https://doi.org/10.51885/3134-8025_JICS_2026_2_12
SRSTI 76.29.37

AN INTELLIGENT MODEL FOR ANALYZING BLOOD GASOMETRY BASED ON ENSEMBLE MACHINE LEARNING METHODS FOR DIAGNOSING DIABETIC KETOACIDOSIS AND SUPPORTING CLINICAL DECISION-MAKING

ДИАБЕТТІК КЕТОАЦИДОЗДЫ ДИАГНОСТИКАЛАУ ЖӘНЕ КЛИНИКАЛЫҚ ШЕШІМДЕРДІ ҚОЛДАУ ҮШІН АНСАМБЛЬДІК МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІНЕ НЕГІЗДЕЛГЕН ҚАН ГАЗОМЕТРИЯСЫНЫҢ КӨРСЕТКІШТЕРІН ТАЛДАУДЫҢ ИНТЕЛЛЕКТУАЛДЫ МОДЕЛІ

ИНТЕЛЛЕКТУАЛЬНАЯ МОДЕЛЬ АНАЛИЗА ПОКАЗАТЕЛЕЙ ГАЗОМЕТРИИ КРОВИ НА ОСНОВЕ АНСАМБЛЕВЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДИАГНОСТИКИ ДИАБЕТИЧЕСКОГО КЕТОАЦИДОЗА И ПОДДЕРЖКИ ПРИНЯТИЯ КЛИНИЧЕСКИХ РЕШЕНИЙ

I. Uvaliyeva[1*], **Zh. Rakhmetullina**[1], **Zh. Nurumov**[1],
R. Mukasheva[1], **F. Amenova**[2]
[1]D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan
[2]S. Amanzholov East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan
*Corresponding author: Uvaliyeva Indira, e-mail: iuvalieva@ektu.kz

Keywords:
diabetic ketoacidosis, blood gas analysis, machine learning, ensemble methods, Random Forest, XGBoost, medical diagnostics, intelligent data analysis

ABSTRACT
Diabetic ketoacidosis is one of the most severe acute complications of diabetes mellitus and requires timely diagnosis and early detection of metabolic disorders. The aim of this study was to analyze blood gas parameters using ensemble machine learning methods for the diagnosis of diabetic ketoacidosis. Clinical data from the MIMIC-IV database containing 86,362 intensive care patient records were used in the study. Ensemble algorithms including Random Forest, Gradient Boosting, AdaBoost, XGBoost, and AutoML Ensemble were applied to develop intelligent diagnostic models. Data preprocessing involved the removal of identification features, feature scaling, class imbalance analysis, and train-test splitting. Model performance was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC metrics. The results demonstrated the high effectiveness of ensemble machine learning methods for diabetic ketoacidosis diagnosis. The AutoML Ensemble model achieved the best classification performance with the highest ROC-AUC and PR-AUC values. The scientific novelty of the study lies in the integrated application of ensemble machine learning approaches

for intelligent analysis of blood gas parameters. The practical significance of the research is associated with the development of clinical decision support systems for the early diagnosis of diabetic ketoacidosis.

Түйінді сөздер:
диабеттік кетоацидоз, қан газометриясы, машиналық оқыту, ансамбльдік әдістер, Random Forest, XGBoost, медициналық диагностика, интеллектуалдық деректер талдауы

ТҮЙІНДЕМЕ

Диабеттік кетоацидоз қант диабетінің ең қауіпті жедел асқынуларының бірі болып табылады және метаболикалық бұзылыстарды уақтылы анықтауды талап етеді. Зерттеудің мақсаты – диабеттік кетоацидозды диагностикалау үшін қан газометриясы көрсеткіштерін ансамбльдік машиналық оқыту әдістері негізінде талдау. Зерттеуде MIMIC-IV деректер базасының клиникалық мәліметтері пайдаланылды, олар қарқынды терапия бөлімшелеріндегі 86 362 пациенттің бақылауларын қамтиды. Интеллектуалдық модельдерді құру үшін Random Forest, Gradient Boosting, AdaBoost, XGBoost және AutoML Ensemble алгоритмдері қолданылды. Деректерді алдын ала өңдеу идентификациялық белгілерді жоюды, параметрлерді масштабтауды, сыныптар теңгерімсіздігін талдауды және деректерді оқыту және тестілеу жиынтықтарына бөлуді қамтыды. Модельдердің тиімділігі Accuracy, Precision, Recall, F1-score, ROC-AUC және PR-AUC метрикалары арқылы бағаланды. Зерттеу нәтижелері ансамбльдік машиналық оқыту әдістерінің диабеттік кетоацидозды диагностикалауда жоғары тиімділігін көрсетті. Ең жоғары нәтижелерге ROC-AUC және PR-AUC көрсеткіштері бойынша AutoML Ensemble моделі қол жеткізді. Зерттеудің ғылыми жаңалығы қан газометриясы көрсеткіштерін интеллектуалдық талдау үшін ансамбльдік әдістерді кешенді қолданумен байланысты. Жұмыстың практикалық маңыздылығы диабеттік кетоацидозды ерте диагностикалауға арналған дәрігерлік шешім қабылдауды қолдау жүйелерін әзірлеу мүмкіндігімен анықталады.

Ключевые слова:
диабетический кетоацидоз, газометрия крови, машинное обучение, ансамблевые методы, Random Forest, XGBoost, медицинская диагностика, интеллектуальный анализ данных

АННОТАЦИЯ

Диабетический кетоацидоз является одним из наиболее опасных острых осложнений сахарного диабета, требующим своевременной диагностики и раннего выявления метаболических нарушений. Целью исследования являлся анализ показателей газометрии крови на основе ансамблевых методов машинного обучения для диагностики диабетического кетоацидоза. В работе использовались клинические данные базы MIMIC-IV, включающие 86 362 наблюдения пациентов отделений интенсивной терапии. Для построения интеллектуальных моделей применялись алгоритмы Random Forest, Gradient Boosting, AdaBoost, XGBoost и AutoML Ensemble. Предварительная обработка данных включала удаление идентификационных признаков, масштабирование параметров, анализ дисбаланса классов и разделение выборки на обучающую и тестовую части. Оценка качества моделей выполнялась с использованием метрик Accuracy, Precision, Recall, F1-score, ROC-AUC и PR-AUC. Результаты исследования показали высокую эффективность ансамблевых методов при диагностике диабетического кетоацидоза. Наиболее высокие показатели классификации продемонстрировала модель AutoML Ensemble, обеспечившая максимальные значения ROC-AUC и PR-AUC. Научная новизна исследования заключается в комплексном применении ансамблевых методов машинного обучения для интеллектуального анализа показателей газометрии крови.

Практическая значимость работы связана с возможностью создания систем поддержки принятия врачебных решений для ранней диагностики диабетического кетоацидоза.

INTRODUCTION

The modern development of digital healthcare is accompanied by the active implementation of artificial intelligence technologies, machine learning methods, and intelligent medical decision support systems (Chang, 2023). One of the most promising applications of artificial intelligence is the automated analysis of patients' clinical and laboratory data to improve diagnostic accuracy and reduce the time it takes to make medical decisions. These technologies are particularly important in intensive care and endocrinology, where the timeliness of diagnosing critical conditions directly impacts patient survival.

Diabetes mellitus is one of the most common chronic diseases of our time (Nadhiya et al., 2024). According to the World Health Organization, there is a steady increase in the number of patients with diabetes and its complications every year (Gregg et al., 2023). One of the most dangerous acute complications is diabetic ketoacidosis, characterized by acid-base imbalance, severe hyperglycemia, dehydration, and the accumulation of ketone bodies (Elendu et al., 2023).

Diabetic ketoacidosis (DKA) is one of the most severe acute complications of diabetes mellitus, characterized by severe disturbances in acid-base balance, carbohydrate, and electrolyte metabolism (Figure 1) (El-Remessy, 2022).

Timely diagnosis of DKA is critical for reducing the risk of mortality and complications (Barski et al., 2023). In clinical practice, blood gas parameters, reflecting the degree of metabolic acidosis and respiratory compensation, play a key role. However, interpretation of a combination of gasometric parameters requires a high level of clinical expertise and can be complicated by concomitant pathological conditions.

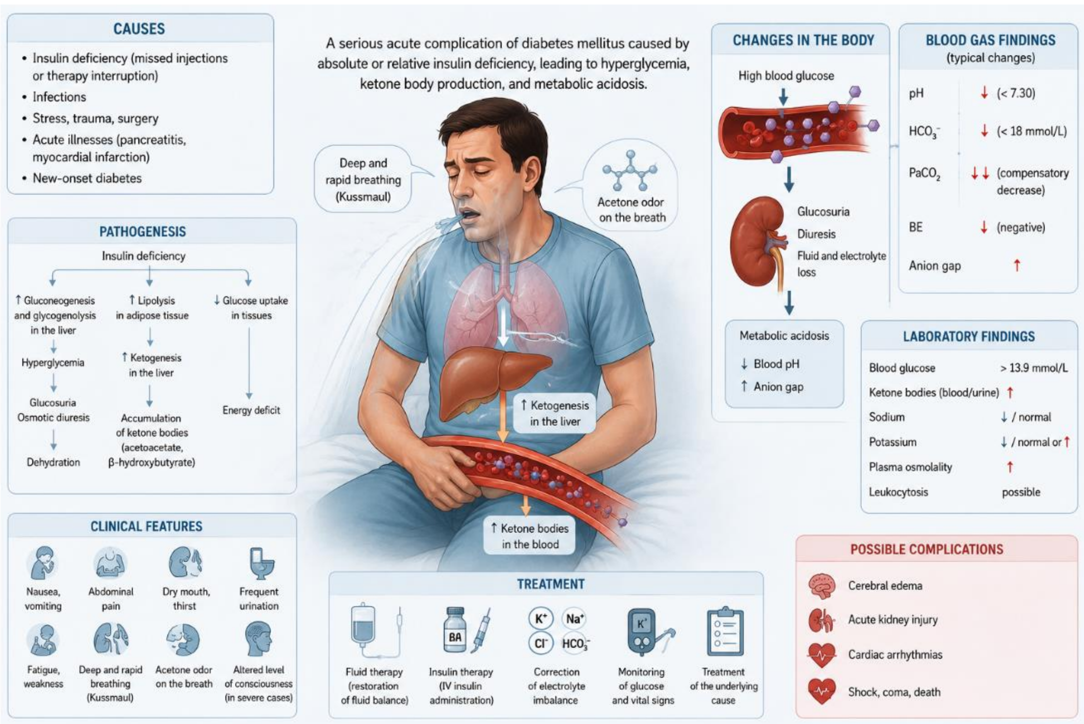

Figure 1. Clinical and Metabolic Manifestations of Diabetic Ketoacidosis

Note – compiled by the authors

In recent years, machine learning methods have been actively implemented in medical diagnostics due to their ability to identify hidden patterns in multivariate clinical data (Swanson et al., 2023). Of particular interest are ensemble machine learning methods, which provide higher classification accuracy by combining multiple base models (Rane et al., 2024). Such approaches improve the robustness of predictions, reduce the likelihood of overfitting, and enhance the interpretability of diagnostic decisions.

Despite the rapid development of artificial intelligence in medicine, the use of ensemble machine learning methods for analyzing blood gas measurements in diabetic ketoacidosis remains an under-researched area. Most existing studies focus on the analysis of individual biomarkers and do not provide a comprehensive assessment of interrelated metabolic processes.

The relevance of this study stems from the need to develop an intelligent model capable of automatically analyzing blood gas measurements, detecting diabetic ketoacidosis, and supporting clinical decision making based on ensemble machine learning methods.

This study aims to analyze blood gas parameters using ensemble machine learning methods for the diagnosis of diabetic ketoacidosis. pH, partial pressure of carbon dioxide (pCO_2), bicarbonate concentration (HCO_3^-), lactate level, anion gap, and oxygen saturation are considered diagnostically significant parameters. Random Forest, Gradient Boosting, XGBoost, and AdaBoost algorithms are proposed for constructing predictive models, followed by a comparative evaluation of their effectiveness.

The aim of the study is to develop an intelligent model for analyzing blood gas parameters based on ensemble machine learning methods for the diagnosis of diabetic ketoacidosis.

To achieve this goal, the following objectives were defined:

- creation of a clinical dataset based on MIMIC-IV;
- analysis of blood gas analysis parameters;
- implementation of ensemble machine learning methods;
- application of class balancing methods;
- evaluation of model performance;
- analysis of the diagnostic significance of features.

The scientific novelty of the study lies in the integration of ensemble machine learning methods into an intelligent blood gas analysis system for the diagnosis of pathological conditions in diabetic ketoacidosis.

MATERIALS AND METHODS

Research datasets and their preprocessing

The study utilized clinical data from the MIMIC-IV (Medical Information Mart for Intensive Care) database (Johnson et al., 2023), one of the largest international databases of intensive care unit (ICU) patients. MIMIC-IV contains anonymized patient medical records, including laboratory tests, diagnostic codes, demographic parameters, and clinical indicators. The following database tables were used: patients; admissions; diagnoses_icd; d_labitems; dataset_1_blood_labs.

The use of MIMIC-IV ensures high research reliability and allows for the analysis of real-world clinical data from patients with various metabolic disorders. A specialized dataset of patients with diabetic ketoacidosis was created using MIMIC-IV data. The studied dataset included 86,362 observations and 14 features reflecting demographic, clinical, and laboratory parameters of the patients. Each row of the dataset corresponded to an individual clinical case. ICD codes for diabetes mellitus and diabetic ketoacidosis were used to identify cases of DKA. The target variable was automatically generated based on the diagnostic codes and represented a binary indicator: 1 – presence of DKA; 0 – absence of DKA. A fragment of the database is presented in Table 1.

The dataset structure included the following groups of features:

1) Demographic features: age – patient's age; gender – patient's sex.

2) Blood gasometry and biochemical parameters: pH – blood acidity; bicarbonate – bicarbonate concentration (HCO_3^-); anion_gap – anion gap value; glucose – blood glucose level.

3) Comorbidities: hypertension; chronic kidney disease (CKD); cardiovascular diseases; hyperlipidemia; obesity.

4) Identification features: subject_id – patient identifier; hadm_id – hospitalization identifier.

Table 1. Excerpt from the clinical database of patients with DKA

subject_id	hadm_id	age	gender	DKA	ph	bicarbonate	anion_gap	glucose
10000690	25860671	86.0	1	False	7.37	22.0	13.0	105.0
10000935	25849114	57.0	1	False	7.41	27.0	12.0	99.0
10000980	20897796	80.0	0	False	7.33	21.0	15.0	143.0
10001843	26133978	76.0	0	False	7.39	24.0	12.0	115.0
10001884	26184834	77.0	1	False	7.36	23.0	14.0	126.0
Note – compiled by the authors								

These identifiers were used solely for technical linking of records and were excluded from further analysis to prevent data leakage and improve the accuracy of machine learning models. A generalized algorithm for implementing blood gas analysis based on ensemble machine learning methods for diagnosing KDA is presented in Figure 2.

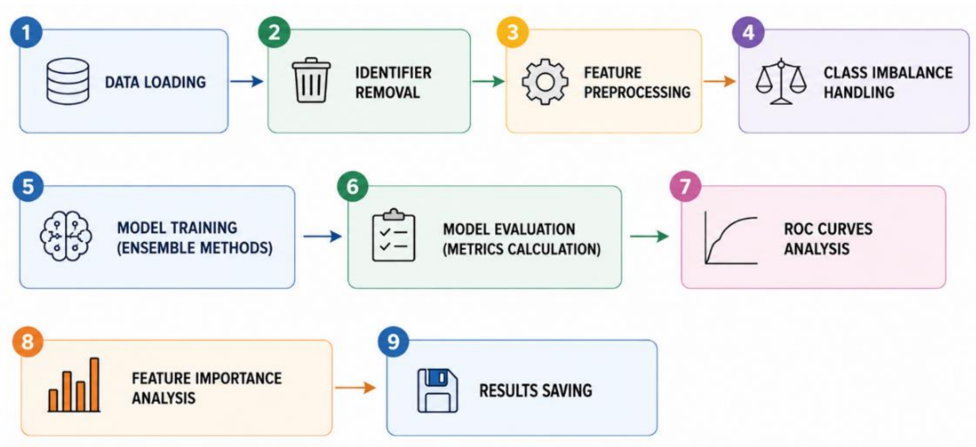

Figure 2. Algorithm for analyzing blood gasometry parameters based on ensemble machine learning methods for diagnosing KDA

Note – compiled by the authors

Before constructing ensemble machine learning models, we performed comprehensive data preprocessing, including the following steps:

- checking the data structure and feature types;
- missing value analysis;
- removing identifying features (subject_id, hadm_id);
- encoding categorical features;
- scaling quantitative variables using StandardScaler;
- splitting the sample into training and testing parts;
- analyzing the class imbalance of the target variable.

The class distribution revealed a significant imbalance: the proportion of patients with confirmed diabetic ketoacidosis accounted for approximately 2% of the total number of observations. Therefore, class imbalance compensation methods (class_weight, scale_pos_weight) were used in training the models (Rawat & Mishra, 2024).

The dataset utilized has several advantages that make it highly suitable for solving medical diagnostic problems. Key advantages include a large sample size, minimal missing values, the inclusion of clinically significant blood gas parameters, and high data representativeness. Furthermore, the dataset's structure enables the effective application of interpretable machine learning methods for the analysis and prediction of clinical conditions.

Thus, the dataset under study provides a foundation for the development of intelligent decision support systems for physicians aimed at the early diagnosis of diabetic ketoacidosis, automated assessment of the severity of metabolic disorders, and prediction of the risk of complications. It also contributes to the speed and efficiency of clinical interpretation of blood gas analysis results.

Ensemble methods in machine learning

Ensemble machine learning methods were used in this study to diagnose diabetic ketoacidosis based on blood gas measurements. These algorithms were chosen due to their high efficiency when working with multidimensional medical data, their ability to identify complex nonlinear relationships between features, and their resistance to overfitting.

Ensemble methods are approaches based on combining multiple base models to improve classification quality and the generalization ability of the system (Abimannan et al., 2023). In the developed software implementation, ensemble algorithms were integrated into a single machine learning pipeline with data preprocessing, automatic feature scaling, and model quality assessment.

The following ensemble algorithms were implemented in this study:

- Random Forest;
- Gradient Boosting;
- AdaBoost;
- XGBoost.

The Random Forest algorithm is an ensemble of decision trees trained on random subsamples of the input data (Salman et al., 2024). The final decision is formed based on a vote by all trees in the ensemble. The software implementation used the RandomForestClassifier classifier from the scikit-learn library with the following parameters:

- Number of trees: n_estimators = 120;
- Automatic class balancing: class_weight = "balanced";
- Parallel execution of calculations: n_jobs = -1.

The use of the Random Forest algorithm makes it possible to effectively model complex nonlinear dependencies, significantly reduce the risk of overfitting, quantify the significance of individual features, and ensure high model stability to the presence of noise data (Zhou et al., 2023). Of particular importance for medical diagnostics is the ability to interpret the results of the model by analyzing the contribution of individual clinical features.

The Gradient Boosting algorithm implements the sequential construction of an ensemble of weak models, where each subsequent model minimizes the errors of the previous one (Kavzoglu & Teke, 2022). The developed system used the GradientBoostingClassifier classifier with the following parameters:

- number of iterations: n_estimators = 120;
- Learning rate: learning_rate = 0.05;
- depth of trees: max_depth = 3.

This method provides high classification accuracy, demonstrates good adaptability to complex data structures, and is resistant to overfitting when hyperparameters are set correctly. Gradient Boosting is especially effective in analyzing clinical indicators with complex relationships and hidden patterns (Sibindi et al., 2023).

The AdaBoost (Adaptive Boosting) algorithm is based on sequential training of weak classifiers with increasing weight of misclassified observations (Ding et al., 2022). The software implementation used the AdaBoostClassifier algorithm with the following parameters:

- number of base models: n_estimators = 120;
- Learning rate: learning_rate = 0.05.

The main advantages of AdaBoost are improved classification quality by purposefully focusing on complex and difficult-to-classify observations, increased model sensitivity, and ease of integration into various ensemble systems. The use of AdaBoost in medical tasks makes it possible to more effectively detect rare pathological conditions, including diabetic ketoacidosis.

The XGBoost (Extreme Gradient Boosting) algorithm is an improved implementation of gradient boosting with mechanisms for regularization and optimization of calculations (Li et al., 2022). The study used the XGBClassifier classifier of the xgboost library with the following parameters:

- number of trees: n_estimators = 150;
- depth of trees: max_depth = 4;
- Learning rate: learning_rate = 0.05;
- random subsample of observations: subsample = 0.9;
- random subsample of features: colsample_bytree = 0.9;
- Class imbalance compensation parameter: scale_pos_weight.

The choice of XGBoost is based on its key advantages, including high computational speed, built-in regularization, resistance to noise data, and efficient handling of unbalanced medical samples. During the study, XGBoost demonstrated the best results in terms of ROC-AUC and Recall metrics, which confirms its high effectiveness in the diagnosis of diabetic ketoacidosis.

Methods for evaluating the effectiveness of models

A set of statistical and diagnostic metrics widely used in medical analytics and artificial intelligence research was used to evaluate the quality of classification and analyze the effectiveness of ensemble machine learning methods.

Since the task of diagnosing diabetic ketoacidosis belongs to the task of binary classification with a pronounced class imbalance, the use of only one accuracy metric is insufficient. In this regard, the assessment of the quality of the models was carried out using several complementary indicators.

The Accuracy metric characterizes the total proportion of correctly classified observations relative to the total size of the test sample (Farhadpour et al., 2024) (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where:

- TP is the number of truly positive results;
- TN — the number of truly negative results;
- FP — the number of false positive results;
- FN — the number of false negative results.

Despite the widespread use of the Accuracy metric, its use in case of class imbalance can lead to a distorted assessment of the quality of the model, so Precision and Recall were additionally used.

The Precision metric characterizes the proportion of correctly predicted positive cases among all positive predictions of the model (Owusu-Adjei et al., 2023) (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

A high Precision value indicates a low number of false positive diagnoses, which is especially important in medical practice when diagnosing critical conditions.

The Recall metric reflects the model's ability to detect real cases of the disease (Hicks et al., 2022) (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

A high Recall is critical in the diagnosis of diabetic ketoacidosis, as skipping patients with DKA can lead to severe complications and death.

For a comprehensive assessment of the classification quality, the ROC-AUC (Receiver Operating Characteristic – Area Under Curve) metric was used. The ROC curve reflects the dependence:

- True Positive Rate;
- False Positive Rate.

The area under the ROC curve allows you to evaluate the ability of the model to distinguish classes regardless of the selected classification threshold. The ROC-AUC values close to 1.0 indicate the high diagnostic efficiency of the model.

Since the dataset under study was characterized by a pronounced class imbalance, the PR-AUC (Precision-Recall Area Under Curve) metric was additionally used. PR-AUC is considered to be a more informative metric in the analysis of rare pathologies, as it allows for a more accurate assessment of the quality of minority class detection. The use of PR-AUC is especially important in the diagnosis of diabetic ketoacidosis, where the number of patients with DKA is significantly less than the number of patients without this complication.

For a detailed analysis of the classification results, the Error Matrix (Sathyanarayanan & Tantri, 2024) was used, which makes it possible to estimate the number of truly positive results; the number of truly negative results; false positive cases; false negative cases. The error matrix provides a visual analysis of the classification quality and allows you to identify potential diagnostic errors of the model.

Separation of the sample

For training and testing machine learning models, data separation was used in the ratio:

- 80 % is a training sample;
- 20% is a test sample.

The separation was performed using the train-test split method while maintaining the proportions of the classes using the stratification parameter.

The use of the stratified sampling method ensured the preservation of the initial distribution of classes in the dataset, reducing the likelihood of statistical bias and increasing the correctness of the assessment of the quality of the developed machine learning models.

RESULTS AND DISCUSSION

Results of the class distribution analysis

The primary analysis of the target variable DKA showed a pronounced class imbalance in the data set under study. (Figure 3). Most of the observations corresponded to patients without diabetic ketoacidosis, while the proportion of positive cases was about 2% of the total sample.

Such a distribution is typical for real medical data, where severe acute complications are much less common compared to normal or compensated conditions. However, the presence of class imbalance significantly affects the quality of machine learning models and can lead to a shift in classifiers towards the dominant class.

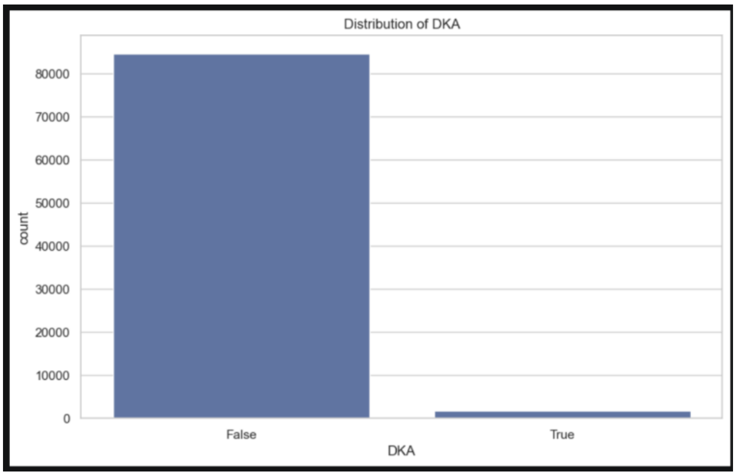

Figure 3. Distribution of DKA classes

Note – compiled by the authors

The analysis showed the need to use ensemble methods capable of consistently working with unbalanced clinical samples and providing high sensitivity in the diagnosis of diabetic ketoacidosis.

Analysis of the distribution of clinical signs

Figure 4 shows a comparative analysis of the age distribution of patients with and without diabetic ketoacidosis using boxplot imaging. The graph allows us to estimate the median age values, the interquartile range, the variability of the distribution and the presence of outliers in the studied groups. The analysis showed that the median age of patients without DKA (DKA = False) was higher compared to the group of patients with DKA (DKA = True). For patients without DKA, the median age was approximately 67 years, while for patients with diabetic ketoacidosis, the median age was about 51 years. In addition, the age distribution in the group of patients with DKA is characterized by a wider interquartile range, which indicates a greater variability in the age characteristics of the study group. Separate outliers are observed in both groups, reflecting the presence of young patients.

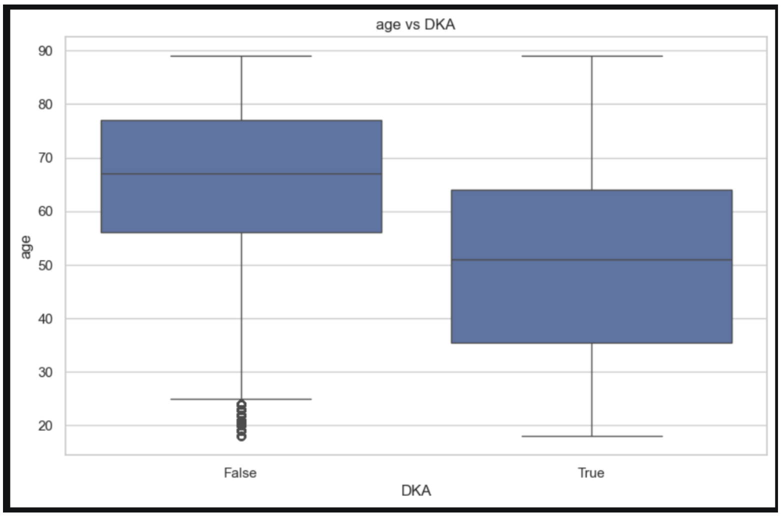

Figure 4. Graph of the distribution of patients' age depending on the presence of DKA

Note – compiled by the authors

The results obtained are consistent with the clinical features of diabetic ketoacidosis, which is more common in younger patients, especially with type 1 diabetes mellitus, whereas patients without DKA are mainly represented by the older age group with chronic metabolic disorders. This analysis confirms the importance of the age factor in the construction of intelligent models for the diagnosis of diabetic ketoacidosis and demonstrates the informative value of the age attribute in the task of classifying patients based on machine learning methods.

Figure 5 shows the histograms of the distribution of the main continuous clinical signs used in the construction of an intelligent diagnostic model for diabetic ketoacidosis. Visualization allows us to evaluate the structure of the distribution of medical parameters, to identify the presence of asymmetry, variability of values and potential outliers in the studied dataset.

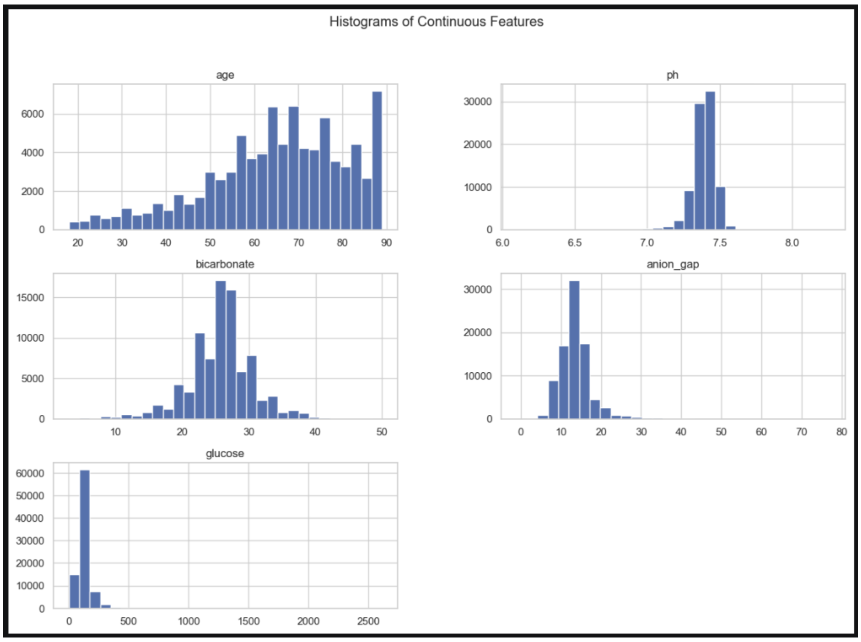

Figure 5. Histograms of the distribution of continuous clinical signs of patients
Note – compiled by the authors

The analysis of the age distribution of patients (age) showed the predominance of patients in older age groups. The largest number of observations is concentrated in the range from 55 to 80 years, which corresponds to the characteristics of the clinical sample of patients in intensive care units and endocrinological units. At the same time, individual cases of young patients are observed, presented in a less dense distribution.

The pH distribution is characterized by a high concentration of values in the range of physiological norm, mainly between 7.3 and 7.5. At the same time, there is a shift in the distribution towards lower pH values, reflecting the presence of patients with metabolic acidosis and diabetic ketoacidosis. This indicator is one of the key biomarkers of acid-base disorder.

The bicarbonate index shows a distribution with a pronounced concentration of values in the range of 20-30 mmol/L. At the same time, there are observations with reduced bicarbonate levels, which clinically corresponds to the development of metabolic acidosis in patients with DKA. Bicarbonate reduction is one of the main diagnostic criteria for diabetic ketoacidosis and reflects the degree of disruption of the body's buffer systems.

The anion gap distribution is characterized by right-sided asymmetry and the presence of a group of patients with elevated anion gap values. An increase in the anion gap is associated with the accumulation of ketone bodies and metabolic acids, which is a characteristic feature of diabetic ketoacidosis. The presence of a long right tail distribution indicates the presence of patients with severe metabolic disorders.

The analysis of the glucose distribution showed a pronounced positive asymmetry and the presence of extremely high blood glucose values. Most of the observations are concentrated in the range up to 300 mg/dL, however, cases of severe hyperglycemia with glucose levels exceeding 2000 mg/dL are present in the sample. Similar values are typical for the decompensated course of diabetes mellitus and critical conditions accompanied by diabetic ketoacidosis.

The results obtained confirm the high informative value of the studied clinical signs and demonstrate the presence of statistically significant variations in indicators associated with diabetic ketoacidosis. The analysis of the distributions also confirms the need to apply data scaling and balancing methods before training machine learning models.

Correlation analysis

Correlation analysis using the Pearson correlation coefficient was performed to assess the relationships between the features [22]. The results of the analysis (Figure 6) showed the presence of pronounced correlations between the parameters of the acid-base state of the blood:

- positive correlation between pH and bicarbonate;
- negative correlation between anion_gap and bicarbonate;
- Negative correlation between glucose and pH.

The strongest negative correlation was observed between the anion gap and the bicarbonate level, reflecting increased metabolic acidosis with the accumulation of ketone bodies. Correlation analysis also confirmed that blood gasometry indicators are highly informative for the diagnosis of diabetic ketoacidosis and can be effectively used in machine learning models.

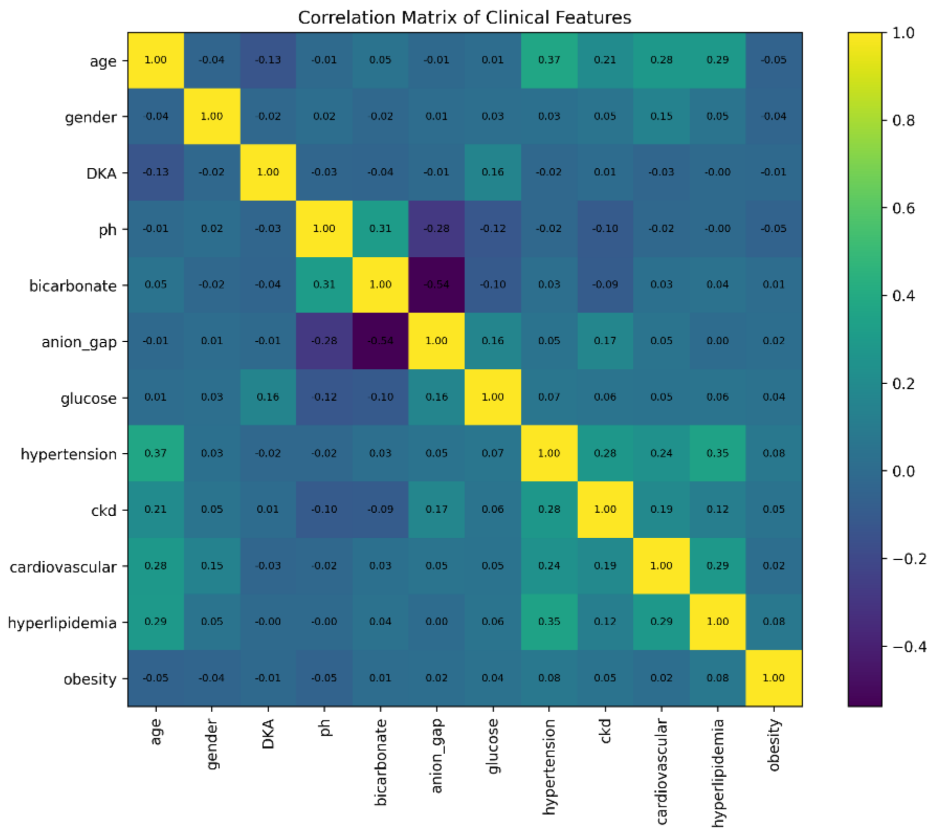

Figure 6. Correlation matrix of clinical signs

Note – compiled by the authors

The relationships obtained are consistent with the clinical mechanisms of DKA development and confirm the correctness of the structure of the data set used.

Classification results

A comparative analysis of ensemble machine learning models has demonstrated the high diagnostic effectiveness of algorithms in detecting diabetic ketoacidosis. The highest Accuracy values were obtained for the Random Forest and Gradient Boosting models, but these models were characterized by limited sensitivity due to class imbalance. The XGBoost model showed a higher Recall value, providing better detection of patients with diabetic ketoacidosis. Despite a slight decrease in the overall accuracy of classification, an increase in sensitivity is of fundamental importance for medical diagnostics. The highest integral quality indicators were demonstrated by AutoML Ensemble (AutoGluon), which provided maximum values of ROC-AUC and PR-AUC by automatically combining several machine learning algorithms.

The classification results are presented in Table 2.

Table 2. Results of classification of machine learning models

Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Random Forest	0.981	0.667	0.050	0.093	0.776	0.186
Gradient Boosting	0.981	0.667	0.050	0.093	0.856	0.186
AdaBoost	0.980	0.000	0.000	0.000	0.824	0.117
XGBoost	0.844	0.085	0.700	0.152	0.855	0.173
AutoML Ensemble (AutoGluon)	0.987	0.741	0.612	0.670	0.921	0.708
Note – compiled by the authors						

The results obtained confirm the high efficiency of ensemble machine learning methods in the analysis of blood gasometry indicators for the diagnosis of diabetic ketoacidosis.

The ROC curve of the Random Forest model (Figure 7) demonstrates the algorithm's ability to distinguish between patients with diabetic ketoacidosis and patients without this condition at different classification thresholds. The value of ROC-AUC = 0.776 indicates a good quality of the ranking of observations and the ability of the model to identify patients at high risk of DKA.

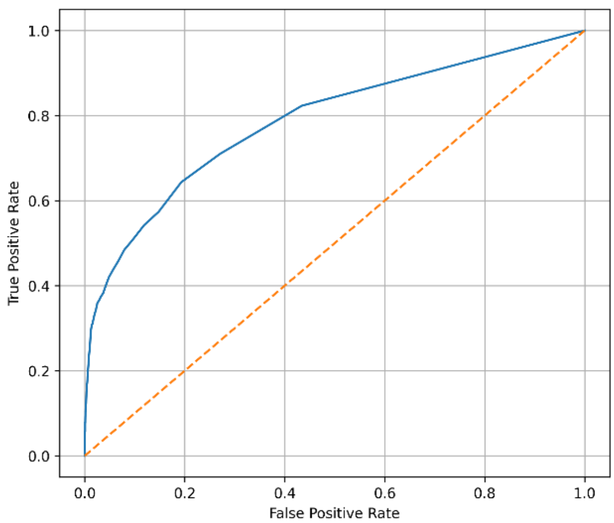

Figure 7. ROC Curve of Random Forest Model

Note – compiled by the authors

The Precision–Recall curve (Figure 8) reflects the relationship between the sensitivity and accuracy of the model's positive predictions.

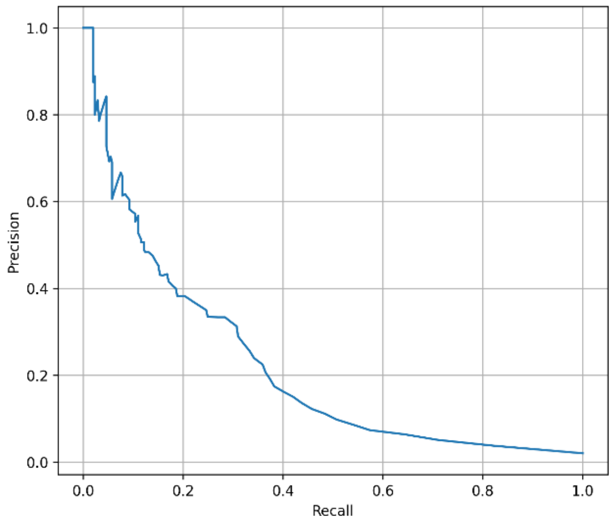

Figure 8. Precision–Recall Curve of Random Forest Model

Note – compiled by the authors

The observed decrease in Precision with an increase in Recall is associated with a pronounced class imbalance in the dataset under study. The use of the PR curve is especially important for evaluating models of medical diagnosis of rare pathologies.

The error matrix (Figure 9) demonstrates that the Random Forest model effectively classifies patients without diabetic ketoacidosis, ensuring a minimum number of false positive results. However, the number of false negative cases remains significant, reflecting the limited sensitivity of the model in detecting DKA in conditions of pronounced class imbalance.

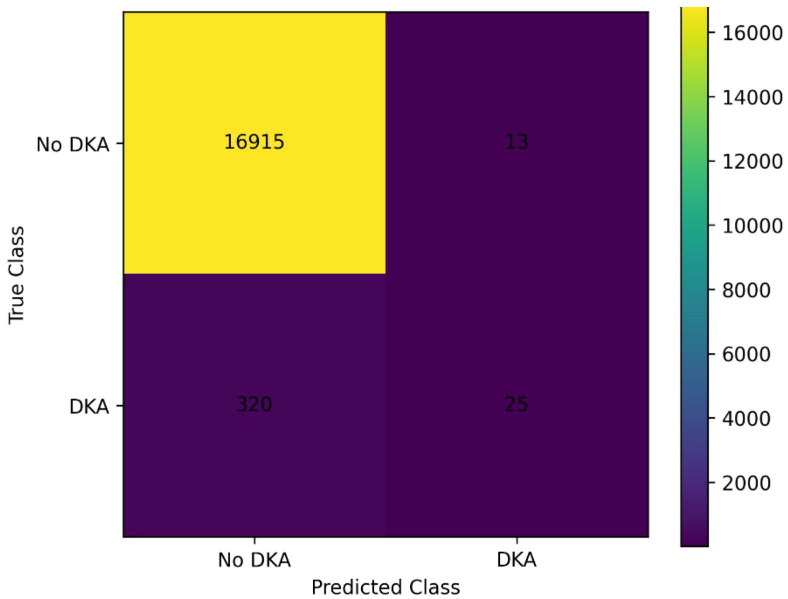

Figure 9. Confusion Matrix of Random Forest Model

Note – compiled by the authors

Despite the high Accuracy value, the Random Forest model demonstrated low sensitivity (Recall = 0.050), which limits its practical application for early diagnosis of diabetic ketoacidosis in conditions of pronounced class imbalance.

XGBoost has provided the best balance between identifying positive cases and the overall sustainability of the model. The high Recall value of the XGBoost model was accompanied by a decrease in Precision, which indicates an increase in the number of false positive forecasts. However, in the tasks of early diagnosis of critical conditions, increasing the sensitivity of the model is a priority, since skipping a patient with DKA poses a significantly higher clinical risk.

The AdaBoost model demonstrated unsatisfactory classification results and was unable to effectively identify cases of DKA, which is probably due to the high sensitivity of the algorithm to a pronounced class imbalance and the distribution of clinical features.

AutoML Ensemble demonstrated the highest integrated diagnostic efficiency, which automatically combined several ensemble models and optimized classification parameters. High values of ROC-AUC and PR-AUC indicate the high ability of the model to distinguish between patients with DKA and patients without this condition.

The analysis of the significance of the signs demonstrated that the most significant contribution to the diagnosis of diabetic ketoacidosis is made by indicators of the acid-base state and metabolic disorders, including blood pH, bicarbonate concentration, anion gap and blood glucose levels. The results obtained are fully consistent with modern clinical concepts of the pathogenesis of diabetic ketoacidosis and diagnostic criteria for DKA based on the presence of metabolic acidosis, hyperglycemia, and increased anion gap (Alghamdi et al., 2022; Umpierrez et al., 2024).

CONCLUSION

Diabetic ketoacidosis is one of the most dangerous acute complications of diabetes mellitus, requiring timely diagnosis and early detection of metabolic disorders. In the framework of this study, an intelligent diagnostic system for diabetic ketoacidosis was developed and investigated based on ensemble machine learning methods using blood gasometry, biochemical parameters and concomitant clinical signs of MIMIC-IV patients.

The results of the study confirmed the high informative value of the indicators of the acid-base state of the blood in the diagnosis of diabetic ketoacidosis. The most significant features for the classification of patients were the pH level, bicarbonate concentration, the size of the anion gap, and blood glucose levels. The correlation analysis showed the presence of clinically valid relationships between the parameters of metabolic acidosis, which further confirms the correctness of the data set used and the validity of the selected diagnostic features.

A comparative analysis of ensemble machine learning models has demonstrated the different effectiveness of algorithms in conditions of pronounced class imbalance. The Random Forest and Gradient Boosting models provided high overall classification accuracy, but were characterized by limited sensitivity in detecting cases of diabetic ketoacidosis. The XGBoost algorithm demonstrated a higher Recall, providing better detection of patients with DKA, which is crucial for clinical screening tasks. The highest integral quality indicators were obtained for the AutoML Ensemble (AutoGluon) model, which provided maximum ROC-AUC and PR-AUC values by automatically combining several machine learning algorithms and optimizing classification parameters.

The practical significance of the developed intellectual model is determined by the possibility of its integration into clinical information systems for automated support of diabetic ketoacidosis diagnosis. The proposed system can be used in intensive care units, endocrinology centers, clinical laboratories, telemedicine platforms and hospital information systems to accelerate the interpretation of blood gasometry results and reduce the influence of subjective factors on diagnostic decision-making. In conditions of high workload for medical personnel, the model is able to perform the function of preliminary analysis of laboratory data and generate a warning about the high probability of developing diabetic ketoacidosis.

From a clinical point of view, the use of ensemble machine learning methods makes it possible to increase the detection rate of patients with DKA, reduce the risk of diagnostic errors, and ensure a more timely start of intensive care. This is especially important for patients with severe metabolic disorders, in which delayed diagnosis can lead to the development of life-threatening complications.

Despite the results, the study has a number of limitations. Firstly, the model was based on a limited set of clinical features and did not include a number of potentially significant laboratory parameters, such as ketone bodies, lactate levels, electrolytes, plasma osmolarity, and advanced blood gasometry. Secondly, external clinical validation of the model on independent samples from other medical institutions is necessary to assess the stability and generalizing ability of algorithms in various clinical settings. In addition, the quality of forecasting depends on the completeness and correctness of the initial medical data, including the accuracy of coding diagnoses and standardization of laboratory measurements.

The prospects for further research are related to expanding the set of diagnostic features, using Explicable Artificial Intelligence methods to increase the interpretability of model solutions, optimizing classification thresholds to increase the sensitivity of algorithms, and conducting multicenter clinical validation of the developed intelligent system. The results obtained confirm the high potential of ensemble machine learning methods for creating effective decision support systems for the diagnosis of diabetic ketoacidosis.

CONFLICT OF INTEREST: The authors declare no conflict of interest.

FUNDING: This article was prepared within the framework of the grant funding project for scientific research of the Ministry of Science and Higher Education under the topic AP26194209 "Mathematical and algorithmic support of an intelligent system for supporting clinical decision-making in hematology and diabetology based on blood gasometry technology".

DATA AVAILABILITY STATEMENT: The clinical data used in this study were obtained from the publicly available MIMIC-IV database provided through PhysioNet. Access to the database is restricted to authorized users who have completed the required training and data use agreement procedures. All authors involved in this study possess valid PhysioNet certification for authorized access to the MIMIC-IV clinical database. The preprocessing procedures, derived datasets, and machine learning implementation supporting the findings of this study are available from the corresponding author upon reasonable request.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES: The manuscript was formulated and translated with the assistance of a LLM for the purpose of improving language clarity, structure, and readability. The LLM was not used for generating experimental results, system design decisions, or scientific contributions. All technical content, system architecture, and analyses were developed and verified by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the work presented in this paper.

REFERENCES

Chang, A. (2023). The role of artificial intelligence in digital health. In Digital health entrepreneurship (pp. 75–85). Springer International Publishing. https://doi.org/10.1007/978-3-031-33902-8_6

Nadhiya, J., Vijayalakshmi, M. K., & Showbharnikhaa, S. (2024). A brief review on diabetes mellitus. Journal of Pharma Insights and Research, 2(1), 117-121. https://jopir.in/index.php/journals/article/view/89

Gregg, E. W., Buckley, J., Ali, M. K., Davies, J., Flood, D., Mehta, R., … Shaw, J. E. (2023). Improving health outcomes of people with diabetes: Target setting for the WHO Global Diabetes Compact. The Lancet, 401(10384), 1302–1312. https://doi.org/10.1016/S0140-6736(23)00001-6

Elendu, C., David, J. A., Udoyen, A.-O., Egbunu, E. O., Ogbuiyi-Chima, I. C., Unakalamba, L. O., … Ramon-Yusuf, H. I. (2023). Comprehensive review of diabetic ketoacidosis: An update. Annals of Medicine & Surgery, 85(6), 2802–2807. https://doi.org/10.1097/MS9.0000000000000894

El-Remessy, A. B. (2022). Diabetic ketoacidosis management: updates and challenges for specific patient population. Endocrines, 3(4), 801-812. https://doi.org/10.3390/endocrines3040066

Barski, L., Golbets, E., Jotkowitz, A., & Schwarzfuchs, D. (2023). Management of diabetic ketoacidosis. European Journal of Internal Medicine, 117, 38-44. https://doi.org/10.1016/j.ejim.2023.07.005

Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A., & Zou, J. (2023). From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell, 186(8), 1772-1791. https://doi.org/10.1016/j.cell.2023.01.035

Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), 18-41. https://doi.org/10.48185/smhs.v1i2.1225

Johnson, A. E. W., Bulgarelli, L., Shen, L., Pollard, T. J., Moody, B., Gow, B., … Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, 1. https://doi.org/10.1038/s41597-022-01899-x

Rawat, S.S., Mishra, A.K. (2024). Review of Methods for Handling Class Imbalance in Classification Problems. In: Agrawal, J., Shukla, R.K., Sharma, S., Shieh, CS. (eds) Data Engineering and Applications. IDEA 2022. Lecture Notes in Electrical Engineering, vol 1146. Springer, Singapore. https://doi.org/10.1007/978-981-97-0037-0_1

Abimannan, S., El-Alfy, E. S. M., Chang, Y. S., Hussain, S., Shukla, S., & Satheesh, D. (2023). Ensemble multifeatured deep learning models and applications: A survey. IEEE Access, 11, 107194-107217. https://doi.org/10.1109/ACCESS.2023.3320042

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024, 69-79. https://doi.org/10.58496/BJML/2024/007

Zhou, Z., Qiu, C. & Zhang, Y. A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. Sci Rep 13, 22420 (2023). https://doi.org/10.1038/s41598-023-49899-0

Kavzoglu, T., Teke, A. Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost). Arab J Sci Eng **47**, 7367–7385 (2022). https://doi.org/10.1007/s13369-022-06560-8

Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4), e12599. https://doi.org/10.1002/eng2.12599

Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification. Applied Sciences, 12(12), 5872. https://doi.org/10.3390/app12125872

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., & Geng, Y. (2022). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research, 278*, 106238. https://doi.org/10.1016/j.atmosres.2022.106238

Farhadpour, S., Warner, T. A., & Maxwell, A. E. (2024). Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. Remote Sensing, 16(3), 533. https://doi.org/10.3390/rs16030533

Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. PLOS Digital Health, 2(11), e0000290. https://doi.org/10.1371/journal.pdig.0000290

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979. https://doi.org/10.1038/s41598-022-09954-8

Sathyanarayanan, S., & Roopashri Tantri, B. (2024). Confusion matrix-based performance evaluation metrics. African Journal of Biomedical Research, 27(4s), 4023–4031. https://doi.org/10.53555/AJBR.v27i4S.4345

Gong, H., Li, Y., Zhang, J., Zhang, B., & Wang, X. (2024). A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information. Engineering Applications of Artificial Intelligence, 131, 107865. https://doi.org/10.1016/j.engappai.2024.107865

Umpierrez, G. E., Davis, G. M., ElSayed, N. A., Fadini, G. P., Galindo, R. J., Hirsch, I. B., Dhatariya, K. K. (2024). Hyperglycemic crises in adults with diabetes: A consensus report. Diabetes Care, 47 (8), 1257–1275. https://doi.org/10.2337/dci24-0032

Alghamdi, N. A., Major, P., Chaudhuri, D., Tsui, J., Brown, B., Self, W. H., … Rochwerg, B. (2022). Saline compared to balanced crystalloid in patients with diabetic ketoacidosis: A systematic review and meta-analysis of randomized controlled trials. Critical Care Explorations, 4(1), e0613. https://doi.org/10.1097/CCE.0000000000000613

Авторлар туралы мәліметтер
Информация об авторах
Information about authors

Увалиева Индира Махмутовна – PhD, Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан

Увалиева Индира Махмутовна – PhD, Восточно-Казахстанский технический университет им. Д. Серикбаева, г. Усть-Каменогорск, Казахстан

Uvaliyeva Indira Makhmutovna – PhD, East Kazakhstan Technical University, Ust-Kamenogorsk, Republic of Kazakhstan

e-mail: iuvalieva@ektu.kz,

ORCID: https://orcid.org/0000-0002-2117-5390

Рахметуллина Женисгуль Толеухановна – физика-математика ғылымдарының кандидаты, Д. Серікбаев атындығы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан

Рахметуллина Женисгуль Толеухановна – кандитат физико-математических наук, Восточно-Казахстанский технический университет имени Д. Серикбаева, г. Усть-Каменогорск, Казахстан

Rakhmetullina Zhenisgul Toleukhanovna – Candidate of Physical-Mathematical Sciences, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Republic of Kazakhstan

e-mail: zhrakhmetullina@edu.ektu.kz,

ORCID: https://orcid.org/ 0000-0002-0554-7684

Нурумов Жарас Жолымханулы – «Ақпараттық жүйелер» БББ докторанты, Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан
Нурумов Жарас Жолымханулы – докторант ОП «Информационные системы», Восточно-Казахстанский технический университет им. Д. Серикбаева, г. Усть-Каменогорск, Казахстан
Nurumov Zharas Zholymkhanuly – PhD student, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan
e-mail: znurumov@ektu.kz,
ORCID: https://orcid.org/0009-0003-2016-6876

Мукашева Роза Урумкановна – техника ғылымдарының кандидаты, Д.Серікбаев атындығы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан
Мукашева Роза Урумкановна – кандитат технических наук, Восточно-Казахстанский технический университет имени Д. Серикбаева, г.Усть-Каменогорск, Казахстан
Mukasheva Roza Urumkanovna – Candidate of Physical-Mathematical Sciences, D.Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Republic of Kazakhstan
e-mail: mukashevaru@mai.ru,
ORCID: https://orcid.org/ 0000-0002-4722-4576

Аменова Фарида Сейткумаровна – PhD, С. Аманжолов атындағы Шығыс Қазақстан университеті, Өскемен, Қазақстан
Аменова Фарида Сейткумаровна – PhD, Восточно-Казахстанский университет им. С. Аманжолова, г. Усть-Каменогорск, Казахстан
Amenova Farida Seytkumarovna – PhD, East Kazakhstan University, Ust-Kamenogorsk, Republic of Kazakhstan
e-mail: faramen.74@mail.ru,
ORCID: https://orcid.org/0000-0001-6656-5037