7 |  | 47.988852 | 0.711148 |
| 7 | 241.9 | 0.899 | 212.4 | 517.9 | 2.3 | 1771.5 | 48.5 |  | 48.136549 | 0.363451 |
| 8 | 241.4 | 0.900 | 212.3 | 517.6 | 2.3 | 1752.8 | 48.8 |  | 48.053057 | 0.746943 |
| 9 | 240.2 | 0.900 | 214.3 | 518.3 | 2.2 | 1767.4 | 48.4 |  | 48.057432 | 0.342568 |
| 10 | 240.8 | 0.901 | 214.1 | 518.1 | 2.3 | 1780.7 | 47.9 |  | 47.774196 | 0.125804 |
| 11 | 155.6 | 0.901 | 196.6 | 515.4 | 2.2 | 1099.6 | 37.1 |  | 37.843743 | -0.743743 |
| 12 | 237.6 | 0.901 | 211.7 | 517.6 | 2.2 | 1745.1 | 47.3 |  | 47.758377 | -0.458377 |
| 13 | 239.2 | 0.900 | 213.0 | 518.1 | 2.2 | 1781.6 | 48.8 |  | 47.896658 | 0.903342 |
| 14 | 239.6 | 0.901 | 217.1 | 518.0 | 2.2 | 1766.9 | 48.4 |  | 47.672751 | 0.727249 |
| 15 | 239.6 | 0.901 | 216.5 | 518.3 | 2.2 | 1774.4 | 48.1 |  | 47.741578 | 0.358422 |
| 16 | 239.4 | 0.901 | 215.6 | 518.3 | 2.2 | 1754.3 | 47.9 |  | 47.856243 | 0.043757 |
| 17 | 239.8 | 0.902 | 213.8 | 518.5 | 2.2 | 1742.1 | 47.9 |  | 48.107864 | -0.207864 |
| 18 | 239.8 | 0.901 | 213.4 | 518.7 | 2.3 | 1748.7 | 48.0 |  | 47.904229 | 0.095771 |
| 19 | 239.8 | 0.901 | 214.5 | 519.0 | 2.2 | 1743.1 | 48.0 |  | 48.175504 | -0.175504 |
| 20 | 239.8 | 0.901 | 211.9 | 519.0 | 2.2 | 1743.0 | 47.5 |  | 48.365504 | -0.865504 |
| 21 | 239.1 | 0.901 | 212.9 | 519.1 | 2.3 | 1735.5 | 47.1 |  | 47.955085 | -0.855085 |
| 22 | 223.8 | 0.900 | 214.7 | 518.9 | 2.2 | 1762.5 | 46.3 |  | 45.514946 | 0.785054 |
| 23 | 216.4 | 0.901 | 209.5 | 519.5 | 2.2 | 1560.2 | 47.5 |  | 45.599599 | 1.900401 |
| 24 | 231.0 | 0.901 | 211.8 | 519.2 | 2.2 | 1689.0 | 50.5 |  | 47.209394 | 3.290606 |
| 25 | 238.0 | 0.900 | 214.9 | 519.0 | 2.2 | 1761.9 | 46.5 |  | 47.810517 | -1.310517 |
| 26 | 237.7 | 0.901 | 214.2 | 519.0 | 2.2 | 1773.7 | 46.7 |  | 47.735090 | -1.035090 |
| 27 | 237.6 | 0.901 | 214.4 | 518.9 | 2.3 | 1754.7 | 46.7 |  | 47.489306 | -0.789306 |
| 28 | 237.5 | 0.901 | 214.1 | 518.9 | 2.3 | 1766.6 | 46.9 |  | 47.446940 | -0.546940 |
| 29 | 237.8 | 0.901 | 213.1 | 518.9 | 2.2 | 1752.4 | 47.0 |  | 47.899089 | -0.899089 |
| 30 | 238.0 | 0.900 | 212.2 | 518.8 | 2.2 | 1739.9 | 47.2 |  | 48.059431 | 0.859431 |
| 31 | 237.7 | 0.900 | 212.9 | 519.0 | 2.2 | 1740.7 | 47.0 |  | 47.993722 | -0.993722 |
| 32 | 237.7 | 0.900 | 212.6 | 518.9 | 2.2 | 1765.0 | 47.1 |  | 47.898868 | -0.798868 |
| 33 | 237.6 | 0.899 | 212.2 | 519.0 | 2.3 | 1757.4 | 47.3 |  | 47.718064 | -0.418064 |
| 34 | 238.6 | 0.899 | 212.5 | 518.9 | 2.2 | 1744.3 | 47.2 |  | 48.165351 | -0.965351 |
| 35 | 240.2 | 0.898 | 212.1 | 518.9 | 2.3 | 1763.5 | 47.3 |  | 48.131774 | -0.831774 |
| 36 | 240.8 | 0.898 | 212.7 | 518.9 | 2.2 | 1775.6 | 48.2 |  | 48.409264 | -0.209264 |
| 37 | 241.4 | 0.899 | 212.9 | 518.9 | 2.2 | 1767.8 | 48.5 |  | 48.492558 | 0.007442 |
| 38 | 241.0 | 0.899 | 212.3 | 518.8 | 2.2 | 1793.4 | 49.0 |  | 48.349838 | 0.650162 |
| 39 | 240.2 | 0.898 | 210.4 | 519.1 | 2.3 | 1770.4 | 49.1 |  | 48.264773 | 0.835227 |
| 40 | 240.9 | 0.899 | 208.4 | 518.9 | 2.2 | 1789.6 | 49.0 |  | 48.651943 | 0.348057 |

Анықтауға арналған негізгі құрал ретінде регрессиялық модель таңдалды. Бұл әдіс оны талдауға ыңғайлы болуы мен қарапайымдылығы байланысты таңдалды. Регрессия моделі Python-дағы scikit-learn кітапханасының көмегімен жүзеге асырылды. Регрессия функциясы тәуелсіз айнымалылар (шикізат көлемі, шикізат тығыздығы, шикізат температурасы, реактор температурасы, реактор қысымы, катализатор шығыны) мен тәуелді айнымалы (бензин шығымы) арасындағы байланыс болды (2-кесте). Модельдің дәлдігін бағалау үшін келесі көрсеткіштер қолданылды:

- Орташа квадраттық қате (MSE): болжау қателіктерінің орташа квадратын өлшеу;

- Детерминация коэффициенті (R2): тәуелді айнымалы модель арқылы түсіндірілетін дисперсияның үлесін көрсететін көрсеткіш.

Талдау мен модельдеуді орындау үшін келесі бағдарламалық құралдар қолданылды:

- Python 3.9-деректерді өңдеуге және модель құруға арналған бағдарламалау тілі;

- Pandas-деректерді тазарту және талдау құралдарын қамтитын деректер кітапханасы;

- Сандық есептеулерді орындауға арналған NumPy-кітапхана;

- scikit-learn-Машиналық оқыту модельдерін құруға және бағалауға арналған кітапхана.

- Matplotlib және Seaborn-деректер мен модельдеу нәтижелерін визуализациялауға арналған кітапханалар.

Болжау нәтижелерін визуализациялау үшін болжамды мәндердің нақты мәндерге тәуелділік графиктері, сондай-ақ болжау қателерінің графиктері қолданылды. Бұл модельдің дәлдігін нақты көрсетуге және ықтимал кемшіліктерді анықтауға мүмкіндік берді.

Нәтижелер мен талқылау. Әдебиеттерді талдау және жүргізілген зерттеулер негізінде реактор мен регенератордың математикалық модельдерінің құрылымын таңдау жүргізілді [17,18,19]. Реактор мен Регенератор блогын басқару алгоритмінде каталитикалық крекинг қондырғысының RFCC қондырғысымен жұмыс жасайтын персоналдың тәжірибесін қолданған жөн. Сараптамалық білімді ресімдеу мақсатында сарапшыларға сауалнама жүргізілді, ол каталитикалық крекинг процесіне әсер ететін негізгі параметрлерді анықтауға мүмкіндік берді. Блоктардағы қалыпты режимді бұзатын негізгі ауытқытушы параметрлер шикізаттың шығыны, тығыздығы , температурасы болып табылады. Процеске әсер ететін негізгі факторлар - реактордағы температура  мен қысым , қалпына келтірілген катализатордың шығыны . Қондырғының кіріс параметрлерінің тиімділігін

сипаттайтын шығыс параметрінің математикалық тәуелділігін алу үшін Шымкент мұнай өңдеу зауытының статистикалық деректері негізінде эксперимент жүргізілді. Реактордағы температура мен қысым мәндері бергіштің көрсеткіштеріне сәйкес келеді. Шикізат пен сұйық газды тұтыну мәндері шығын өлшегіштердің көмегімен анықталады. Алайда, реактордан шыққан кезде бензинді өлшейтін шығын өлшегіштің болмауына байланысты үлкен қиындықтар болды. y параметрінің мәні субъективті түрде, әр бақылау кезінде қондырғы операторының технологына сауалнама жүргізу арқылы алынды. Бұл y параметрінің мәндеріндегі айқын емес ретінде түсіндірілетін белгісіздікке әкелді.

Процесті сипаттайтын регрессия теңдеуі мына түрде таңдалды:

 (1)

1. бағалау коэффиценттерін анықтау үшін Python тілінде бағдарламалық код жазылды:
2. *import numpy as np*
3. *from sklearn.linear\_model import LinearRegression*
4. *from sklearn.preprocessing import PolynomialFeatures*
5. *import matplotlib.pyplot as plt*
6. *import pandas as pd*
7. *from sklearn.datasets import make\_regression*
8. *df = pd.read\_excel('StableGasoline.xlsx')*
9. *x = df[['x1']] # , 'x2', 'x3', 'x4', 'x5', 'x6']]*
10. *y = df['y']*
11. *model = LinearRegression()*
12. *model.fit(x, y)*
13. *LinearRegression()*
14. *predictions = model.predict(x)*
15. *y\_pred = model.predict(x)*
16. *intercept = model.intercept\_*
17. *coefficients = model.coef\_*
18. *from sklearn.metrics import r2\_score*
19. *r2 = r2\_score(y, predictions)*
20. *n = len(y)*
21. *p = x.shape[1]*
22. *adjusted\_r2 = 1 - (1 - r2) \* (n - 1) / (n - p - 1)*
23. *sse = np.sum((y - predictions) \*\* 2)*
24. *ssr = np.sum((y\_pred - np.mean(y)) \*\* 2)*
25. *residuals = y – predictions*
26. *plt.xlabel("Шикізат")*
27. *plt.ylabel("Бензин шығымы")*
28. *plt.scatter(x, y)*
29. *plt.plot(x, y\_pred);*

Регрессия моделі шикізат құрамы, процесс шарттары және катализатордың сипаттамалары туралы мәліметтер негізінде сәтті құрылды. Талдау нәтижесінде каталитикалық крекинг қондырғысынан бензиннің шығуына факторлардың әрқайсысының әсерін көрсететін регрессия коэффициенттері алынды

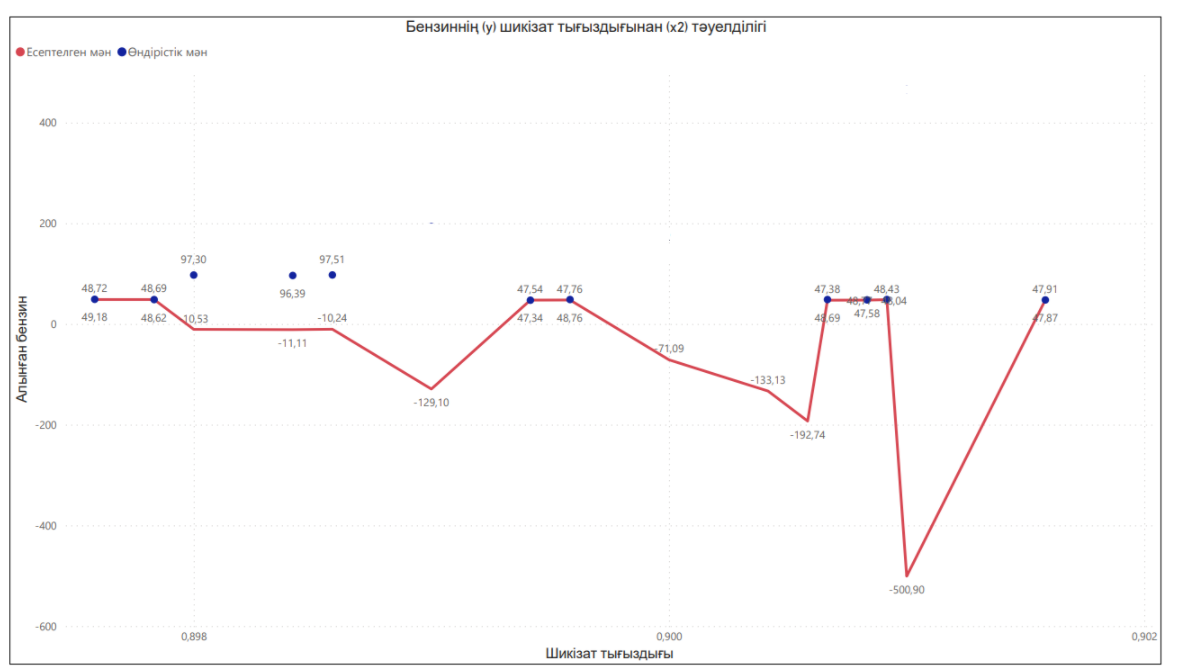
Статистикаға сәйкес (1) теңдеу коэффициенттерін есептеу нәтижелері келесідей:

c0=-30,242; c1=0,161; c2=-30,391; c3=-0,730; c4=0,185; c5=-2,734; c6=-0,004.

Осылайша, бензин шығысының параметрлерге айқын емес тәуелділігін сипаттайтын теңдеу келесідей болады:

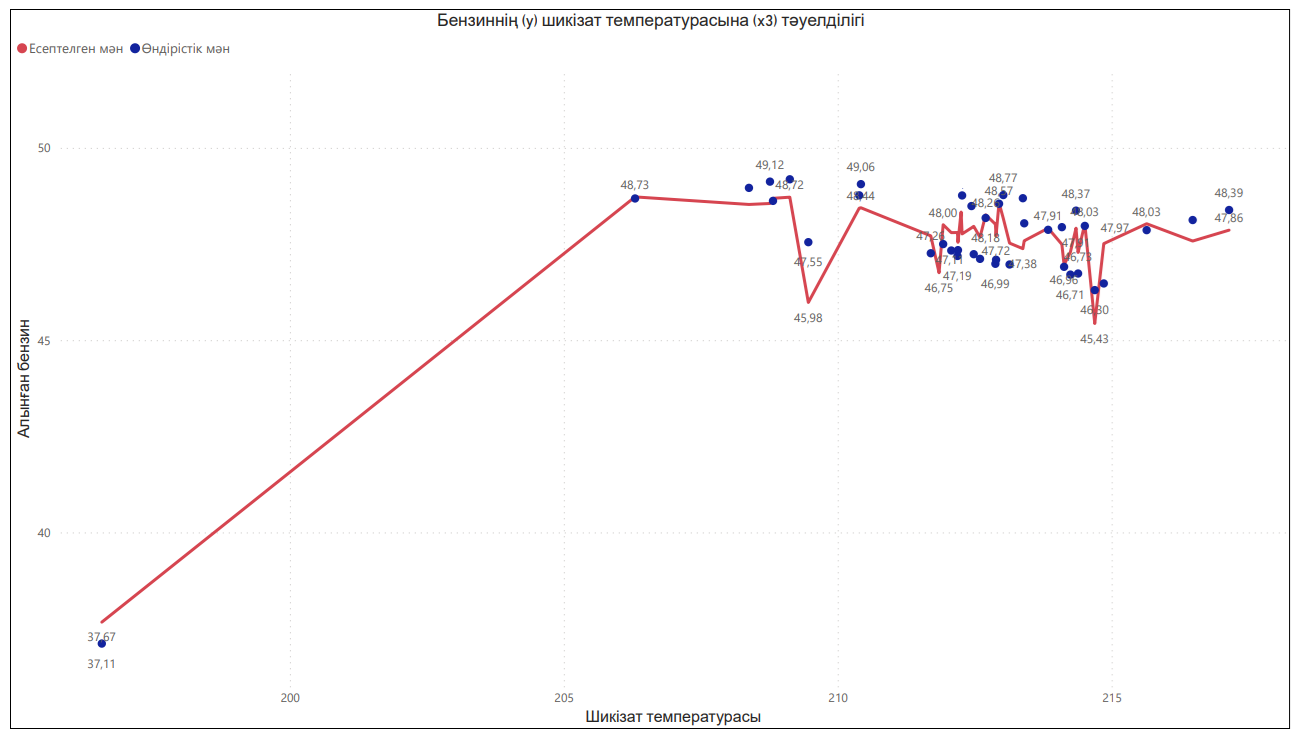
 (2)

Тұрғызылған модельге сәйкес шикізат шығыны, катализатор, реактор мен регенератордың температурасы, реактордағы қысым сияқты кіріс параметрлеріне өнімнің шығысының (тұрақты бензин) графиктері алынды (2-5-суреттер).



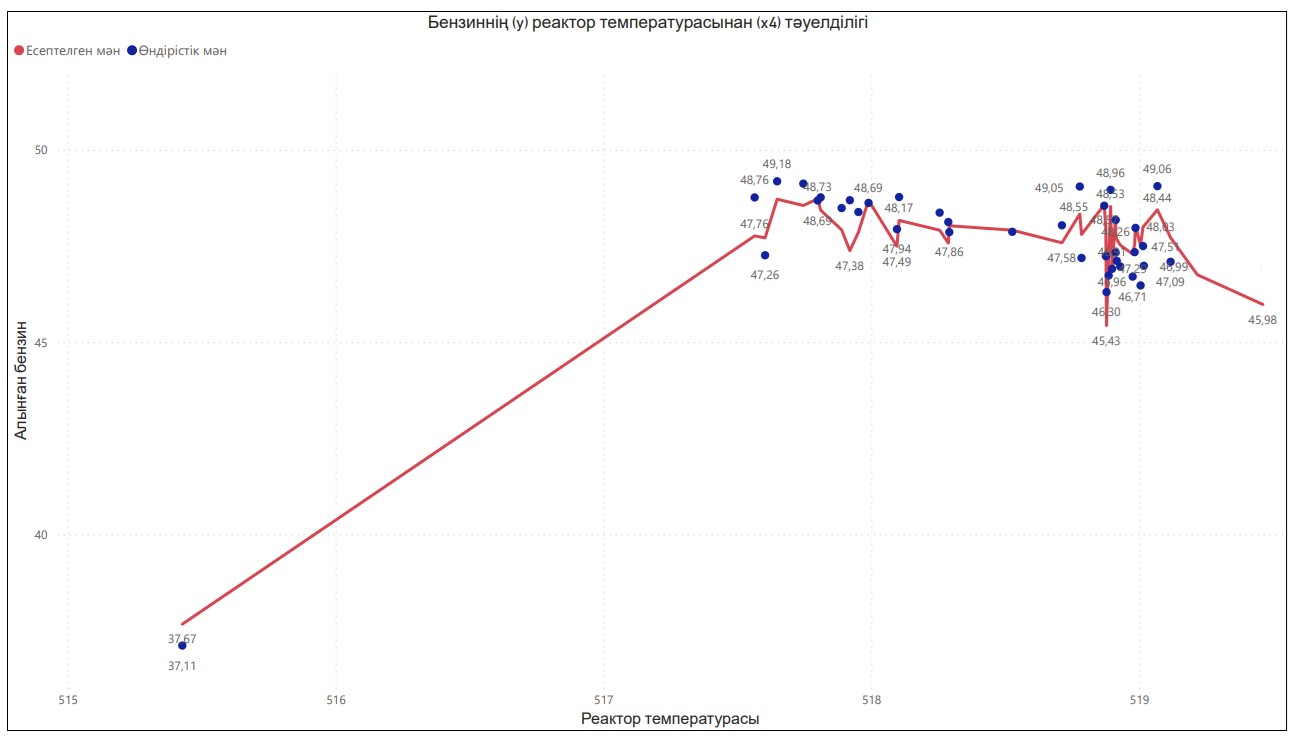
**2-сурет. Бензин шығымының шикізат тығыздығынан тәуелділігі**

2-суретте бензин шығымының шикізат тығыздығына тәуелділік графигі тұрғызылған. График бензин тығыздығының бензин шығымына кері тәуелділігін көрсетеді.

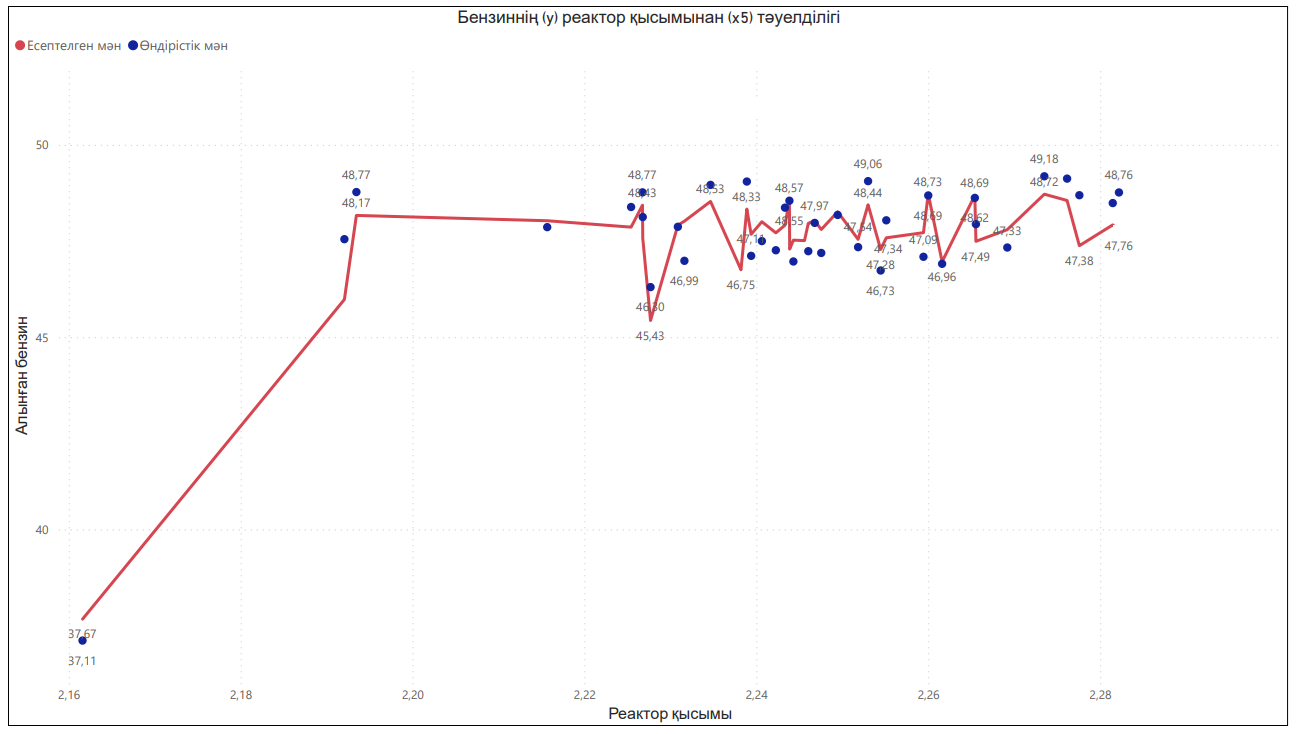


**3-сурет. Бензин шығымының шикізат температурасынан тәуелділігі**

3-суретте бензин шығымының шикізат температурасынан тәуелділігінің өндірістік және есептелген мәндерінің сәйкестігін көруге болады. Шикізат температурасын суретте көрсетілген мәнге дейін арттыру арқылы бензин шығымын айтарлықтай көбейтуге болады. Сондай-ақ тұрғызылған графиктен модель арқылы есептелген мәннің өндірістік мәндерге жақын болуы модельдің адекваттылығын көрсетеді.



**4 - сурет. Бензин шығымының реактор температурасынан тәуелділігі**



**5-сурет. Бензин шығымының реактордағы қысымнан тәуелділігі**

4,5-суреттерде бензин шығымының сәйкесінше реактор температурасы мен қысымынан тәуелділік графигін көруге болады. Каталитикалық крекинг бойынша негізгі процестер реактор бөлігінде жүргізілетіндіктен реактордың параметрлерін есепке алу маңызды. Алынған графиктен өндірістік және есептелген мәннің шамаларының жақындығын байқауға болады.

Сынауға арналған деректер жиынтығындағы модельдің дәлдігін бағалау орташа қалдық қате (MSE) 0.754-ке тең екенін көрсетті. Бұл мән болжанған мәндердің нақты мәндерден ауытқуының рұқсат етілген деңгейін көрсетеді. Детерминация коэффициенті (R2) 0.792-ге тең. R2 мәні модель тәуелді айнымалының 79% вариациясын көрсетеді, бұл көптеген әсер ететін факторлары бар өнеркәсіптік процесс үшін жақсы нәтиже. Регрессия коэффициенттері өнімнің шығуына әсер ететін ең маңызды факторлар мыналар екенін көрсетті:

- реактордың температурасы: температураның жоғарылауы бензин сияқты жеңіл өнімдердің шығуына оң әсер етеді;

- катализатордың сипаттамалары: катализатордың белсенділігі күткенге сәйкес келетін бензин шығымымен тікелей корреляцияны көрсетті.

Регрессия моделі өзін өнімнің шығуын болжаудың сенімді құралы ретінде көрсетті. Алайда, сызықтық емес тәуелділіктер немесе ескерілмеген айнымалылар арасындағы өзара әрекеттесулер болған кезде оның дәлдігі төмендеуі мүмкін. Болашақта болжамдардың дәлдігін жақсарту үшін гибридті модельдерді қолдануды қарастырған жөн.

**Қорытынды.** Бұл жұмыста Python бағдарламалық ортасында сызықтық регрессия әдістерін қолдана отырып, каталитикалық крекинг қондырғысынан өнімнің шығуын анықтау әдісі жасалды және бағаланды. Зерттеу нәтижелері ұсынылған модельдің жоғары дәлдігін көрсетті, бұл оның өнеркәсіптік тәжірибеде қолдану тиімділігін растайды. Регрессия моделін қолдану өнімнің шығымына әсер ететін негізгі технологиялық параметрлерді анықтауға және каталитикалық крекинг процесін оңтайландыруға мүмкіндік берді. Детерминация коэффициентін (R2) және қалдық қателерді талдауды қолдана отырып, модельдің сапасын бағалау болжамды мәндердің нақты деректерге сәйкестігінің жоғары дәрежесін көрсетті. Осылайша, ұсынылған тәсілді мұнай өңдеудегі технологиялық процестерді одан әрі оңтайландыру үшін қолдануға болады, бұл каталитикалық крекинг қондырғыларының өнімділігі мен тиімділігін арттырады. Болашақта болжамдардың дәлдігін жақсарту үшін машиналық оқытудың күрделі әдістерін біріктіруге болады.

**Әдебиеттер**

1. Han I.-S., Chung C.-B. Dynamic modeling and simulation of a fluidized catalytic cracking process. Part II: Property estimation and simulation // Chem. Eng. Sci. -2022. -Vol. 56. -P. 1973-1990. [DOI 10.1016/S0009-2509(00)00494-2](https://doi.org/10.1016/S0009-2509(00)00494-2)
2. Emberru R.E., Patel R., Mujtaba I.M., John Y.M. A Review of Catalyst Modification and Process Factors in the Production of Light Olefins from Direct Crude Oil Catalytic Cracking // Chemical Engineering, Faculty of Engineering & Digital Technologies. -2024. –Vol. 6(11). DOI 10.3390/sci6010011
3. Palos R., Rodríguez, E., Gutiérrez A., Bilbao J., & Arandes, J. M. Kinetic modeling for the catalytic cracking of tires pyrolysis oil // Fuel. -2022. -Vol. 309. [DOI 10.1016/j.fuel.2021.122055](https://doi.org/10.1016/j.fuel.2021.122055)
4. Orazbayev B.B., Shangitova Z.Y., Orazbayeva K.N., Serimbetov B.A., Shagayeva A.B. Studying the Dependence of the Performance Efficiency of a Claus Reactor on Technological Factors with the Quality Evaluation of Sulfur on the Basis of Fuzzy Information //Theor. Found. Chem. Eng., -2020.-Vol. 54. - P. 1235–1241. DOI 10.1134/S0040579520060093
5. Taşkin H., Kubat C., Uygun Ö., Arslankaya S. FUZZYFCC: Fuzzy logic control of a fluid catalytic cracking unit (FCCU) to improve dynamic performance // Computers & chemical engineering. -2006. - Vol. 30(5). - P. 850-863. [DOI 10.1016/j.compchemeng.2005.12.016](https://doi.org/10.1016/j.compchemeng.2005.12.016)
6. Precup R.E., Nguyen A.T., Blažič S. A survey on fuzzy control for mechatronics applications //International Journal of Systems Science.- 2024.- Vol.55(4).-P. 771-813.

DOI [10.1080/00207721.2023.2293486](https://doi.org/10.1080/00207721.2023.2293486)

1. Hе G., Zhou C., Luo T., Zhou L., Dai Y., Dang, Y., Ji, X. Online optimization of Fluid Catalytic Cracking process via a Hybrid model based on Simplified structure-Oriented Lumping and case-based reasoning // Industrial & Engineering Chemistry Research. -2020.-Vol. 60(1). - P. 412-424. DOI [10.1021/acs.iecr.0c04109](http://dx.doi.org/10.1021/acs.iecr.0c04109)
2. Beatriz Flamia Azevedo, Ana Maria A.C. Rocha, Ana I. Pereira Hybrid approaches to optimization and machine learning methods: a systematic literature review. Machine Learning. -2024. -Vol 113. - P. 4055 - 4097. DOI [10.1007/s10994-023-06467-x](http://dx.doi.org/10.1007/s10994-023-06467-x)

9.Orazbayev B.B., Kenzhebayeva S.T., Orazbayeva K. N. Development of Mathematical Models and Modelling of Chemical Technological Systems using Fuzzy-Output Systems//Applied Mathematics & Information Sciences -2019. -Vol. 13(4).-P.653-664.DOI 10.18576/amis/130417

10.Yang F., Xu M., Lei W., Lv J. Artificial intelligence methods applied to catalytic cracking processes//Big Data Mining and Analytics.-2023.-Vol. 6(3).-P. 361-380.

DOI [10.26599/BDMA.2023.9020002](http://dx.doi.org/10.26599/BDMA.2023.9020002)

11.Santander V. Kuppuraj, C.A. Harrison, M. Baldea An open source FCC model to support developing and benchmarking process control and operations strategies // Computers & Chemical Engineering. -2022. -Vol. 164. DOI [10.1016/j.compchemeng.2022.107900](https://doi.org/10.1016/j.compchemeng.2022.107900)

12.Josiah P.N., Otaraku I.J., Evbuomwan B.O. Servo and Regulatory Response of an Industrial Fluid Catalytic Cracking (FCC) Unit under Fuzzy Logic Supervisory Control // Eng. Technol. J. -2023. -Vol.41. -P. 1139-1151. DOI 10.30684/etj.2023.139485.1432

13.Orazbayev B., Boranbayeva N., Makhatova V., Rzayeva L., OspanovY., Kurmashev I., Kurmangaziyeva L. Development and Synthesis of Linguistic Models for Catalytic Cracking Unit in a Fuzzy Environment. Processes. -2024. -Vol. 12(8). DOI [10.3390/pr12081543](https://doi.org/10.3390/pr12081543)

14. Oloruntoba A., Zhang Y., Hsu C.S. State-of-the-Art Review of Fluid Catalytic Cracking (FCC) Catalyst Regeneration Intensification Technologies//Energies.-2022.-Vol.15.

DOI 10.3390/en15062061

15.Letzsch W. Fluid Catalytic Cracking (FCC) in Petroleum Refining. In: Treese, S., Pujadó, P., Jones, D. (eds) Handbook of Petroleum Processing // Springer, Cham. -2015. DOI10.1007/978-3-319-14529-7\_2

16.Amblard, B., Singh R., Gbordzoe E., Raynal L.CFD modeling of the coke combustion in an industrial FCC regenerator // Chemical Engineering Science.- 2016. - P. 731- 742.

DOI [10.1016/j.ces.2016.12.055](http://dx.doi.org/10.1016/j.ces.2016.12.055)

17. Idris M., Burn A. CFD Modelling Gas-Solid Flows in CFB // FCC Riser Reactors: Simulation Using Kinetic Theory of Granular Flow (KTGF) in a Fully Developed Flow Situation. -American institute of chemical engineers, 2008. 1-19

18. Barbosa et.al., A. C., Three-Dimensional Simulation of Catalytic Cracking Reactions in an Industrial Scale Riser Using a 11-lump Kinetic Model. Chem. Eng. Trans. -2013.- 32. -P. 637-642 DOI [10.3303/ACOS1311004](http://dx.doi.org/10.3303/ACOS1311004)

19. Sabzi H.Z. Developing an intelligent expert system for streamflow prediction, integrated in a dynamic decision support system for managing multiple reservoirs: a case study // Expert systems with applications. -2017. -Vol. 82. -№3. -C.145–163. DOI 10.1016/j.eswa.2017.04.039

***Авторлар туралы мәліметтер***

Боранбаева Н.Б. – докторант, Л.Н.Гумилев атындағы Еуразия Ұлттық университеті, сеньор лектор, Astana IT University, Астана, Қазақстан, e-mail: ades\_98@mail.ru;

Оразбаев Б.Б.- профессор, т.ғ.д., Л.Н.Гумилев атындағы Еуразия Ұлттық университеті, Астана, Қазақстан, e-mail: batyr\_o@mail.ru;

Рзаева Л.Г.- қауымдастырылған профессор, Astana IT University, Астана, Қазақстан, e-mail: [l.rzayeva@astanait.edu.kz](mailto:l.rzayeva@astanait.edu.kz);

Карабаев Ж.А. -PhD, аға оқытушы, М.Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан, e-mail:jalal45@mail.ru

Серимбетов Б.А.- техникалық ғылымдар кандидаты, қауымдастырылған профессор, Қ.Құлажанов атындағы Қазақ технология және бизнес университеті ,Астана, Қазастан, e-mail: [sba\_rnmc@mail.ru](mailto:sba_rnmc@mail.ru).

***Information about the authors***

Boranbayeva N.- doctoral student at L. N. Gumilyov Eurasian National University, senior lecturer, Astana IT University, Astana, Kazakhstan, e-mail: ades\_98@mail.ru;

Orazbayev B.- Doctor of Technical Sciences, Professor, L.N.Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: batyr\_o@mail.ru;

Rzayeva L.- Associate Professor, Astana IT University, Astana, Kazakhsta, e-mail: [l.rzayeva@astanait.edu.kz](mailto:l.rzayeva@astanait.edu.kz);

Karabayev Zh.- PhD, senior lecturer, Mukhtar Ayezov South Kazakhstan University, Shymkent, Kazakhstan, e-mail: [jalal45@mail.ru](mailto:jalal45@mail.ru).

Serimbetov B.- Candidate of Technical Sciences, associate Professor,K.KulazhanovKazakh University of Technology and Business, Astana, Kazakhstan, e-mail: [sba\_rnmc@mail.ru](mailto:sba_rnmc@mail.ru).

ҒТАМР 20.23.17

**БІЛІМ БЕРУГЕ АРНАЛҒАН АҚПАРАТТЫҚ ЖҮЙЕ ӘЗІРЛЕУ**

**1Д.К.Даркенбаев🖂, 2Г.Ж. Жакшиликова, 2Н.О.Мекебаев**

1әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

2Қазақ ұлттық қыздар педагогикалық университеті, Алматы, Қазақстан

**🖂**Корреспондент-автор: dauren.kadyrovich@gmail.com

Қазіргі ақпараттық технологиялардың дамуынан туындаған білім беру процесіндегі өзгерістер, білім беру саласындағы өте өзекті мәселелердің бірі болып табылады. Білім беру үшін ақпараттық жүйелерді дамытудың маңыздылығы, олардың функциялары, соның ішінде оқу процесін басқару, деректерді сақтау және қашықтықтан оқытуды ұйымдастыру, өз кезеңінде оқытуды білім алушылар мен педагогтардың қажеттіліктеріне бейімдеу арқылы қол жетімді және тиімді екенін атап өтеді. Мақалада білім беру мекемелерінің ерекшеліктері мен білім алушылардың талаптарын ескеретін ақпараттық жүйелерді мұқият жобалау қажеттілігі зерттелген. Жүйелердің сенімділігі, қауіпсіздігі және ыңғайлылығы сияқты негізгі аспектілер, виртуалды және кеңейтілген шындық сияқты заманауи білім беру технологиялары және олардың оқытудың интерактивтілігіне әсері де салыстырыла талданған. Талдау нәтижесінде ақпараттық жүйелерге қойылатын негізгі талаптар және олардың оқу процесін қарқындатудағы рөлі ерекшеленеді. Мақалада цифрлық трансформация жағдайында білім беру мекемелерінің алдында тұрған өзекті міндеттер талқыланған, қазіргі заманғы технологияларды бағдарлай алатын ақпараттық сауатты мамандарды даярлау мәселелері зерттелген.Зерттеу барысында білім беру жүйелерін жобалаудың теориялық негіздері талданып, ақпараттық жүйе құруға талаптар мен техникалық шешімдер қабылдау мәселелері сараланған. Бұл жүйе білім алушылардың оқыту процесін жеңілдетіп, білім алушылардың интеллектуалды дамуына оң ықпал етеді.Мақала авторлары алдағы уақытта зерттеу жұмыстарын әрі қарай жалғастыруды жоспарлап отыр.

**Түйін сөздер:** ақпарат, жүйе, технология, диаграмма, компьютер, модель, өңдеу, сервер.

**РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ОБРАЗОВАНИЯ**

**1Д.К.Даркенбаев🖂, 2Г.Ж. Жакшиликова, 2Н.О.Мекебаев**

1Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,

2Казахский национальный женский педагогический университет,Алматы, Казахстан,

е-mail: [dauren.kadyrovich@gmail.com](mailto:dauren.kadyrovich@gmail.com)

Изменения в образовательном процессе, связанные с развитием современных информационных технологий, являются одной из наиболее актуальных тем в образовании. Важность разработки информационных систем для образования обуславливает их такие явления, как управление научным процессом, хранение данных и организация дистанционного обучения, которые становятся более доступными и эффективными благодаря адаптации к потребностям образования и образования. В статье обоснована необходимость информационного проектирования с системами государственных учреждений и запроса обучающихся. Внимание также уделяется таким аспектам, как надежность, безопасность и удобство использования систем, а также современным образовательным технологиям, таким как виртуальная и дополненная реальность, и их перспективность в интерактивности текущего процесса. По результатам анализа определяются основные требования к информационным системам и их важная роль в ускорении обучения. Рассматриваются задачи, стоящие перед образовательными преобразованиями в условиях цифровой трансформации, а также вопросы подготовки специалистов, способных эффективно работать с современными технологиями. В ходе анализа разрабатываются теоретические основы проектирования образовательных систем, определяются требования к созданию информационных систем и обсуждаются сложности с использованием технических решений. Эти системы способствуют облегчению нынешних процессов и положительно влияют на интеллектуальное развитие учащихся. Авторы планируют продолжить свои исследования в данной области.

**Ключевые слова:** информация, система, технология, диаграмма, компьютер, модель, обработка, сервер.

**DEVELOPMENT OF AN INFORMATION SYSTEM FOR EDUCATION**

**1D.K.Darkenbayev🖂,** **2G.Zh. Zhakshilikova,** **2N.O.Mekebayev,**

1Al-Farabi Kazakh National University, Almaty, Kazakhstan,

2Kazakh National Women's Pedagogical University, Almaty, Kazakhstan,

е-mail: dauren.kadyrovich@gmail.com

Changes in the educational process associated with the development of modern information technologies are one of the most pressing topics in education. The importance of developing information systems for education determines such phenomena as scientific process management, data storage and distance learning organization, which become more accessible and effective due to adaptation to the needs of education and learning. The article substantiates the need for information design with the systems of state public institutions and the request of students. Attention is also paid to such aspects as reliability, security and usability of systems, as well as modern educational technologies such as virtual and augmented reality and their prospects in the interactivity of the current process. Based on the analysis, the main requirements for information systems and their important role in accelerating learning are determined. The challenges facing educational transformations in the context of digital transformation are considered, as well as the issues of training specialists who can effectively work with modern technologies. During the analysis, the theoretical foundations of designing educational systems are developed, the requirements for the creation of information systems are determined and the difficulties in using technical solutions are discussed. These systems help to facilitate current processes and have a positive effect on the intellectual development of students. The authors plan to continue their research in this area.

**Keywords:** information, system, technology, diagram, computer, model, processing, server.

**Кіріспе**.Ақпараттық технологиялардың қарқынды даму дәуірінде білім айтарлықтай өзгерістерге ұшырады. Дәрістер мен қағаз материалдарына негізделген дәстүрлі оқыту әдістері цифрлық технологияларды қолдануға негізделген интерактивті және тиімді тәсілдерге мүмкіндік береді. Бұл тұрғыда білім беру үшін ақпараттық жүйелерді әзірлеу аса маңызды. Ақпараттық жүйелер білім беру процесін өзгерте алады, оны қол жетімді, тиімді және білім алушылар мен педагогтардың қажеттіліктерін қамтамасыз ете алады.

Білім берудегі ақпараттық жүйелердің қажеттілігі оқу процесін басқарудан және білім алушылардың деректерін сақтаудан бастап, білім беру ресурстарына қол жетімділікті қамтамасыз етуге және қашықтықтан оқытуды ұйымдастыруға дейінгі көптеген функцияларды қамтиды. Бұл оқу орнын басқаруды оңтайландыруға ғана емес, сонымен қатар білім алушылардың жеке қажеттіліктеріне бейімдеу арқылы білім сапасын жақсартуға мүмкіндік береді.Алайда, тиімді ақпараттық жүйені құру дамудың барлық кезеңдерінде мұқият қарауды қажет етеді. Ақпараттық жүйені жобалау барысында, білім беру мекемесінің ерекшелігі, білім алушылардың талаптары, технологиялар мен педагогика тенденциялары сияқты факторларды ескеру қажет. Талаптарды талдау кезеңінде жүйенің олардың қажеттіліктері мен мүмкіндіктерін қанағаттандыруы үшін білім алушыларға не қажет екенін түсіну маңызды. Ақпараттық жүйенің архитектурасын жобалауда басқа платформалармен интеграциялау мүмкіндігін ескеру керек, бұл әсіресе білім беру стандарттары мен технологияларын үнемі жаңартып отыру жағдайында маңызды.

Ақпараттық жүйе (АЖ) – басқару немесе шешім қабылдау процесінде қолданылатын ақпаратты жинау, сақтау, жинақтау, іздеу және беру жүйесі. Ол әдетте ақпараттық-анықтамалық қорды (құжаттар, мәліметтербазасы, ақпараттыққоймалар), ақпаратты өңдеу және жүйемен байланыс тілін, ақпарат тасымалдаушыларды, сондай-ақ жүйенің жұмыс істеуін қамтамасыз ететін модельдер кешенін қамтиды.

Ақпараттық жүйе (АЖ) – деректерді автоматтандырылған жинау, өңдеу және басқару үшін қолданылатын техникалық және бағдарламалық құралдардың, әдістердің және қызмет көрсету персоналының өзара байланысты жиынтығы [1]. Ақпараттық және коммуникациялық технологиялар білім беруді әр түрлі деңгейде өзгертеді. Білім беру қызметіндегі жаңа басымдықтар білім беру ұйымдарының басқару қызметін ақпараттық қолдау болып табылады[2]. Мақалада жазу барысында зерттеліп жобаланған жүйе білім берудің цифрлық трансформациясы білім беру мекемесін ақпараттандырудан әлдеқайда асып түседі. Цифрлық білім берудің белгісі ақпараттық-коммуникациялық технологиялар (бұдан әрі АКТ технологиялар) негізінде кешенді көп деңгейлі интеграцияланған орта екенін атап өту қажет: білім беру контентін әзірлеу және басқару, әртүрлі деңгейдегі ақпараттық жүйелермен интеграциялау, білім алушылардың қажеттіліктеріне сәйкес білім беру траекторияларын құру [3]. Білім беру саласына ақпараттық жүйені енгізу білім алушыларды оқыту қажеттілігімен қатар жүреді, бұл жұмыстың жаңа форматына біркелкі көшуді қамтамасыз етеді. Сонымен қатар, жүйені іске қосқаннан кейін бақылау және қолдау оның ұзақ мерзімді тиімділігін қамтамасыз етуде маңызды рөл атқарады[4].

Білім беру жүйесіндегі ақпараттық жүйенің мақсаты:

* білім беру тиімділігін арттыру;
* білім берудің икемділігі мен қолжетімділігін арттыру;
* ақпараттық мәдениетті дамыту [5].

Ақпараттық жүйені құрудың маңыздылығы қазіргі білім беру мекемелерінің басты міндеттерінің бірі - білім беру үдерісіне цифрлық ортаны енгізуге бағытталған педагогтардың цифрлық сауаттылығын дамыту қажеттілігімен айқындалады. Цифрлық орта педагогтар мен білім алушыларға жаңа бағыт пен үнемі біліктілікті көтеріп отуды талап ететіні белгілі.Оқытушы цифрлық технологияларды қолдана отырып, компьютерлік бағдарламалау, ақпаратты іздеу, алмасу және коммуникация дағдыларын меңгеріп, үнемі жетілдіріп отыруы қажет [6].

**Материалдар мен тәсілдер.** Қазіргі білім – білім мен заманауи технологиялардың интеграциясы [7]. Білім және заманауи технологиялардың интеграциясы педагогпен білім алушылардың арасындағы оқытумен өзара әрекеттесу тәсілдерін өзгертетін динамикалық және көпқырлы жүйе. Виртуалды шындық (VR) және кеңейтілген шындық (AR) сияқты технологиялар интерактивті және иммерсивті білім беру орталарын құруға мүмкіндік береді. Төменде көрсетілген суретте (1-сурет) ақпараттық жүйе негізінде басқарылатын жалпы білім беретін орындардың білім беру процесінің жалпы моделі бейнеленген.



**1-сурет. Білім беру процесінің жалпы моделі.**

Білім беру процесінің жалпы моделі – білім беру процесінің әртүрлі компоненттерінің ақпараттық технологияларды қолдану арқылы бір-бірінен қалай әрекеттесеттінін сипаттайтын құрылым. Модель ақпараттық жүйелермен бірге білім беру процесіне қатысушылар арасындағы оқыту, басқару және өзара әрекеттесу процесін ұйымдастыруға және оңтайландыруға ықпал ету көрінісін бейнелеп тұр.

Білім беруге негізделген ақпараттық жүйелерге қойылатын талаптар:

1. Ақпаратты өңдеудің толықтығы – қажетті ақпаратпен қажетті көлемде қамтамасыз ету.
2. Сапа – жүйенің берілген функцияларды орындауға жарамдылық дәрежесі.
3. Ақпараттық жүйеге деген сенімділік – техникалық құралдары мен бағдарламалық жасақтамасының арасында қателіктердің туындамауы.
4. Ақпараттық жүйенің қауіпсіздігі.
5. Пайдаланудың ыңғайлылығы – әзірленетін жүйеде тиісті бөлімшенің ерекшеліктерін, онда өңделетін ақпарат пен әзірленетін құжаттардың құрамын, мазмұны мен нысандарын барынша есепке алуды талап ету.
6. Модульдік – білім беруге арналған ақпараттық жүйенің элементтерін салыстырмалы түрде тәуелсіз және функционалды түрде аяқталған стандартты бөліктерден-модульдерден құру мүмкіндігі.
7. Сүйемелдеу – сапаны жақсарту мүддесінде жүйенің ұзақ мерзімді пайдалануға және жетілдіруге жоғары дәрежеде бейімделуін қамтамасыз ету.

Білім берудің заманауи ақпараттық жүйелері оқыту және оқу процесін басқаруды айтарлықтай өзгертеді. Осы негізде білім беруге арналған ақпараттық жүйені бірнеше категорияларға бөліп қарастырып, категорияға сәйкес төмендегі 1-кестеде мысал ретінде бірнеше ақпараттық жүйелер көрсетілген.

**1-кесте. Білім беру арналған ақпараттық жүйелер**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Оқытуды басқару жүйесі | | | | | | |
| Атауы | | | Артықшылығы | Кемшілігі | | |
| Moodlе | | | Жаңа курс құру  Оқу процесін басқару  Мультимедиялық материалдар  Форумдар мен талқылаулар  Онлайн тестілеу  Оқушы үлгерімін бағалау | Платформаны орнату қиындығы  Платформаны жаңарту кезіндегі қиындықтар  Пайдаланушы интерфейсі күрделі  Шектеулі интеграциялар  Техникалық қиындықтар | | |
| Canvas | | | Интуитивті интерфейс  Курстардың икемділігі  Бағалау және кері байланыс  Басқа платформалармен интеграция | Орнатудың қиындылығы  Тегін нұсқаның шектеулі мүмкіндік-тері  Оқу уақыты | | |
| Онлайн оқыту платформасы | | | | | | |
| Coursera | | | Технологияның әртүрлі бағыттары бағыттары бойынша көпетеген курстар ұсынады  Белгілі университеттер мен компаниялардан сертификаттар мен дипломдар алу мүмкіндігі  Белсенді оқытуға ықпал ететін бейне жазбалар мен тесттер, тәжірибелік сабақтар | Көптеген курстар ақылы  Сапалы және сапасыз курстардың санының көптігі  Оқытушылармен тарапынан жеткілікті қолдаудың болмағаны | | |
| edX | | | Беделді университеттердің курстары  Көптеген курстарды тегін оқу мүмкіндігі  MicroMasters және Professional Certificates  Іс жүзінде қолдануға мүмкіндік беретін тәжірибелік тапсырмалар | Курстардың күрделілігі  Шектеулі интерактивті курстар  Платформадағы техникалық мәселелер | | |
| Бірлескен жұмыс құралдары | | | | | | |
| Google workspace | | Gmail, Google Docs, Sheets, Slides, Drive жүйелермен бір уақытта жұмыс жасау мүмкіндігі  Бірнеше пайдаланушыға құжаттарды бір уақытта өңдеуге мүмкіндік беруі | | | | Интернетке тәуелділік  Деректердің құпиялылығы  Ақылы функцияларда мүмкіндіктер көптігі |
| Microsoft Teams | | Басқа Microsoft құралдарымен интегра­ция  Бірлескен мүмкіндіктер  Командалар мен сыныптарды басқару  Бейне конференциялармен сыныптарды басқару | | | | Платформаны орнатудың қиындығы  Тегін нұсқаның шектеулі мүмкіндік­тері  Пайдаланушы интерфейсінің күрделі­лігі  Деректерді қорғау және құпиялылық |
| Виртуальды және толықтырылған шындық | | | | | | |
| Google Expeditions | Көптеген виртуальды турлармен экскурсияларға мүмкіндік  Интуитивтіинтерфейс  Мазмұнныңәртүрлілігі  виртуалдынысандарменөзараәрекеттесу мүміндігі | | | | Жақсы интернет жүйесінің болуын қажет етеді  Виртуальды мүмкіндіктердің аздығы  Техникалық ақаулардың орын алу мүмкіндігі | |
| zSpace | Тақырыптарды кеңінен түсінуге ықпал ететін виртуальды және толықтырылған шығдық тапсырмалары  Тақырыптардың кең ауқымы | | | | Платформаның оқу орындары үшйн қымбаттылығы  Аппараттық және бағдарламалық жасақтаманы орнату қиындығы  Шектеулі қолжетімділік | |

Бұл технологиялар білім беруді жаңа мазмұнмен толтыру үшін, білім алушылардың шығармашылығы мен шығармашылық қабілетін дамытатын оқытудың жаңа формалары мен әдістерін қолдану үшін қолданылған кезде тиімді болады.

Ашық білім беру жүйелерін дамыту білім беру мекемелерінің әлемдік білім беру қызметтері нарығындағы бәсекелестігінің жаңа нысандарын тудырады.

Қазіргі білім саласы заманауи технологиялардың дамуымен, интернеттің қолжетімділігінің арқасында айтарлықтар өзгерістерге ұшырауда. Білім саласында қолданысқа ие платформалардың функциональдылығы барлық білім алушыларға бір уақытта бірнеше адаммен байланыс орнатуға, ақпараттық ресурстарға, конференциялар мен вебинарларға қатысуға, туындаған мәселеге байланысты онлайн пікір таластырулар ұйымдастыруға мүмкіндік береді (2-сурет).



**2-сурет. Білім беруге арналған ақпараттық жүйенің фукнционалдылығы**

Ақпараттық жүйені әзірлеу білім беру процесін ұйымдастырудың заманауи тәсілінің ерекшелігін көрсетеді, бұл білім алушының жаңа білімді өз бетінше алу қабілетін қалыптастыруға және оларды болашақ кәсіби қызметтің міндеттерін шешуге қолдануға дайын болуға бағытталған қол жетімді интерактивті білім беру ортасын құруды қамтиды [8]. Бүгінгі таңда білім берудің әртүрлі салаларына ақпараттық жүйелерді енгізу кезінде келесі міндеттер орындалуы тиіс. Білім алушының интеллектуалды тұлғасын дамыту, жеке тұлғаны ақпараттық қоғам жағдайында жайлы өмір сүруге бейімдеу, ақпаратты қабылдау және өңдеу қабілеттерін дамыту, әлеуметтік дағдыларды жетілдіру, сондай-ақ қойылған міндетті тез талдау және оңтайлы шешім табу мүмкіндігі [9].

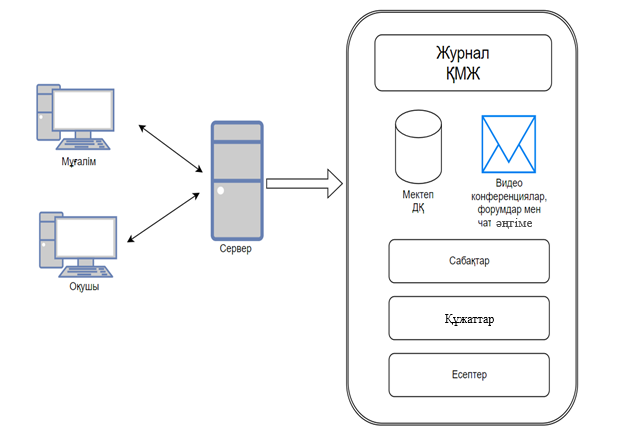
Оқу-тәрбие процесін қарқындату. Бұл тұжырымдама білім алушының алған білім сапасын жоғалтпай және оқу ұзақтығын өзгертпестен білім алушыға берілетін ақпарат көлемін ұлғайтуды білдіреді. Заманауи технологияларға бағдарланған және болашақта еңбек нарығында өзін сенімді сезінуге қабілетті ақпараттық сауатты адамды дайындау [10].

**Нәтижелер мен талдау.**Зерттеу барысында білім беру мекемелері үшін ақпараттық жүйелерді әзірлеуде қолданылатын ағымдағы тенденцияларды, әдістер мен құралдарға талдау жүргізілді.

Білім беруге арналған ақпараттық жүйелер мынадай қызмет түрлерін жүзеге асыру мүмкіндігін қамтамасыз етеді:

* білім беру процесін жоспарлайды;
* білім беру процесінің барысын және білім беру бағдарламаларын игеру нәтижелерін белгілейді;
* білім алушылар арасындағы өзара іс – қимыл, оның ішінде-интернет желісі арқылы қашықтықтан оқытады;
* білім беру қызметін басқару міндеттерін шешу үшін білім беру процесі барысында қалыптастырылатын деректерді пайдалану мүмкіндігін жоғарылатады;
* білім алушылар интернет желісінің ақпараттық білім беру ресурстарына бақыланатын қолжетімділікті арттырады;
* білім беру мекемесінің білім беру саласындағы басқаруды жүзеге асыратын органдармен, сондай-ақ басқа да білім беру мекемелерімен және ұйымдарымен өзара іс-қимылды жеңілдетеді.

Білім алушылар мен педагогтардың өзара байланысын визулизациялауға көмектесу үшін ақпараттық жүйенің архитектурасы жобаланды. Төменде 3-суртетте көрсетілген.

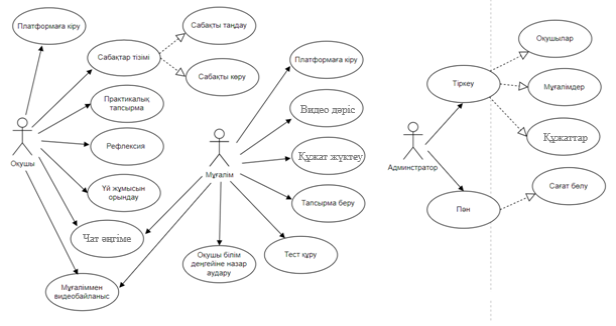


**3-сурет. Ақпараттық жүйе архитектурасы**

Білім беруге арналған ақпараттық жүйенің архитектурасы білім беру процесін қолдауға бағытталған компоненттердің құрылымдық сипаттамасы және олардың өзара әрекеттесуі көруге көмектеседі (3-сурет). Ақпараттық жүйенің әрбір бөлігін сипаттау үшін UML бірыңғай модельдеу тілі қолданылды.

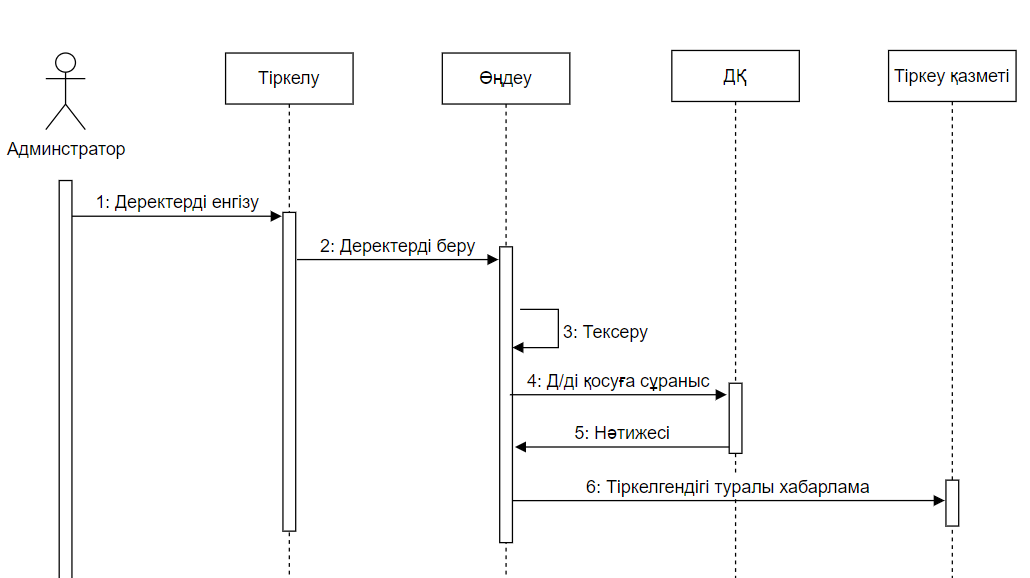
UML модельдеу тілі объектілерді, бизнес процестерді модельдеуге және жүйелік жобалауға және ұйымдық құрылымдарды көрсетуге арналған графикалық сипаттау тілі және ол бағдарламалаушыларға, жүйе аналитиктеріне, дизайнерлерге жүйенің күрделі аспектілерін түсінуге көмектеседі.UML тілінің басты мақсаты – күрделі жүйелерді визуализациялау және түсіну.

Білім беруге арналған ақпараттық жүйенің «Use case» диаграммасы – ақпараттық жүйені жобалау процесінің бір бөлігі болып табылады және жүйенің функционалдық талаптарын анықтау мен сипаттау үшін қолданылады. Диаграмма – ақпараттық жүйенің пайдаланушылармен қалай әрекеттесетінін және оның қандай функцияларды орындайтынын көрсетеді (4-сурет).



**4-сурет. Use case диаграмма**

Ақпараттық жүйенің белгілі бір сценарий шеңберінде объектілер немесе компоненттер арасындағы өзара әрекеттесуді визуализациялауға реттілік диаграммасын қолдандық (5-сурет). Реттілік диаграммасы пайдаланушылар мен ақпараттық жүйенің өзара байланысын қалыптастыруға негіз болды.



**5-сурет. Ақпараттық жүйенің реттілік диаграммасы**

Диаграмма жүйенің компоненттері арасындағы әрекеттер мен хабарламалар тізбегін көрсету арқылы пайдалану процестерін егжей-тегжейлі көрсетуге көмектеседі. Бұл use case диаграммаларында анықталған функциялардың қалай жүзеге асырылатынын жақсы түсінуге мүмкіндік береді.

Білім беруге арналған ақпараттық жүйе өте күрделі және көп уақытты талап ететін процесс. Білім беру саласы дамып келе жатқандықтан көптеген процестер жаңашылдықты, жаңа технологиялар мен жаңа ресурстарды талап етеді.

**Қорытынды.** Мақалада білім беруге арналған ақпараттық жүйені әзірлеу мәселесі зерттелген және оның білім беру саласында заманауи білім беру процесін ұйымдастыруға қажетті құрал екені дәлелденген. Ақпараттық жүйелерді білім саласында сәтті енгізу білім алушылардың да, педагогтардың да қажеттіліктерін қанағаттандырады және білім берудің тиімділігі мен икемділігін едәуір арттыруға мүмкіндік береді.

Жобаланған ақпараттық жүйенің негізгі артықшылықтары:

1. Оқытуды даралау. Жүйе білім алушының жеке қажеттіліктер мен қалауларына сәйкес материалды зерттеудің қарқыны мен дәйектілігін таңдай отырып, өздерінің білім беру траекториясын дербес құруға мүмкіндік береді.
2. Интерактивтілік және тәжірибелік фокус. Платформа білім алушылардың алған білімдері мен дағдыларын оқу процесінде белсенді қолдануға мүмкіндік беретін интерактивті жаттығулардың, тапсырмалардың және құралдардың кең ауқымын қамтиды.
3. Қашықтан және асинхронды өзара әрекеттесу. Білім алушылар теориялық материалды оқи алады, тапсырмаларды орындай алады және кезкелген педагогтан өздеріне ыңғайлы уақытта аудиториялық сабақтар шеңберімен шектелмей кері байланыс ала алады.
4. Үлгерім мониторингі және жеке қолдау. Жүйе білім алушылардың үлгерімін тиімді бақылауды, тапсырмаларды автоматтандырылған тексеруді және одан әрі оқыту бойынша жекелендірілген ұсыныстарды қамтамасыз етеді.

Ақпараттық жүйені одан әрі жетілдіру оның функционалдық мүмкіндіктерін кеңейтуге, басқа білім беру ресурстарымен интеграциялауға, сондай-ақ әртүрлі оқу бағдарламалары мен білім беру ұйымдарының ерекшеліктеріне бейімделуге бағытталуы мүмкін. Мұндай шешімдерді енгізу білім беруді неғұрлым қолжетімді және тиімді етуге мүмкіндік береді.

**Әдебиеттер**

1. G.T.Balakayeva,D.K.Darkenbayev, M.Zhanuzakov.Development of a software system for predicting employee ratings // Informatyka, Autovatyka, Pomiary w Gospodarce I Pchronie Srodowiska, 2023. – Vol.13(3). -P. 121–124.DOI 10.35784/iapgos.3723
2. Ni Wayan Aprillia Pratiwi,Ida Bagus Surya Manuaba. The Effectiveness Of A Concrete Media Assisted Project Based Learning Model On Students 'Science Competency // Journal of Education Technology. -2020. -Vol.4(4). - P. 465-470. [DOI 10.23887/jet.v4i4.27112](https://doi.org/10.23887/jet.v4i4.27112)
3. Isa, W.A.R.W.M., Suhaimi, A.I.H., Noordin, N. *et al.* Hunger hero mobile application: applying soft system methodology at a local orphanage // International Journal of Information Technology. - 2023. -Vol.15, -P.691–696. DOI 10.1007/s41870-022-01101-w
4. Ade Mukhfir Guswara, Wawan Purwanto. The Contribution of Google Classroom Application and Motivation to The Learning Outcomes of Web Programming //Journal of Education Technology. -2020. -Vol.4(4). - P. 424 - 432. [DOI 10.23887/jet.v4i4.29896](https://doi.org/10.23887/jet.v4i4.29896)
5. G.T.Balakayeva, P.Ezhichelvan, M.K.Tursynkozha. Analysis Research and Development of an Innovative Enterprise Digitalization System for Remote Work // International Journal of Mathematics and Physics.-2022. -Vol.13(1). -P.19-29.DOI 10.26577/ijmph.2022.v13.i1.02

6.D.Darkenbayev. BigData processing on the example of credit scoring //Journal of problems in computer science and information technologies. - 2023. -Vol.1.-P.50 – 61.

[DOI 10.26577/1i32jpcsit2307](https://doi.org/10.26577/1i32jpcsit2307)

7.[Yue Huang](https://link.springer.com/article/10.1007/s40593-024-00443-9#auth-Yue-Huang-Aff1), [Joshua Wilson](https://link.springer.com/article/10.1007/s40593-024-00443-9#auth-Joshua-Wilson-Aff2), [Henry May](https://link.springer.com/article/10.1007/s40593-024-00443-9#auth-Henry-May-Aff2). Exploring the Long‑Term Efects of the Statewide Implementation of an Automated Writing Evaluation System on Students’ State Test ELA Performance. // International Journal of Artifcial Intelligence in Education. -2024. -Vol.34(3). -P. 1- 30. DOI 10.1007/s40593-024-00443-9

8.D.K.Darkenbayev, A.Altybay, Zh.Darkenbayeva, N.O.Mekebayev. Intelligent Data Analysis on an Analytical platform // Informatyka, Autovatyka, Pomiary w Gospodarce I Pchronie Srodowiska. - 2024. -Vol.14(1). - P. 119-122. DOI 10.35784/iapgos.5423

9.Schwab M., Strobelt H., Tompkin J., Fredericks C., Huff C., Higgins D., Strezhnev A., Komisarchik M., King, G., &Pfister H. An Education System with Hierarchical Concept Maps // IEEE Transactions on Visualization and Computer Graphics.- 2016. -Vol.22(9). -P. 2111- 2124. DOI 10.1109/TVCG.2016.2598518

10.*S*hi, D., Cui W., Huang D., Zhang H., & Cao N. Reverse-Engineering Information Presentations: Recovering Hierarchical Grouping from Layouts of Visual Elements // Association for Computing Machinery. - 2022. -Vol.1(1). - P.1-21. DOI 10.1007/s44267-023-00010-1

***Авторлар туралы мәліметтер***

ДаркенбаевД.К. - PhD, әл-Фараби атындағы Қазақ ұлттық университетінің доцент м.а., Алматы, Қазақстан, e-mail: [dauren.kadyrovich@gmail.com](mailto:dauren.kadyrovich@gmail.com);

Жакшиликова Г.Ж. -Қазақ ұлттық қыздар педагогикалық университетінің магистранты, Алматы, Қазақстан,e-mail: gulnur201801@gmail.com;

Мекебаев Н.О. - PhD, Қазақ ұлттық қыздар педагогикалық университетінің қауымдастырылған профессор м.а., Алматы, Қазақстан, e-mail: nurbapa@gmail.com

***Information about the authors***

D.Darkenbayev - PhD, Acting Associate Professor Al-Farabi Kazakh National University,Almaty, Kazakhstan,e-mail: [dauren.kadyrovich@gmail.com](mailto:dauren.kadyrovich@gmail.com);

G.Zhakshilikova - Master's student, Kazakh National Women's Pedagogical University, Almaty, Kazakhstan,e-mail: gulnur201801@gmail.com;

N.Mekebayev - PhD, Acting Associate Professor Kazakh National Women's Pedagogical University, Almaty, Kazakhstan,e-mail: [nurbapa@gmail.com](mailto:nurbapa@gmail.com)

ГРНТИ 28.23.39

**APPLICATION OF AGGLOMERATIVE CLUSTERING FOR FORMING SKILL COMMUNITIES OF JOB VACANCIES**

**V. Ramazanova, M.Sambetbayeva,** A.Tokhmetov,**Zh. Lamasheva, S.Serikbayeva🖂,**

1L. N. Gumilyov Eurasian National University, Astana, Kazakhstan,

**🖂**Correspondent-author: [inf\_8585@mail.ru](mailto:inf_8585@mail.ru)

One of the traditional methods for community detection in knowledge graphs is agglomerative clustering. Agglomerative hierarchical clustering is a widely used type of hierarchical clustering for grouping objects based on their similarity. This method follows a bottom-up approach, beginning with each individual data point considered as an independent cluster, which are then continuously merged based on a similarity threshold between clusters. This paper focuses on the use of agglomerative clustering for analyzing skills extracted from job postings on an online recruitment platform. It describes the approach to data collection, processing, and subsequent clustering, providing an overview of linkage methods between clusters and examples of the application of various coefficients for quantitative assessment of cluster quality. An analysis of bilingual clusters in Russian and English is conducted, al-lowing for an evaluation of the versatility and adaptability of the proposed approach to analyzing the multilingual labor market in Kazakhstan. It was found that agglomerative clustering methods hold significant potential for identi-fying structured groups of skills, which can enhance the understanding of labor market trends and needs. The analysis of clusters formed in different languages confirmed the universality and adaptability of the proposed ap-proach to multilingual data.

**Keywords:** Sentence transformers, skills clustering, agglomerative clustering, silhouette coefficient, skills communities.

**ПРИМЕНЕНИЕ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ ДЛЯ ФОРМИРОВАНИЯ СООБЩЕСТВ НАВЫКОВ ВАКАНСИЙ**

**В.С.Рамазанова, М.А.Самбетбаева, А.Т. Тохметов, Ж.Б. Ламашева, С.К.Серикбаева🖂,**

1Евразийский национальный университет им. Л. Н. Гумилева, Астана, Казахстан,

e-mail:inf\_8585@mail.ru

Одним из традиционных методов обнаружения сообществ в графах знаний является агломеративная кластеризация. Агломеративная иерархическая кластеризация – это широко используемый тип иерархической кластеризации для группировки объектов на основе их сходства. Этот метод следует подходу «снизу-вверх», начиная с того, что каждая отдельная точка данных рассматривается как независимый кластер, которые затем непрерывно объединяются на основе порога сходства между кластерами. В данной статье основное внимание уделяется использованию агломеративной кластеризации для анализа навыков, извлеченных из объявлений о вакансиях на онлайн-платформе по подбору персонала. В ней описывается подход к сбору, обработке и последующей кластеризации данных, дается обзор методов связи между кластерами и примеры применения различных коэффициентов для количественной оценки качества кластера. Проводится анализ двуязычных кластеров на русском и английском языках, что позволяет оценить универсальность и адаптивность предлагаемого подхода к анализу многоязычного рынка труда в Казахстане. Было обнаружено, что методы агломеративной кластеризации обладают значительным потенциалом для выявления структурированных групп навыков, которые могут улучшить понимание тенденций и потребностей рынка труда. Анализ кластеров, сформированных на разных языках, подтвердил универсальность и адаптивность предлагаемого подхода к многоязычным данным.

**Ключевые слова:** Трансформеры предложений, кластеризация навыков, агломеративная кластеризация, коэффициент силуэта, сообщества навыков.

**ЖҰМЫС ДАҒДЫЛАРЫ ҚАУЫМДАСТЫҒЫН ҚАЛЫПТАСТЫРУ ҮШІН АГЛОМЕРАТИВТІ КЛАСТЕРЛЕУДІ ҚОЛДАНУ**

**В.С.Рамазанова, М.А.Самбетбаева, А.Т.Тохметов, Ж.Б. Ламашева, С.К.Серикбаева🖂,**

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

e-mail: inf\_8585@mail.ru

Білім графтарында қауымдастықты анықтаудың дәстүрлі әдістерінің бірі агломеративті кластерлеу болып табылады. Агломеративті иерархиялық кластерлеу объектілерді ұқсастығына қарай топтастыру үшін иерархиялық кластерлеудің кеңінен қолданылатын түрі болып табылады. Бұл әдіс әрбір жеке деректер нүктесін тәуелсіз кластер ретінде өңдеуден бастап, төменнен жоғарыға қарай әрекет етеді, содан кейін олар кластерлер арасындағы ұқсастық шегіне негізделген үздіксіз біріктіріледі. Бұл жұмыс онлайн жалдау платформасында жұмыс туралы хабарландырулардан алынған дағдыларды талдау үшін агломеративті кластерлеуді пайдалануға бағытталған. Ол деректерді жинау, өңдеу және кейіннен кластерлеу тәсілін сипаттайды, кластерлер арасындағы байланыс әдістеріне шолу жасайды және кластердің сапасын сандық бағалау үшін әртүрлі коэффициенттерді қолдану мысалдарын береді. Орыс және ағылшын тілдеріндегі қостілді кластерлерге талдау жүргізілді, бұл Қазақстанның көптілді еңбек нарығын талдауға ұсынылып отырған тәсілдің әмбебаптығы мен бейімділігін бағалауға мүмкіндік береді. Агломеративті кластерлеу әдістері еңбек нарығының тенденциялары мен қажеттіліктерін түсінуді жақсартатын құрылымдық дағдылар топтарын анықтау үшін маңызды әлеуетке ие екендігі анықталды. Әртүрлі тілдерде қалыптасқан кластерлерді талдау көптілді деректерге ұсынылған тәсілдің әмбебаптығы мен бейімделгіштігін растады.

**Түйін сөздер:** Сөйлем трансформерлері, дағдыларды кластерлеу, агломеративті кластерлеу, силуэт коэффициенті, дағдылар қауымдастықтары.

**Introduction.**The rapid growth in the volume of data on skills and qualifications available through online employment platforms represents a rich source of information on current and future labor market trends. However, due to their volume and diversity, clustering methods play a crucial role in structuring this information. In the context of graphs, communities, also known as clusters or modules, are groups of vertices that typically share common characteristics and/or perform similar functions within the graph [1].

One of the methods for community detection in graphs is hierarchical clustering. Applying agglomerative clustering methods to skills can reveal not only existing groups of similar skills but also uncover communities within knowledge graphs. Understanding and analyzing the structures of skill communities can significantly enhance career development strategies by fostering closer integration between educational and professional training and market needs.

The goal of this work is to create clusters that group semantically similar skills, which, though they may be phrased differently, retain the same overall meaning. The results of such clustering are intended to form communities of nodes in knowledge graphs, allowing for a deeper exploration of skill structures in the market.

This paper focuses on the use of agglomerative clustering for analyzing skills extracted from job postings on the online recruitment platform HeadHunter in Kazakhstan. It describes the approach to data collection, processing, and subsequent clustering, providing an overview of methodologies and examples of applying various metrics to assess cluster quality. An analysis and comparison of clusters obtained with different agglomerative clustering parameters for sets of phrases in Russian and English were conducted, allowing for an evaluation of the versatility and adaptability of the proposed approach to analyzing the modern labor market.

**Literature review.** Hierarchical clustering is highlighted as one of the traditional methods for community detection in graph structures in article [1]. Hierarchical clustering is widely used in various fields such as social networks and biology. This method begins with defining a measure of similarity between all pairs of vertices. Agglomerative algorithms, which merge clusters based on similarity, do not require a pre-specified number of clusters, which is an advantage. However, the main drawback of agglomerative hierarchical clustering is poor scalability.

Articles [2] and [3] provide a detailed overview of various agglomerative hierarchical clustering methods, such as single linkage, complete linkage, average linkage, and Ward's method. These methods are used to group objects based on their similarity, which is applicable to the analysis of skills from job postings. As noted in [3], the choice of metric and linkage method significantly impacts clustering results, necessitating a careful approach to parameter tuning. This is particularly important for text data analysis, where selecting an appropriate metric can substantially affect the quality of clusters and, consequently, skill analysis. The algorithm proposed in article [4] presents a distributed approach to agglomerative clustering that can be efficiently scaled to handle billions of objects. Article [5] introduces the concept of using the Ordered Weighted Averaging (OWA) operator to modify traditional linkages in agglomerative hierarchical clustering. This method is of interest for research aimed at enhancing the flexibility and robustness of clustering algorithms.

The use of the silhouette coefficient to evaluate cluster quality, as demonstrated in [6], allows for assessing the degree of separation of clusters and determining the optimal number of clusters. On the other hand, studies [7] highlight the limitations of the silhouette coefficient and the Calinski-Harabasz index, which can lead to misinterpretation of clustering results. In such cases, an additional expert review of the data may be helpful to improve understanding of the internal structure of the clusters.

Article [8] describes in detail how proper text preprocessing, such as sentence segmentation, tokenization, stop-word removal, and lemmatization, can significantly improve the quality and accuracy of machine learning algorithms. The study [9] examines sentence vector representations based on Transformer models in combination with various clustering methods. Article [10] discusses the semantic accuracy of models. The authors compare four freely available pre-trained sentence transformer models (all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-mpnet-base-v2, and all-distilroberta-v1) on a sample of 6,110 articles and select the most efficient model, all-mpnet-base-v2.

Article [11] addresses the complex task of clustering categorical data in data analysis. The author proposes an algorithm for clustering job vacancies based on required skills using hierarchical clustering and the Girvan-Newman method to identify job clusters. Article [12] explores the use of two popular data clustering methods: K-Means and agglomerative clustering, on data related to seafarer certification skills. The authors analyze the advantages of each approach, including K-Means ability to quickly process large volumes of data and provide clearly delineated clusters, while agglomerative clustering offers a detailed, hierarchical view of the data, useful for understanding complex structures of certification requirements. The study [13] analyzes the performance of text data clustering using the TF-IDF method, fuzzy K-means, and hierarchical agglomerative clustering on datasets such as News 20, Reuters, and email collections. The results show that hierarchical agglomerative clustering provides better performance compared to fuzzy K-means, including lower entropy and higher F-measure values, indicating clearer separation and consistency of clusters.

Article [14] describes the use of the Hierarchical Dirichlet Process (HDP) for clustering documents that define skills in an industrial environment for large IT companies. The article [15] studies thematic clustering using TF-IDF and K-Means methods in the field of information technology. The articles [16-17] consider clustering of vacancies based on skills.

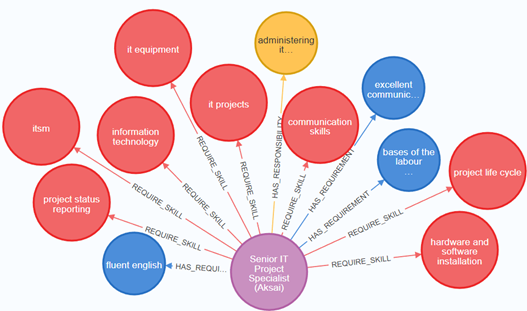
Thus, the literature review emphasizes the importance of selecting the appropriate clustering methods and metrics in data analysis. Studies show that hierarchical agglomerative clustering offers its advantages depending on data characteristics and analysis goals. Clustering efficiency increases with the use of adapted metrics and methods, which is crucial for achieving accuracy and validity of results. The implementation of innovative approaches enhances the flexibility and scalability of methods, which is critically important when working with large volumes of data. Assessing cluster quality through silhouette coefficients and other indices helps determine the optimal number of clusters, ensuring clearer data separation and consistency.

**Materials and methods.** *Collection, processing and preparation of data embeddings.*

The dataset on skills was downloaded from the online recruitment site hh.kz in Russian and English using the provided API (https://api.hh.ru/vacancies) and the Python programming language. Skills from job vacancies were collected over two months, from February 1, 2024, to March 31, 2024, from 177 localities in Kazakhstan, covering 25 professions in the information technology sector: analyst, art director, creative director, game designer, designer, artist, chief information officer (CIO), product manager, programmer, developer, development team lead, project manager, network engineer, system administrator, system engineer, information security specialist, technical support specialist, tester, chief technology officer (CTO), technical writer, system analyst, business analyst, methodologist, BI analyst, data analyst, head of analytics department, DevOps engineer, product analyst, and data scientist. The number of vacancies amounted to 5248. The number of unique skills totaled 3047.Skill phrases were treated as sentences.

Next, from the data set of vacancies and skills, a knowledge graph was built in the neo4j graph DBMS. Let's consider the ontological model of the graph of skills and vacancies. The graph includes a set of entity classes: ERecruitment = {Vacancy, Skill, Requirement, Responsibility} and a set of relationships: RRecruitment = {REQUIRE, HAS\_REQUIREMENT, HAS\_RESPONSIBILITY}.

Figure 1 presents a subgraph centered on the node representing the vacancy "Senior IT Project Specialist (Aksai)" and other related skills and vacancies. This node is connected to other nodes representing skills, requirements, and responsibilities that describe various aspects of its role and qualifications.



**Fig. 1. - Example of a subgraph of vacancies and skills. The node representing the vacancy is marked in purple, skill nodes are marked in red, requirement nodes are marked in blue, and responsibility nodes are marked in yellow**

A notable feature of this graph is the presence of many paraphrased skills that share the same meaning. Subsequently, it is necessary to construct communities of semantically similar paraphrased skills for the job vacancy skills graph.

Duplicates were removed from the set of skill phrases, all skills were converted to lower case, and phrase embeddings were generated using the paraphrase-multilingual-mpnet-base-v2 model. Paraphrase-multilingual-mpnet-base-v2 is a pre-trained transformer model designed to generate text embeddings, mapping sentences and paragraphs into a 768-dimensional dense vector space. The model is trained on over fifty languages and optimized to create vectors that effectively represent the semantic content of text, making the embeddings particularly useful for tasks related to paraphrase identification, semantic search, and text clustering. The embeddings generated by this model capture the semantic and contextual features of the text, making them suitable for grouping texts by meaning.

*Selection of clustering method, visualization and evaluation*. Agglomerative clustering is a type of hierarchical clustering used to group objects into clusters based on their similarity. It is considered a "bottom-up" method because it starts with each object initially considered as a separate cluster, and then, step by step, clusters are merged until a specified number of clusters or a distance threshold is reached (Fortunato, 2010).

The configuration for agglomerative clustering involves several parameters: n\_clusters - the number of clusters to find, should be None if distance\_threshold is specified; distance\_threshold - this parameter sets the distance threshold to stop clustering; metric - a “cosine” or “Euclidean” metric was used to measure the distance between clusters, if the relationship is “ward”, only the "Euclidean" metric is accepted; linkage - the link criterion determines what distance to use between sets of observations. The algorithm will combine pairs of clusters that minimize this criterion: “ward” minimizes the variance of the clusters being merged, “average” uses the average of the distances of each observation of two sets, “full” or “maximum” link uses the maximum distances between all observations of two sets, “single” uses the minimum distance between all observations of two sets.

The visualization of the results was performed using the scikit-learn and matplotlib libraries and the t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction technique (Scikit-learn user guide, n.d.). The main advantage of t-SNE is its ability to preserve local data structures, which makes it possible to identify clusters and groups that may be hidden in high-dimensional data.

If the true cluster labels are unknown, clustering evaluation must be performed using the model itself. The silhouette coefficient is an example of such a score, where a higher silhouette coefficient score refers to a model with more clearly defined clusters. Silhouette coefficient is a metric that measures how well each data point fits into its assigned cluster. It combines information about both the connectivity (how close data point *a* is to other points in its own cluster) and separation (how far data point b is from points in other clusters) of the data point technique (Scikit-learn user guide, n.d.).

The Calinski-Harabasz index, also known as the Variance Ratio Criterion, can also be used to evaluate the model, where a higher Calinski-Harabasz score corresponds to a model with well-defined clusters. The score is higher when clusters are dense and well-separated, which aligns with the standard concept of a cluster technique (Scikit-learn user guide, n.d.).

Our approach to determining the optimal silhouette coefficient is based on an iterative experimental programmatic process in which the distance threshold between clusters was cyclically set for different proximity metrics and linkage types to determine the maximum value of the silhouette coefficient. Ranges of the «Distance\_threshold» parameter were tested to determine the point at which a balance among the average silhouette coefficient (should be maximum), the percentage of clusters with negative silhouette coefficient (should be closer to zero), and visual inspection of the clusters was observed. The program for the cyclic process was implemented in Python using the Scikit-learn library.

**Results and discussions.** As a result of the experiments, 3047 skill phrases were processed. Table 1 presents data on various metrics (cosine, euclidean), linkage types (average, complete, single, ward), distance threshold values, the number of clusters, the average silhouette coefficient of all clusters, the percentage of clusters with a silhouette coefficient above 0.03 (the most successful clusters), the percentage of clusters with a negative silhouette coefficient, and the average Calinski-Harabasz index.

The data for Table 1, starting from the fourth column, were obtained as a result of cyclic operation of the agglomerative clustering algorithm for three initial parameters (the first three columns), with programmatically evaluation of cluster quality by silhouette and Calinski-Harabasz coefficient algorithms.

**Table 1 - Parameters and results of experiments**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Clustering parameters** | | | **Number of clusters** | **Average silhouette coefficient of all clusters** | **Percentage of clusters with silhouette coefficient greater than 0.03** | **Percentage of clusters with negative silhouette coefficient** | **Average Calinski-Harabasz index** |
| **Metric** | **Linkage** | **Distance\_**  **threshold** |
| cosine | average | 0.2 | 1669 | 0.25506 | 33.97% | 0% | 9.60789 |
| 0.24 | 1403 | 0.26764 | 40.77% | 0.07% | 9.03752 |
| 0.25 | 1335 | 0.27106 | 42.77% | 0.07% | 9.03694 |
| 0.26 | 1280 | 0.26711 | 43.44% | 0.07% | 8.88383 |
| 0.27 | 1221 | 0.26728 | 44.31% | 0% | 8.83944 |
| 0.28 | 1166 | 0.27018 | 46.14% | 0% | 8.77975 |
| 0.29 | 1121 | 0.27126 | 47.55% | 0% | 8.76839 |
| 0.3 | 1077 | 0.27041 | 48.75% | 0.09% | 8.76821 |
| 0.31 | 1023 | 0.26873 | 49.76% | 0.09% | 8.74085 |
| 0.32 | 978 | 0.26798 | 51.43% | 0.1% | 8.77425 |
| 0.4 | 636 | 0.23460 | 61.16% | 0.31% | 8.96599 |
| 0.5 | 357 | 0.17152 | 72.55% | 0.84% | 9.38007 |
| complete | 0.29 | 1342 | 0.28119 | 49.93% | 2.6% | 9.45677 |
| 0.3 | 1295 | 0.28338 | 51.43% | 2.78% | 9.48204 |
| 0.31 | 1257 | 0.28413 | 52.51% | 2.86% | 9.43280 |
| 0.32 | 1222 | 0.28468 | 53.68% | 2.86% | 9.43222 |
| 0.33 | 1175 | 0.28383 | 55.15% | 2.89% | 9.33287 |
| 0.34 | 1143 | 0.28272 | 55.47% | 3.32% | 9.31358 |
| 0.35 | 1099 | 0.28272 | 57.23% | 3.37% | 9.30610 |
| 0.4 | 916 | 0.27294 | 63.54% | 3 .93% | 9.39409 |
| single | 0.2 | 1045 | -0.18031 | 19.14% | 0.27% | 2.39889 |
| 0.29 | 406 | -0.28775 | 17.24% | 0.49% | 1.84882 |
| 0.39 | 106 | -0.19031 | 17.92% | 0.94% | 1.89379 |
| euclidean | ward | 2 | 1316 | 0.25526 | 47.72% | 4.79% | 10.69402 |
| 2.5 | 894 | 0.26155 | 62.08% | 7.49% | 10.48007 |
| 2.9 | 666 | 0.24942 | 72.67% | 9.46% | 10.89370 |
| **Note -**[**Compiled by the author**](https://context.reverso.net/%D0%BF%D0%B5%D1%80%D0%B5%D0%B2%D0%BE%D0%B4/%D0%B0%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9-%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9/Compiled+by+the+author) | | | | | | | |

When using cosine similarity with average linkage, a significant reduction in the number of clusters from 1669 to 357 is observed as the distance threshold increases from 0.2 to 0.5. In this parameter configuration, the silhouette coefficient is of primary interest, and the optimal threshold for balanced clustering is found to be between 0.28 and 0.3, where good silhouette scores are achieved while maintaining a reasonable number of clusters.

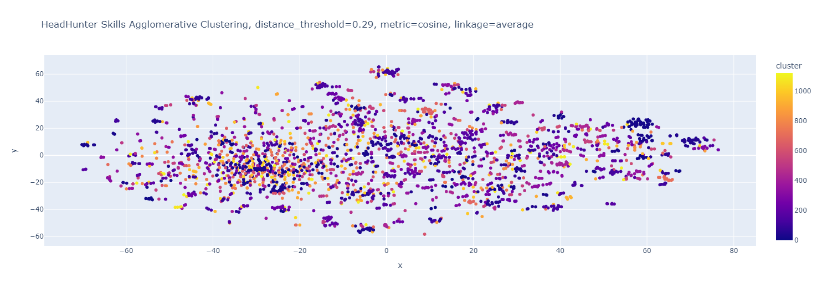
For cosine similarity with complete linkage, the silhouette coefficient varies from 0.28119 to 0.27294. The complete linkage method exhibits a significant increase in the percentage of negative silhouettes (up to 3.93% at a threshold of 0.4), which may be attributed to its tendency to form clusters by combining distantly located elements, thereby increasing group overlap.

When using cosine similarity with single linkage, poor clustering quality is evident, with negative average silhouette values and low Calinski-Harabasz index scores, indicating an unsuitable linkage method for this type of data. The method exhibits negative average silhouette values, indicating extremely weak separation between clusters.

For Euclidean similarity with Ward linkage, a threshold of 2.5 is optimal in terms of the silhouette coefficient, providing a good balance between the number and quality of clusters. The method shows an increase in the percentage of negative silhouettes as the threshold increases, reaching 9.46% at a threshold of 2.9, indicating that as the threshold increases, cluster quality deteriorates, and clusters become more diffuse and overlapping.

The average and complete methods with cosine similarity generally yield better silhouette coefficient results, particularly at intermediate threshold values, allowing for a balance between the number of clusters and their quality. The percentage of negative silhouette coefficients indicates the degree of overlap or misclassification within groups, serving as a crucial indicator of cluster structure quality. Further examination of the data leads to the selection of cosine similarity with average linkage and a distance threshold of 0.29, as this option demonstrates the best silhouette coefficient results.

Figure 2 depicts the distribution of skill clusters using average linkage with cosine similarity at a distance threshold of 0.29, where each point represents a specific skill, and the color indicates membership in one of the clusters.



**Fig. 2 - Distribution of skill clusters using average linkage with cosine similarity at a distance**

**threshold of 0.29**

The points are evenly distributed across the graph, although there is some concentration in the center. This may indicate common or frequently occurring skills that do not exhibit distinct unique properties, in contrast to rarer or more specialized skills found on the periphery.

Grouped points represent niche skills that are divided into several clusters. For instance, Table 2 shows several clusters related to testing. As described in the Materials and methods section, the clustering data was collected in Russian and English. In this case, the phrases in the data could be completely in the same language or contain elements of both languages at the same time.

**Table 2 - Example of clusters**

|  |  |  |
| --- | --- | --- |
| **Cluster number** | **Skills in the cluster** | **Silhouette coefficient** |
| 248 | тестирование мобильного приложения, тестирование мобильных приложений, диагностика смартфонов | 0.53794 |
| 130 | автоматизированное тестирование, автотестирование, автоматизация тестирования, automation test, auto-testing | 0.5163 |
| 214 | тестирование пользовательского интерфейса, тестирование новых девайсов | 0.18981 |
| 478 | проведение ux/ui тестирований, unit testing, ui test | 0.1366 |
| 116 | test case, проведение тестирований, тестирование api, функциональное тестирование, нагрузочное тестирование, диагностика, a/b testing, ручное тестирование, тестирование, модульное тестирование, регрессионное тестирование, a/b тесты, диагностика пк, a/b тестирование, регресcионное тестирование, functional testing, кроссбраузерное тестирование, тестовая документация, тест-кейсы, testing framework, тест-дизайн, a/b-тестирование | 0.09042 |
| 976 | интеграционное тестирование | 0.0 |
| 1057 | beta-тестирования | 0.0 |
| 818 | тестирование qa | 0.0 |
| **Note -**[**Compiled by the author**](https://context.reverso.net/%D0%BF%D0%B5%D1%80%D0%B5%D0%B2%D0%BE%D0%B4/%D0%B0%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9-%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9/Compiled+by+the+author) | | | |

The list of clusters containing both Russian and English terms highlights the bilingual nature of the data. Including skills in both languages can enhance the understanding of professional competency requirements across different regions, thereby increasing the universality of research in employment and education.

Some clusters overlap with each other, which may result from fuzzy boundaries between skills or the proximity of their content. Cluster 33 is the last cluster with a silhouette coefficient greater than zero.Table 3 shows cluster 33, which includes many unrelated abbreviations and monosyllabic terms.

**Table 3 - Cluster 33**

|  |  |  |
| --- | --- | --- |
| **Cluster number** | **Skills in the cluster** | **Silhouette coefficient** |
| 33 | java, верстка, vue, лвс, ui, vuejs, veam, sap fi, osi, скд, с#, релиз, jpa, lte, асутп, сапр, asana, kotlin, впр, pestel, ge/m, vlan, java ee, сопровождение по, снип, itil, otrs, presales, nat, онбординг, sono, siem, unix, java se, idef, trello, глонасс, rtos, mbedos, срм, iis, koin di, vue js, pascal, kafka для очередей, fastapi, vuetify, фронтенд, удаленка, абис, uix, jest, nestjs, nodejs, цод, vrrp, айдентика, aiohtpp, парсер, сбор трейсов, zbrush, спецпроекты, ar, lld, асу тп, стпо, ai, impa, ррл, vite, бэм, плк, геттеры, смр, olap, aris, erspan, cacti, xunit, gradle, prd, gke, maven, ремонт п, пайка, переустановка по, trc, bem, сметы, гис, эсф, perco, vr, по, опс, java/kotlin, xamarin, лендинг, vtiger, биот, эдо, эквайринг, предовос, пэк, ппм, ндв, пдс, сзз, palo alto, jvm, суиб, волс, talend, пресейл, presale, pam, estaff, актуарий, uart, serdes, отладка, saml, элеткрик, пк, екв, в2в., wcdma | 0.0063 |
| **Note -**[**Compiled by the author**](https://context.reverso.net/%D0%BF%D0%B5%D1%80%D0%B5%D0%B2%D0%BE%D0%B4/%D0%B0%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9-%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9/Compiled+by+the+author) | | | |

An analysis of the table for Cluster 33 reveals a large number of diverse skills, ranging from programming (Java, Kotlin, Node.js) to specialized technical and engineering areas (soldering, repair, software reinstallation). The very low silhouette coefficient (0.0063) indicates that this cluster is poorly separated from others, possibly due to its excessive heterogeneity.

The remaining skills do not form clusters and have a silhouette coefficient of zero. These isolated skills constitute approximately 19% of the total number of skills.

Analyzing the results, it can be stated that the distribution of silhouette coefficients reflects the quality of the clustering. Typically, the silhouette coefficient ranges from 1 to -1. In our case, clusters with a high silhouette coefficient, ranging from 1 to 0.04, exhibit acceptable quality upon inspection and can be used to create skill communities within a knowledge graph. Clusters with a silhouette coefficient between 0.03 and 0 are not suitable for community creation within the graph, as they have very blurred boundaries and include overly heterogeneous terms. Clusters with a zero silhouette coefficient include unique skills that do not cluster with other skill formulations. Clusters with a negative silhouette coefficient are of very poor quality and may include highly heterogeneous data. However, in our selected configuration, such clusters were not observed as a result of the experiments.

When the silhouette coefficient threshold for each cluster is set above 0.03, these clusters will subsequently be used to identify skill communities within job and skill knowledge graphs.

In analyzing clusters formed from skills in both Russian and English, it is important to note that this approach allows for the consideration of linguistic diversity, making the clustering process more adaptive and accurate for multilingual data. This approach enables a better understanding of data structure and provides a higher-quality analysis of skills or qualifications.

When analyzing potential clustering errors, a key pattern can be observed that may affect cluster quality, particularly in relation to phrase length. Short phrases or abbreviations often carry less unique information and may be erroneously grouped due to similarities in spelling or pronunciation. For example, abbreviations such as SCCM and SCTP might be clustered with Scrum, despite differences in their functions and applications.

For short words and abbreviations, it was necessary to improve embedding generation algorithms and conduct additional clustering. The additional training of the multilingual Paraphrase-multilingual-mpnet-base-v2 model used for generating embeddings based on the paraphrase data set "short abbreviation-full form" partially solved the problem.

Let's consider the ratio of individual skills and erroneous clusters with a positive silhouette coefficient obtained before and after training. The correctness of cluster formation was evaluated by an expert.

**Table 4 - Comparison of erroneous clusters before and after training the embedding model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Embedding model** | **Silhouette coefficient** | **Number of clusters (including single terms)** | **Percentage of Single Terms from Total Skills (3047)** | **Percentage of erroneous clusters** |
| Before training | 0,27126 | 1121 | 19,17% | 19,74% |
| After training | 0,2851 | 1407 | 24% | 13,97% |

Fine-tuning the model led to an increase in the number of clusters, which may indicate more detailed segmentation of skills. The increase in the average silhouette coefficient and the percentage of successful clusters points to improved clustering quality after fine-tuning the model. The percentage of erroneous clusters decreased, further confirming the model's improved performance. After further training, there was a partial improvement in results, but for better results it is necessary to create a separate dictionary of abbreviations.

To address the potential for scalability of our approach beyond the IT sector, we can expand its applicability to other professional domains, such as finance, healthcare, and engineering. The methodology can be adapted to these areas by incorporating domain-specific skills. Potential adaptations include the use of specialized embedding models trained on sector-relevant text corpora, which would enable more accurate clustering of context-specific skills. For instance, in the healthcare sector, medical terminologies and technical jargon could be incorporated into the embedding space to improve the identification of skill communities. Similarly, in the finance sector, the inclusion of financial industry terms and competencies could enhance the clustering process.

**Conclusions.** This study explored the application of agglomerative hierarchical clustering for analyzing skills extracted from an online recruitment platform. The research encompassed a broad range of linkage methods and distance metrics, allowing for the assessment and comparison of the effectiveness of each approach in the context of clustering semantically diverse data.

It was found that agglomerative clustering methods hold significant potential for identifying structured groups of skills, which can enhance the understanding of labor market trends and needs. The analysis of clusters formed in different languages confirmed the universality and adaptability of the proposed approach to multilingual data.

However, certain issues were also identified during the analysis, such as overlapping clusters and low silhouette coefficient values in some clustering configurations. These observations have indicated the need for further refinement of data preprocessing techniques to improve the quality of results. After further training of the embedding model on paraphrases, a partial improvement in results was observed, however, a separate dictionary of abbreviations will be created for the best results.

The findings of this study can be utilized to develop more effective data analysis tools, including the creation of educational programs and skill development strategies tailored to the evolving conditions of the labor market. The proposed methods and approaches may also find application in other areas where analyzing large volumes of textual information is required to identify hidden patterns and relationships.

In conclusion, despite certain challenges, the results achieved confirm the value and effectiveness of hierarchical agglomerative clustering as a tool for deep data analysis, opening new avenues for further research and practical applications in the field of data analytics.

***Financing.*** *This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP22783030).*

**References**

1. Fortunato, S. (2010). Community detection in graphs// Physics Reports.- 2010.-Vol.486(3-5).-P. 75-174. [DOI 10.1016/j.physrep.2009.11.002](https://doi.org/10.1016/j.physrep.2009.11.002)

2. Oti, E., & Olusola, M. (2024). Overview of agglomerative hierarchical clustering methods //British Journal of Computer, Networking and Information Technology.-2024.-Vol., 7(2).-P 14-23. [DOI 10.52589/BJCNIT-CV9POOGW](https://doi.org/10.52589/BJCNIT-CV9POOGW)

3.Tokuda, E. K., Comin, C. H., & Costa, L. F. (2022). Revisiting agglomerative clustering// Physica A: Statistical Mechanics and Its Applications.- Vol.585, Article 126433.

DOI [10.1016/j.physa.2021.126433](https://doi.org/10.1016/j.physa.2021.126433)

1. Sumengen, B. (2021). Scaling hierarchical agglomerative clustering to billion-sized datasets. arXiv, 2021. DOI [10.48550/arXiv.2105.11653](https://doi.org/10.48550/arXiv.2105.11653)
2. Cena, A., & Gagolewski, M. (2020). Genie+OWA: Robustifying hierarchical clustering with OWA-based linkages// Information Sciences.-2020.-Vol. 520.-P. 324-336.

[DOI 10.1016/j.ins.2020.02.025](https://doi.org/10.1016/j.ins.2020.02.025)

1. Gagolewski, M., Bartoszuk, M., & Cena, A. Are cluster validity measures (in)valid? //Information Sciences.-2021.-Vol. 581.- P. 620-634. [DOI 10.1016/j.ins.2021.10.004](https://doi.org/10.1016/j.ins.2021.10.004)
2. Söküt Açar, T., & Öz, N. A. The determination of optimal cluster number by silhouette index at clustering of the European Union member countries and candidate Turkey by waste indicators// Politeknik Dergisi.-2020.-Vol. 26(3). [DOI 10.5505/pajes.2019.49932](https://doi.org/10.5505/pajes.2019.49932)
3. TabassumA., Patil R. R. A survey on text pre-processing & feature extraction techniques in natural language processing.//International Research Journal of Engineering and Technology.-2020ю-Vol.7(6).- P.4864-4867. <https://www.semanticscholar.org/paper/A-Survey-on-Text-Pre-Processing-%26-Feature-in-Tabassum-Patil/f308488e996599115fe478c03b74a0b19b9a8f06>
4. Pugachev L., BurtsevM. (2021). Short text clustering with transformers, ArXiv:21.02.00541v1

DOI [10.28995/2075-7182-2021-20-571-577](http://dx.doi.org/10.28995/2075-7182-2021-20-571-577)

1. GalliC., Donos N.,CalciolariE. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis//Information.- 2024.- Vol.15(2), Article 68. [DOI 10.3390/info15020068](https://doi.org/10.3390/info15020068)
2. Ternikov A. A. (2022). Skill-based clustering algorithm for online job advertisements// Izvestiya of Saratov University. Mathematics. Mechanics. Informatics.-2022.- Vol.22(2).-P. 250-265. [DOI 10.18500/1816-9791-2022-22-2-250-265](https://doi.org/10.18500/1816-9791-2022-22-2-250-265)
3. Karthikeyan, B. A comparative study on K-means clustering and agglomerative hierarchical clustering// International Journal of Emerging Trends in Engineering Researc.-2020.-Vol. 8(5).-P. 1600-1604. [DOI 10.30534/ijeter/2020/20852020](https://doi.org/10.30534/ijeter/2020/20852020)
4. Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61-66). [DOI 10.1109/ICEEOT.2016.7754750](https://doi.org/10.1109/ICEEOT.2016.7754750)
5. Srivastava, R., Hingmire, S., & Palshikar, G. K. (2020). Clustering skills for industrial learning. TCS Research, Tata Research Design & Development Centre, Hadapsar, Pune. <https://www.cse.msu.edu/~wangzh65/AI4EDU/papers/15.pdf>
6. Biloshchytskyi, A., Shamgunova, . M. ., & Biloshchytska , S. . (2024). Exploration of the thematic clustering and collaboration opportunities in Kazakhstani research//Scientific Journal of Astana IT University.-2024.-Vol. 17(17).-P.106-121. [DOI 10.37943/17ALVR8114](https://doi.org/10.37943/17ALVR8114)
7. Ternikov, A. A. Skill-based clustering algorithm for online job advertisements// Izvestiya of Saratov University. Mathematics. Mechanics. Informatics,.-2022.- 22(2).-P. 250 - 265.

[DOI 10.18500/1816-9791-2022-22-2-250-265](https://doi.org/10.18500/1816-9791-2022-22-2-250-265)

15. Dikov M.E., Shirobokova S.N. O variante formalizacii zadachi izvlechenija kljuchevyh navykov i klasternogo analiza vakansij pri realizacii kompleksnogo instrumentarij cifrovoj proforientacii.// Inzhenernyj vestnik Dona.-2024.- № 3.- S.1-10.[in Russian]

<https://cyberleninka.ru/article/n/o-variante-formalizatsii-zadachi-izvlecheniya-klyuchevyh-navykov-i-klasternogo-analiza-vakansiy-pri-realizatsii-kompleksnogo>

***Сведение об авторах***

Рамазанова В. С.- магистр технических наук, Евразийский национальный университет им. Л. Н. Гумилева, Астана, Казахстан, е-mail: [ramazanovavs@gmail.com](mailto:ramazanovavs@gmail.com);

Самбетбаева М. А.- PhD, Евразийский национальный университет им. Л. Н. Гумилева Астана, Казахстан,

е-mail: [madina\_jgtu@mail.ru](mailto:madina_jgtu@mail.ru);

Тохметов А. Т. - доцент, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: [attohmetov@mail.ru](mailto:attohmetov@mail.ru);

Ламашева Ж. Б. - PhD, Евразийский национальный университет им. Л. Н. Гумилева, Астана, Казахстан,

е-mail: [lamasheva\_zhb@enu.kz](mailto:lamasheva_zhb@enu.kz) ;

Серикбаева С. К. - PhD, Евразийский национальный университет им. Л. Н. Гумилева, Астана, Казахстан,

е-mail: [inf\_8585@mail.ru](mailto:inf_8585@mail.ru)

***Information about the authors***

Ramazanova V. - master of technical sciences, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan, е-mail: ramazanovavs@gmail.com;

Sambetbayeva M. - PhD, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan, е-mail: [madina\_jgtu@mail.ru](mailto:madina_jgtu@mail.ru);

Tokhmetov A. - associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, е-mail: [attohmetov@mail.ru](mailto:attohmetov@mail.ru);

Lamasheva Zh. - PhD, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, е-mail: [lamasheva\_zhb@enu.kz](mailto:lamasheva_zhb@enu.kz);

Serikbayeva S. - PhD, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, е-mail: [inf\_8585@mail.ru](mailto:inf_8585@mail.ru)