Information about the authors

Saltanat Adilzhanova – doctor of technical sciences, lecturer at the department of Cryptology and Cybersecurity of the Faculty of Information Technology; Al-Farabi Kazakh National University, Republic of Kazakhstan; e-mail: asaltanat81@gmail.com. ORCID: https://orcid.org/0000-0003-1768-064X.

Murat Kunelbayev – senior research fellow at the Institute of Information and Computational Technologies of the Ministry of Science and Higher Education of the Republic of Kazakhstan; e-mail: murat7508@yandex.kz. ORCID: https://orcid.org/0000-0002-5648-4476.

Dana Sybanova* – cybersecurity analyst, master's student at the department of Cryptology and Cybersecurity of the Faculty of Information Technology; Al-Farabi Kazakh National University, Republic of Kazakhstan; e-mail: dsybanovaa@gmail.com. ORCID: https://orcid.org/0009-0003-4684-5586.

Поступила в редакцию 04.02.2025
Поступила после доработки 05.04.2025
Принята к публикации 09.04.2025

https://doi.org/10.53360/2788-7995-2025-2(18)-3

IRSTI: 20.19.27

A.M. Amantay*, Zh.M. Makhambetali
Kazakh-British Technical University,
050000, Republic of Kazakhstan, Almaty, Tole bi street, 59
*e-mail: ai_amantay@kbtu.kz

CLUSTERING AND CLASSIFICATION OF DISEASES USING STOCHASTIC DYNAMIC OPTIMIZATION

Abstract: *This study presents a new approach to the optimization of Natural Language Processing (NLP) techniques for medical entity recognition and disease classification. By leveraging patient queries and PubMed article abstracts, the research uses advanced extraction methods to identify biomedical entities and diseases from medical texts. Diseases are grouped using a combination of TF-IDF and K-means clustering, and classification models are then applied to predict disease clusters based on known entities. A key innovation of this work is the use of Stochastic Dynamic Optimization to fine-tune parameters, significantly enhancing clustering and classification performance.*

Experimental results demonstrate that the proposed method improves the accuracy of extraction and classification, outperforming traditional methods in terms of precision and scalability. This scalable and efficient approach to biomedical data analysis has the potential to support future clinical decision-making, enable personalized medicine, and provide valuable healthcare insights, ultimately contributing to improved patient outcomes and more effective research workflows.

Key words: *Machine Learning, Stochastic Dynamic Optimization, Disease clustering, PubMed abstracts, Medical Entity Recognition, Data Extraction, Healthcare data optimization.*

Introduction

The rapid expansion of biomedical literature and the increasing volume of patient-generated data present significant challenges for healthcare systems. Understanding unstructured medical texts from the growing medical information becomes an inefficient task for traditional data analysis methods, which struggle to keep up with expansion rates [1, 2]. The fundamental technology Named Entity Recognition (NER) encounters difficulties when processing medical entities because medical language proves complex along with its domain-specific jargon and information presentation irregularities across medical sources. Medical institutions require modern techniques that produce scalable evaluations using efficient methods with high accuracy to analyze their extensive healthcare databases. Current methods fail to build interconnected frameworks, which unite patient-dependent inquiries with medical contents to provide meaningful results while being easy to interpret. The existing gap prevents healthcare providers from using personalized and efficient decision-making approaches during clinical situations.

Our investigation establishes a new medical entity detection system with disease classification capabilities through the integration of modern extraction tools alongside clustering algorithms and operation optimization. The alignment between patient queries and disease clusters creates more precise disease predictions that professionals can easily understand through

enhanced interpretability. The research develops systems, which link unstructured biomedical information to clinical use resulting in solutions able to support both better medical decisions and customized treatments and expanded healthcare delivery. A more powerful adaptive tool emerges from our study, which will facilitate the handling of complex biomedical data thus leading to superior patient care and streamlined healthcare study processes.

Literature review

Similar research works have been done in this sphere. Da Silva et al. [3] propose a method that jointly performs NER and relation extraction (RE) on oncology. In their work, they process unstructured clinical data coming from Brazilian Portuguese hospitals and clinics, using BERT and BiLSTM models to identify entities and categorize relations accurately. Nevertheless, the model is adaptable to other clinical specialties, but it requires re-annotation and adaptation if used in domains without annotated resources or pre-trained models.

Miah et al [4] introduced MedNER, a BiLSTM based framework to identify drug and disease entities in biomedical texts. On a COVID-19 dataset, MedNER was tested and achieved F1 score of 98%. Although the model integrates domain specific embeddings, the fact that it uses COVID-19 datasets introduces a risk of lack of generalizability to the model and needs to be fine-tuned on broader and more diverse corpora for broader applicability.

Chen et al. [5] and Tang et al. [6] addressed the boundary ambiguity for Chinese medical texts with boundary-enhanced framework. Performance across models was demonstrated, with the model performing very well on entities with observable start and stop points. While this approach is effective with Chinese medical text, other languages could benefit from a fine tuning for datasets with different syntactic structure.

In their work, Zhang et al. [7] applied BiLSTM+CRF model to entity recognition in gastrointestinal endoscopy reports. Using BIOES encoding, they were able to fine tune the model on annotated reports and obtain F scores of 97,71% for Position and 95,57% for Organs. Nevertheless, it relies on annotated data and is not feasible for environments with scarce data.

The authors of "Patient Clustering Optimization with K-Means in Healthcare Data Analysis" [8] use the K-Means clustering algorithm to improve grouping methods for healthcare data. The researchers use interpretive thinking with deductive methods to enhance efficiency and test resilience of their algorithm. The research demonstrates how demographic patient groups require proper clustering which receives its accuracy validated through strict verification procedures. Future research could pursue scalable models combined with hybrid approaches alongside interdisciplinary partnerships. Ershadi et al. [9] presents a novel approach for medical classification enhancement through their paper “Applications of Dynamic Feature Selection and Clustering Methods to Medical Diagnosis” that combines dynamic feature selection with adaptive clustering techniques. The study implements K-means and fuzzy c-means and particle swarm optimization clustering techniques, which address imbalanced medical datasets characteristics.

Materials and methods

Our workflow included parsing of the data, medical entity recognition out of queries and parsed abstracts, vectorization, disease clustering and classification and optimization. Full methodology is shown in Figure 1:

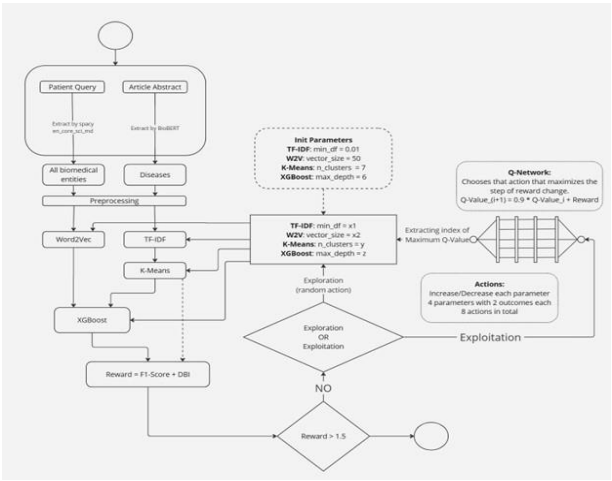

Figure 1 – Proposed methodology

Dataset. In this work we used the PubMed Central (PMC) dataset conducted by Zhao et al. [10]. This dataset includes over 167,000 patient summaries and over 3.1 million annotations connecting patients to relevant articles, as well as over 293,000 annotations connecting patients to similar cases. The dataset structure consists of patient_id, patient_uid, PMID, file_path, title, patient summary, age, gender, relevant_articles, similar_patients.

Entity recognition and Text Vectorization. Our first step in this research after parsing all the articles presented in PMC dataset is extracting structured information from unstructured textual data – named entity recognition. We identified key biomedical entities including diseases, organs, pathological formations, etc. and other critical terms by analyzing patient queries and biomedical literature abstracts. These entities are used as important features for downstream analytical tasks such as clustering, classification and understanding disease characteristics.

In this research work, as shown in Figure 1, we used the pre trained NER model *en_ner_bc5cdr_md* from scispaCy that is specifically designed for biomedical text processing. The model was trained on corpora like PubMed and excels at detecting and classifying biomedical terms into pre-defined categories like DISEASE, CHEMICAL and ANATOMICAL_ENTITY. It is lightweight and computational efficient and perfectly suitable for tasks done on CPUs.

We used two complementary approaches for text vectorization: Diseases extracted from abstracts and BioBERT embeddings of entities in user queries. TF-IDF also highlighted statistically significant disease terms by highlighting different terms and down-weighting less informative ones. Through this dual strategy, we were able to take advantage of the advantages of both methods. TF-IDF allowed us to discover previously hidden patterns in disease data that bridge gaps and connect. User queries were given depth through BioBERT embeddings, which made their complexities into representations between queries and the disease clusters.

K-Means Clusterization and Annotation. We conducted entity recognition for the user queries and the abstracts of the PMC dataset. The main goal was to identify mentions of diseases so that particular attention could be given to related medical concepts. We used a state-of-the-art Named Entity Recognition model fine-tuned on biomedical text, which shows good performance in the detection of disease-related entities. The disease names were, therefore, extracted for each article in the dataset, hence forming a list to be used for the foundation of clustering and analysis thereafter. Articles sharing similar disease characteristics were grouped together. Importantly, the optimal number of clusters was determined using SDO. This method allowed the systematic exploration and fine-tuning of cluster configurations to maximize the cohesion and separation of the clusters.

Disease Classification with Cluster Labels. The proposed classification pipeline executes KMeans clustering with biomedical entity extraction from patient queries to achieve precise disease predictions (Figure 2). This system utilizes disease classifications alongside particular features in queries to achieve comprehensive diagnosis capabilities while analyzing patient cases.

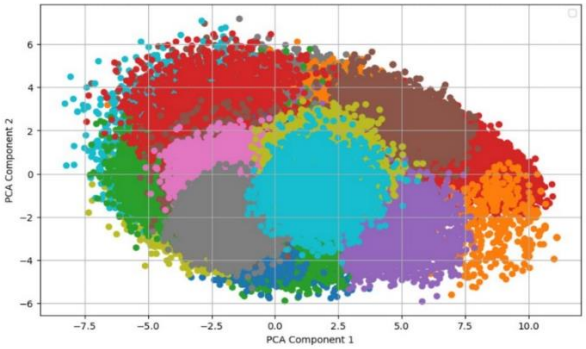

Figure 2 – KMeans clustering of diseases with PCA visualization

The KMeans cluster labels act as the target variable (y) while representing the disease groups which were identified through biomedical abstracts. The input features (X) consist of biomedical entities obtained from patient inquiries which are transformed into BioBERT embeddings to preserve domain-related contextual information.

The Decision Tree Classifier served for classification due to its interpretation abilities and performance benefits. The model parameter max_depth was optimized through SDO by finding the best balance between F1-score and understanding of the model to prevent either overfitting or underfitting.

The integration of cluster-based labels together with query-specific entities brought better classification precision while providing easier understandability to results. The model detects patient queries in relation to disease groupings, which results in interpretable prediction outputs. The SDO optimization process delivered optimal efficiency together with accurate performance.

Stochastic dynamic optimization. In order to refine these biomedical NLP pipelines, we offer a method that is driven by optimization, and applies in the domains of NER, clustering, and classification. In our approach, we use Stochastic Dynamic Optimization (SDO) to maximize exploration and exploitation in tuning min_df, vector_size, n_clusters, max_depth and other parameters. For classification, it optimizes the F1-score, and for clustering, the Davies Breident Index (DBI) is maximized which helps in the downstream tasks, not only precision but also cohesion.

The network architecture consists of:

- Input Layer: state vector, $s \in R^4$;
- Hidden Layers: Two fully connected layers with ReLU activation;
- Output Layes: $Q(s, \alpha) \in R^s$.

The Q-Network is trained to minimize the error between the predicted Q-value and the target Q-value, defined as:

$$Loss = E\left[\left(Q(s,\alpha) - \left(r + \gamma \max_{\alpha'} Q(s',\alpha')\right)^2\right)\right] \quad (2)$$

This iterative update ensures that the Q-Network learns to approximate the true Q-values, improving the agent's ability to make optimal decisions over time.

To balance exploration (discovering new actions) and exploitation (leveraging known information), we employ an ε-greedy policy. The agent selects an action α as follows:

$$\alpha = \begin{cases} argmax_{\alpha'} Q(s,\alpha'), with\ probability\ 1-\varepsilon\ (exploitation) \\ Random\ action,\ with\ probability\ \varepsilon\ (exploration) \end{cases} \quad (3)$$

The parameter ε decays over time, enabling the agent to prioritize exploration in the initial training phases and shift toward exploitation as it gains experience. This decay is critical to ensuring the agent avoids suboptimal solutions while converging to high-reward actions.

The reward function, central to guiding the agent's learning, combines clustering and classification performance:

$$Reward = F_1 score + \frac{1}{1+DBI} \quad (4)$$

This approach allows the agent to dynamically discover optimal parameter configurations, ensuring balanced performance across both clustering and classification tasks. By approximating Q-values with a neural network and systematically exploring the parameter space, the agent effectively learns a policy that maximizes the reward function and achieves consistent improvements in biomedical pipeline performance.

Results

Overall, using TF-IDF and BioBERT together for vectorization provided a robust representation of the dataset, balancing between precision and semantic richness. The one mapped the terrain, the other traced paths; always communicating user queries to disease clusters in the language of context and meaning.

The number of clusters were chosen by the optimization algorithm. The graph on Figure 3 shows how the DBI value changes when different numbers of clusters are used to group diseases. DBI measures cluster quality and shows better results when the value stays low. DBI shows its lowest numbers of 7 and 8 clusters which proves strong cluster unity and distinct cluster divisions:

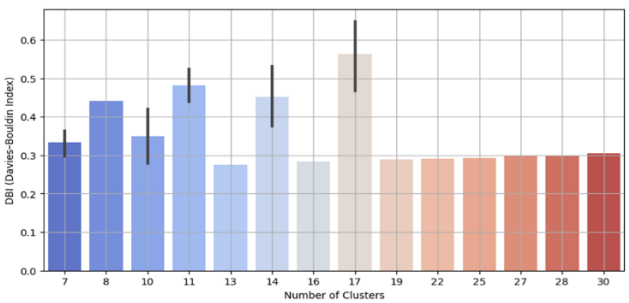

Figure 3 – DBI and number of clusters

The graphs below, in Figure 4, show how reward performance changes when vector size, min_df, n_clusters, and max_depth values change in the system. The series of plots show how these settings impact the reward system's results which demonstrate system performance. The top-left graph shows the connection between reward and vector dimension size. The reward score grows steadily up to 100 dimensions before showing a rapid rise. A vector size of 200 delivers the optimal performance because it provides better representation strength than smaller or larger embedding dimensions. At 120 dimensions, the embeddings show clear improvements because they achieve a sufficient level of semantic understanding.

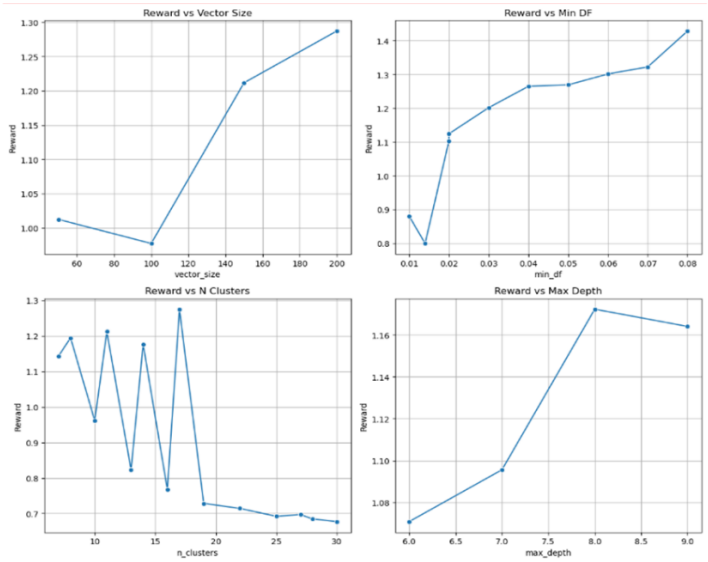

Figure 4 – Dependency of Reward from each parameter

Figure 5 illustrates the relationship between Word2Vec vector size and classification performance (Accuracy and F1-score). Our study evaluates vector sizes from 50 to 200, showing that performance drops when the size decreases below 100, as reduced dimensionality leads to a loss of semantic information. Performance improves as vector size increases from 150 to 200, enhancing contextual understanding. The highest Accuracy (*0.75*) and F1-score (*0.72*) are achieved at 200 dimensions, demonstrating that larger vector spaces contribute to better disease classification. Shaded bars indicate measurement variations.

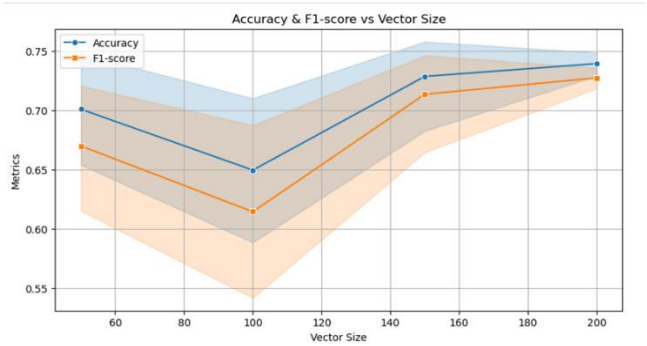

Figure 5 – Accuracy, F1 score and Vector Size

Discussion

Our research proves that disease classification together with clustering needs hyperparameter tuning to achieve significant results. Performance improves when using bigger word vector dimensions although additional increases after 200 dimensions do not significantly benefit accuracy versus computational cost.

Clustering and classification achieve their best results through identifying proper cluster numbers. If cluster numbers exceed their optimal limit they split essential disease traits while lower cluster counts fail to recognize significant medical distinctions. The study demonstrates 17 clusters represent the most clinically relevant solution. Max_depth controls Decision Tree depth in a way that

optimal results emerge after seven or eight levels while demonstrating the necessity of cross-validation to prevent overfitting.

Medical text processing techniques for extraction of biomedical entities from patient data along with PubMed articles create essential fundamental elements for disease classification systems. SDO-based hyperparameter optimization serves as an enhancement for classification results by finding balanced solutions between accuracy rates and interpretability alongside computational demands.

Conclusion

This work shows that robust feature extraction techniques and fine-tuned hyperparameters are essential for efficient disease classification and clustering. Our method combines word vector embeddings, clustering, classification, and biomedical entity recognition, optimizing each stage for increased scalability and accuracy.The main conclusions indicate that the best results are obtained with a vector size of 200, min_df between 0,03 and 0,05, 17 clusters, and a Decision Tree depth of 7–8. Model tuning is further improved by stochastic dynamic optimization (SDO), which maintains a balance between computational efficiency and accuracy.

This system facilitates automated clinical decision-making and biomedical research by organizing PubMed abstracts and patient queries into meaningful illness clusters. For even more accurate categorization, future research should concentrate on better illness mention monitoring and increased contextual representation using BERT-based embeddings.

References

1. Choudhary P. An Intelligent Chatbot Design and Implementation Model Using Long Short-Term Memory with Recurrent Neural Networks and Attention Mechanism / P. Choudhary, S. Chauhan // Decision Analytics Journal. – 2023. – Vol. 9, № 100359. https://doi.org/10.1016/j.dajour. 2023.100359.
2. Online Biomedical Named Entities Recognition by Data and Knowledge-Driven Model / Cao Lulu et al // Artificial Intelligence in Medicine. – 2024. – Vol. 150, № 102813. https://doi.org/10.1016/ j.artmed.2024.102813.
3. Exploring Named Entity Recognition and Relation Extraction for Ontology and Medical Records Integration / D.P. da Silva et al // Informatics in Medicine Unlocked. – 2023. – Vol. 43, № 101381. https://doi.org/10.1016/j.imu.2023.101381.
4. Medical Named Entity Recognition (MedNER): A Deep Learning Model for Recognizing Medical Entities (Drug, Disease) from Scientific Texts / Miah M.S. Ullah et al // IEEE EUROCON 2023 – 20th International Conference on Smart Technologies. – 2023. – P. 158-162. https://doi.org/10.1109/EUROCON56442.2023.10199075.
5. Named Entity Recognition Based on Boundary Enhanced for Chinese Electronic Medical Records / X. Chen et al // 2023 12th International Conference of Information and Communication Technology (ICTech). – 2023. – P. 73-77. https://doi.org/10.1109/ICTECH58362.2023.00025.
6. Tang L. Named Entity Recognition in Chinese Medical Texts Based on RoBERTa-WWM-IDCNN-CRF / L. Tang, J. Kong, L. Xu // 2024 IEEE 25th China Conference on System Simulation Technology and Its Application (CCSSTA). – 2024. – P. 315-319. https://doi.org/10.1109/CCSSTA 62096. 2024.10691735.
7. Zhang Y. Named Entity Recognition of Medical Examination Reports Based on BiLSTM+CRF Model / Y. Zhang, F. Zhang // 2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). – 2023. – P. 340-344. https://doi.org/10.1109/AINIT59027. 2023.10212675.
8. Patient Clustering Optimization with K-Mean in Healthcare Data Analysis / A.K. Rai et al // 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI). – 2023. – P. 1-7. https://doi.org/10.1109/ICAIIHI57871.2023.10489428.
9. Ershadi Mohammad Mahdi. Applications of Dynamic Feature Selection and Clustering Methods to Medical Diagnosis / Ershadi Mohammad Mahdi, Seifi Abbas // Applied Soft Computing. – 2022. – Vol. 126, № 109293. https://doi.org/10.1016/j.asoc.2022.109293.
10. A Large-Scale Dataset of Patient Summaries for Retrieval-Based Clinical Decision Support Systems / Z. Zhao et al // Scientific Data. – 2023. – Vol. 10, № 909. https://doi.org/10.1038/s41597-023-02814-8.

А.М. Амантай*, Ж.М. Махамбетали
Казахстанско-Британский технический университет,
050000, Республика Казахстан, улица Толе би, 59
*e-mail: ai_amantay@kbtu.kz

КЛАСТЕРИЗАЦИЯ И КЛАССИФИКАЦИЯ ЗАБОЛЕВАНИЙ С ПОМОЩЬЮ СТОХАСТИЧЕСКОЙ ДИНАМИЧЕСКОЙ ОПТИМИЗАЦИИ

В данном исследовании представлен новый подход к оптимизации методов обработки естественного языка (NLP) для распознавания медицинских объектов и классификации заболеваний. Используя запросы пациентов и аннотации статей в PubMed, в статье применяются передовые методы извлечения информации для выявления биомедицинских сущностей и заболеваний. Заболевания группируются с помощью комбинации TF-IDF и кластеризации K-means, а затем применяются модели классификации для предсказания кластеров заболеваний на основе известных сущностей.

Ключевым новшеством данной работы является использование стохастической динамической оптимизации для точной настройки параметров, что значительно повышает эффективность кластеризации и классификации. Кроме того, исследование анализирует влияние размеров векторных представлений слов, количества кластеров и глубины дерева решений на итоговую точность модели. Экспериментальные результаты показывают, что предложенный метод повышает точность извлечения и классификации медицинских знаний, превосходя традиционные методы по точности и масштабируемости. Этот масштабируемый и эффективный подход к анализу биомедицинских данных может помочь в принятии клинических решений, обеспечить персонализированную медицину и предоставить ценные сведения о здравоохранении, что будет способствовать улучшению состояния пациентов и повышению эффективности исследовательских процессов.

Ключевые слова: *Машинное обучение, Стохастическая динамическая оптимизация, Кластеризация заболеваний, PubMed статьи, Распознавание медицинских объектов, Оптимизация медицинских данных.*

А.М. Амантай*, Ж.М. Махамбетали
Қазақстан-Британ техникалық университеті,
050000, Қазақстан Республикасы, Төле би көшесі, 59
*e-mail: ai_amantay@kbtu.kz

СТОХАСТИКАЛЫҚ ДИНАМИКАЛЫК ОҢТАЙЛАНДЫРУ АРҚЫЛЫ ДИАГНОЗДЕРДІ КЛАСТЕРЛЕУ ЖӘНЕ ЖІКТЕУ

Бұл зерттеу медициналық нысандарды тану және диагноздерді жіктеу үшін табиғи тілді өңдеу тәсілдерін. (NLP) оңтайландырудың жаңа әдістерін ұсынады. Пациенттердің сұраныстары мен PubMed мақалаларының аннотацияларын пайдалана отырып, мақалада биомедициналық нысандар мен ауруларды анықтау үшін ақпаратты алудың озық әдістері қолданылады. Диагноздер TF-IDF және K-means кластерлеу комбинациясы арқылы топтастырылады, содан кейін белгілі субъектілерге негізделген ауру кластерлерін болжау үшін жіктеу үлгілері қолданылады.

Бұл жұмыстың негізгі жаңалығы – параметрлерді дәл баптау үшін стохастикалық динамикалық оңтайландыруды қолдану. Бұл кластерлеу мен жіктеудің дәлдігін едәуір арттырады. Сонымен қатар, зерттеу сөздердің векторлық көріністерінің өлшемдерінің, кластерлер санының және шешім ағашының тереңдігінің модельдің соңғы дәлдігіне әсерін талдайды. Эксперименттік нәтижелер ұсынылған әдіс дәлдік пен масштабталу бойынша дәстүрлі әдістерден асып түсу арқылы медициналық нысандарды алу және жіктеу дәлдігін арттыратынын көрсетеді. Биомедициналық деректерді талдаудың бұл ауқымды және тиімді тәсілі клиникалық шешімдер қабылдауға, жекешелендірілген медицинаны қамтамасыз етуге, пациенттердің жағдайын жақсартуға сондай-ақ осы бағыттағы зерттеу процестерінің тиімділігін арттыруға ықпал ететін денсаулық сақтау туралы құнды зерттеулер жасауға көмектеседі.

Түйін сөздер: *Машиналық оқыту, Стохастикалық динамикалық оңтайландыру, Диагноздарды кластерлеу, PubMed мақалалары, Медициналық нысанды тану, Денсаулық сақтау саласындағы деректерді оңтайландыру.*

Information about the authors

Aigerim Amantay* – master's student, Kazakh-British Technical University, Almaty, Kazakhstan; e-mail: ai_amantay@kbtu.kz. ORCID: https://orcid.org/0009-0007-0172-6400.

Zhandos Makhambetali – master's student, Kazakh-British Technical University, Almaty, Kazakhstan; e-mail: z_makhambetali@kbtu.kz. ORCID: https://orcid.org/0009-0006-5656-4253.

Сведения об авторах

Айгерим Амантай* – магистрант, Казахстанско-Британский технический университет, Алматы, Казахстан; e-mail: ai_amantay@kbtu.kz. ORCID: https://orcid.org/0009-0007-0172-6400.

Жандос Махамбетали – магистрант, Казахстанско-Британский технический университет, Алматы, Казахстан; e-mail: z_makhambetali@kbtu.kz. ORCID: https://orcid.org/0009-0006-5656-4253.

Авторлар туралы мәліметтер

Айгерим Амантай* – магистрант, Қазақстан-Британ техникалық университеті, Алматы, Қазақстан; e-mail: ai_amantay@kbtu.kz. ORCID: https://orcid.org/0009-0007-0172-6400.

Жандос Махамбетали – магистрант, Қазақстан-Британ техникалық университеті, Алматы, Қазақстан; e-mail: z_makhambetali@kbtu.kz. ORCID: https://orcid.org/0009-0006-5656-4253.

Received 18.02.2025
Accepted 14.04.2025

https://doi.org/10.53360/2788-7995-2025-2(18)-4

МРНТИ: 50.47.29

B.A. Mailykhanova[1]*, Sh.K. Koshimbayev[2], A. Khabay[2], R.A. Jamasheva[1], S. Abdukarimov[1]
[1]Almaty technological university,
050012, Republic of Kazakhstan, Almaty, str. Tole bi, 100
[2]Satbayev University,
050013, Republic of Kazakhstan, Almaty, str. Satbayev, 22
*e-mail: bulgyn@mail.ru

AI-DRIVEN OPTIMIZATION OF CRUDE OIL REFINING PROCESSES

Abstract: *The integration of Artificial Intelligence (AI) in industrial automation has led to significant improvements in efficiency, predictive maintenance, and cost reduction. This study investigates the application of AI-based control systems in crude oil refining, focusing on optimizing process efficiency, minimizing maintenance costs, and improving system reliability. Traditional control methods, which rely on pre-defined rules and manual intervention, often lead to inefficiencies and unplanned downtime. In contrast, AI-driven automation enables real-time data analysis, predictive decision-making, and adaptive control mechanisms.*

Our research utilizes advanced machine learning models, including artificial neural networks (ANNs) and gradient boosting algorithms, to optimize process parameters. These models were trained using historical operational data and validated through simulation-based testing. Results demonstrate that AI-driven systems reduce maintenance costs by up to 30%, improve predictive accuracy by 25%, and enhance energy efficiency by 15%. Furthermore, intelligent control systems show high adaptability to variations in crude composition, enabling more robust and sustainable operations.

To address the challenge of AI model transparency, the study incorporates explainable AI (XAI) techniques such as SHAP and LIME to improve interpretability and support trust in automated decision-making – particularly in safety-critical refinery processes. These tools provide insights into feature importance and model behavior, facilitating better understanding by engineers and operators.

Despite the performance benefits, the adoption of AI in industrial environments faces challenges, including high initial investment costs, integration with legacy systems, and cybersecurity risks. The paper proposes strategies to mitigate these barriers, such as phased deployment, secure system architecture, and hybrid control models combining AI with rule-based logic.

This research underscores the transformative potential of AI in refining operations and contributes to the development of reliable, transparent, and cost-effective automation solutions for the energy sector.

Key words: *Artificial intelligence, industrial automation, process optimization, predictive maintenance, explainable AI, energy efficiency, machine learning, refining control process.*

Introduction

The growing complexity of industrial operations and increasing energy and cost efficiency demands have accelerated the integration of Artificial Intelligence (AI) into automation systems. In crude oil refining, one of the most technically and economically intensive industries, conventional control systems often fail to adapt dynamically to fluctuations in feedstock composition and process disturbances. These limitations result in suboptimal performance, increased energy consumption, and unplanned downtimes.