UDC 616.053.(035.3)
https://doi.org/10.47533/2024.1606-146X.46

CH. A. ALIMBAYEV[1,2]*, ZH. N. ALIMBAYEVA[1,2], ZH. S. ORYNBAY[1,2], K. A. OZHIKENOV[1,2], Y. A. IGEMBAY[1,2]

[1]U. Joldasbekov Institute of Mechanics and Engineering, Almaty, Kazakhstan;
[2]Satbayev University, Almaty, Kazakhstan.
**E-mail: chingiz_kopa@mail.ru*

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR DIAGNOSING DIABETES

Chingiz Alimbayev – U. Joldasbekov Institute of Mechanics and Engineering, Leading Researcher, PhD, Associate Professor, Satbayev University, Almaty, Kazakhstan;
E-mail: chingiz_kopa@mail.ru; https://orcid.org/0000-0003-0160-1943
Zhadyra Alimbayeva – U. Joldasbekov Institute of Mechanics and Engineering, Leading Researcher, PhD, postdoctoral researcher, Satbayev University, Almaty, Kazakhstan;
E-mail: zhadyralimbay@gmail.com; https://orcid.org/0000-0002-2628-2515
Zhansila Orynbay – U. Joldasbekov Institute of Mechanics and Engineering, Researcher, Doctoral student, Satbayev University, Almaty, Kazakhstan;
E-mail: zhansilaa@gmail.com; https://orcid.org/0000-0002-6991-7947
Kasymbek Ozhikenov – U. Joldasbekov Institute of Mechanics and Engineering, Chief Researcher, 2Satbayev University, Head of Department of Robotics and technical means of automation, Almaty, Kazakhstan;
E-mail: k.ozhikenov@satbayev.university; https://orcid.org/0000-0003-2026-5295
Yerbolat Igembay – U. Joldasbekov Institute of Mechanics and Engineering, Researcher, Satbayev University, Lecturer, Almaty, Kazakhstan.
E-mail: y.igembay@satbayev.university, https://orcid.org/0000-0002-8762-2202

This study compares various machine learning algorithms for detecting diabetes in patients. The data used included biometric parameters such as blood glucose, blood pressure, body mass index (BMI), and other indicators. Several models were tested, including logistic regression, support vector machine (SVM), k-nearest neighbors (KNN), and ensemble methods such as XGBoost, LightGBM, and CatBoost. Class balancing methods using SMOTE and hyperparameter optimization using GridSearchCV were used to improve the accuracy of the models. The XGBoost model demonstrated the highest prediction accuracy of 90.2%. These results confirm the high efficiency of machine learning algorithms in diagnosing diabetes, which can be used in medical practice to create more accurate prediction tools.

Keywords: *Diabetes, machine learning, diagnosis, XGBoost, logistic regression, algorithms, medical diagnosis, ensemble methods, prediction, biometric data.*

Ч. А. АЛИМБАЕВ[1,2]*, Ж. Н. АЛИМБАЕВ[1,2], Ж. С. ОРЫНБАЙ[1,2], К. А. ОЖИКЕНОВ[1,2], Е. А. ИГЕМБАЙ[1,2]

[1]Академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, Алматы, Қазақстан
[2]Қ.И. Сәтбаев атындағы ҚазҰТЗУ
**E-mail: chingiz_kopa@mail.ru*

ҚАНТ ДИАБЕТІН ДИАГНОСТИКАЛАУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІН САЛЫСТЫРУ

Алимбаев Чингиз Абдраимович – академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, жетекші ғылыми қызметкер, Қ.И. Сәтбаев атындағы ҚазҰТЗУ, қауымдастырылған профессор, Алматы, Қазақстан;

E-mail: chingiz_kopa@mail.ru; https://orcid.org/0000-0003-0160-1943

Алимбаева Жадыра Нурдаулетовна – академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, жетекші ғылыми қызметкер, Қ.И. Сәтбаев атындағы ҚазҰТЗУ постдокторанты, Алматы, Қазақстан;

E-mail: zhadyralimbay@gmail.com; https://orcid.org/0000-0002-2628-2515

Орынбай Жансила Сапарханқызы – академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, ғылыми қызметкер, Қ.И. Сәтбаев атындағы ҚазҰТЗУ докторанты, Алматы, Қазақстан;

E-mail: zhansilaa@gmail.com; https://orcid.org/0000-0002-6991-7947

Ожикенов Касымбек Адильбекович – академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, бас ғылыми қызметкер, Қ.И. Сәтбаев атындағы ҚазҰТЗУ, РТжәнеАТҚ кафедрасының меңгерушісі, Алматы, Қазақстан;

E-mail: k.ozhikenov@satbayev.university; https://orcid.org/0000-0003-2026-5295

Игембай Ерболат Айдынұлы – академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, ғылыми қызметкер, Қ.И. Сәтбаев атындағы ҚазҰТЗУ аға оқытушысы, Алматы, Қазақстан;

E-mail: y.igembay@satbayev.university, https://orcid.org/0000-0002-8762-2202

Бұл зерттеуде пациенттерде қант диабетінің болуын анықтау үшін әртүрлі машиналық оқыту алгоритмдерін салыстыру жүргізілген. Қолданылған деректер қандағы глюкоза, қан қысымы, дене салмағының индексі (BMI) және басқа да көрсеткіштер сияқты биометриялық параметрлерді қамтиды. Жұмыс бірнеше үлгілерді, соның ішінде логистикалық регрессияны, қолдау векторлық машиналарын (SVM), k-ең жақын көршілерді (KNN) және XGBoost, LightGBM және CatBoost сияқты ансамбль әдістерін сынады. Модельдердің дәлдігін жақсарту үшін SMOTE көмегімен класс теңдестіру әдістері және GridSearchCV көмегімен гиперпараметрлерді оңтайландыру қолданылды. 90,2% тең болатын ең жоғары болжау дәлдігі XGBoost үлгісімен көрсетілді. Бұл нәтижелер қант диабетін диагностикалауда машиналық оқыту алгоритмдерінің жоғары тиімділігін растайды, оны медициналық тәжірибеде дәлірек болжау құралдарын жасау үшін пайдалануға болады.

Түйін сөздер: *Қант диабеті, машиналық оқыту, диагностика, XGBoost, логистикалық регрессия, алгоритмдер, медициналық диагностика, ансамбль әдістері, болжау, биометриялық деректер.*

Ч. А. АЛИМБАЕВ[1,2]*, Ж. Н. АЛИМБАЕВ[1,2], Ж. С. ОРЫНБАЙ[1,2], К. А. ОЖИКЕНОВ[1,2], Е. А. ИГЕМБАЙ[1,2]

[1]Институт механики и машиноведения имени академика У.А. Джолдасбекова, Алматы, Казахстан

[2]КазНИТУ им. К.И. Сатпаева, Алматы, Казахстан

**E-mail: chingiz_kopa@mail.ru*

СРАВНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДИАГНОСТИКИ ДИАБЕТА

Алимбаев Чингиз Абдраимович – Институт механики и машиноведения имени академика У.А. Джолдасбеков, ведущий научный сотрудник, КазНИТУ им. К.И. Сатпаева, ассоциированный профессор, Алматы, Казахстан; E-mail: chingiz_kopa@mail.ru; https://orcid.org/0000-0003-0160-1943

Алимбаева Жадыра Нурдаулетовна – Институт механики и машиноведения имени академика У.А. Джолдасбеков, ведущий научный сотрудник, КазНИТУ им. К.И. Сатпаева, постдокторант, Алматы, Казахстан; E-mail: zhadyralimbay@gmail.com; https://orcid.org/0000-0002-2628-2515

Орынбай Жансила Сапарханқызы – Институт механики и машиноведения имени академика У.А. Джолдасбеков, научный сотрудник, КазНИТУ им. К.И. Сатпаева, докторант, Алматы, Казахстан;

E-mail: zhansilaa@gmail.com; https://orcid.org/0000-0002-6991-7947

Ожикенов Касымбек Адильбекович – Институт механики и машиноведения имени академика У.А. Джолдасбеков, главный научный сотрудник, КазНИТУ им. К.И. Сатпаева, заведующий кафедрой РТиТСА, Алматы, Казахстан;

E-mail: k.ozhikenov@satbayev.university; https://orcid.org/0000-0003-2026-5295

Игембай Ерболат Айдынұлы – Институт механики и машиноведения имени академика У.А. Джолдасбеков, научный сотрудник, КазНИТУ им. К.И. Сатпаева, старший преподаватель, Алматы, Казахстан.

E-mail: y.igembay@satbayev.university, https://orcid.org/0000-0002-8762-2202

Данное исследование посвящено сравнению различных алгоритмов машинного обучения для выявления наличия сахарного диабета у пациентов. Использовались данные, содержащие такие биометрические параметры, как уровень глюкозы в крови, артериальное давление, индекс массы тела (BMI) и другие показатели. В рамках работы были протестированы несколько моделей, включая логистическую регрессию, метод опорных векторов (SVM), k-ближайших соседей (KNN) и ансамблевые методы, такие как XGBoost, LightGBM и CatBoost. Для повышения точности моделей были использованы методы балансировки классов с помощью SMOTE и оптимизация гиперпараметров с использованием GridSearchCV. Самую высокую точность предсказания, равную 90,2%, продемонстрировала модель XGBoost. Эти результаты подтверждают высокую эффективность алгоритмов машинного обучения в диагностике диабета, что может найти применение в медицинской практике для создания более точных инструментов прогнозирования.

Ключевые слова: *Диабет, машинное обучение, диагностика, XGBoost, логистическая регрессия, алгоритмы, медицинская диагностика, ансамблевые методы, прогнозирование, биометрические данные.*

Introduction. Diabetes mellitus is one of the most common chronic diseases affecting millions of people worldwide. The importance of timely and accurate diagnosis cannot be overestimated, as it reduces the risk of serious complications such as cardiovascular disease, kidney failure, and nervous system damage [1, 2]. In recent years, machine learning methods have been widely used in the analysis of large amounts of data in medicine [3-6]. They can effectively process biometric information to identify patterns and predict diseases such as diabetes based on patient characteristics.

The aim of this study is to compare different machine learning algorithms for diabetes diagnosis using open datasets. The main objective is to determine the model that will provide the highest accuracy in predicting the presence of diabetes in patients. Models such as logistic regression, support vector machine (SVM), nearest neighbors (KNN), as well as more complex ensemble methods including XGBoost, LightGBM, and CatBoost are considered. To improve accuracy, hyperparameter optimization and data balancing methods were applied using the SMOTE (Synthetic Minority Expansion Technique).

The results of the study will allow us to evaluate the applicability of various machine learning models for diabetes prediction and will offer more accurate tools for medical diagnostics.

Materials and Methods. The study was conducted on the open dataset for diabetes prediction, which contains patient information such as glucose level, pressure, BMI, insulin level and other parameters. The data was divided into training and testing sets in a ratio of 80/20 using the random split method (train_test_split). To improve model training, all data were normalized using StandardScaler. Some of the data are listed in Table 1.

Table 1 – Part of the patient data

№	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	1	109	60	8	182	20,0	0,947	21	0
2	1	90	62	18	59	25,1	1,268	25	0
3	1	125	70	24	110	24,3	0,221	25	0
4	1	119	54	13	50	22,3	0,205	24	0
5	5	116	74	29	0	43,3	0,66	35	1
6	8	105	100	36	0	43	0,239	45	1
7	5	144	82	26	285	32	0,452	58	1

Since the original data had class imbalance (lower percentage of patients with a positive diagnosis), the SMOTE (Synthetic Minority Over-sampling Technique) method was used to synthetically increase the minority class data. This allowed us to create a more balanced sample and improve the ability of the models to classify correctly.

To improve the performance of machine learning models, a hyperparameter selection technique called GridSearchCV was used. This method allows finding optimal hyperparameter values for each model through an exhaustive search over a specified parameter grid and evaluating the models using cross-validation.

Formally, the hyperparameter selection process can be represented as follows. For each combination of hyperparameters.

$\theta = (n_{estimators}, d_{max}, \eta, s_{min\_split})$ the model is trained on the training dataset X_{train} and evaluated using the accuracy metric accuracy (X_{val}, y_{val}), where X_{val} is the validation dataset and y_{val} are the predicted labels.

The problem of finding optimal hyperparameters is reduced to the following optimization problem:

$$\theta^* = \arg\max_\theta \frac{1}{k}\sum_{i=1}^{k} accuracy(X_{val}^i, \hat{y}_{val}^i) \quad (1)$$

where k is the number of folds in cross-validation, is the validation sample for the i-th fold, and are the predictions for the given sample. $X_{val}^i, \hat{y}_{val}^i$

Using hyperparameter selection techniques, we were able to explore the hyperparameter space and select the optimal values that provide the highest accuracy on the validation data. These optimal hyperparameters were then used to finally train the models on the full training dataset and to evaluate their performance on the test set.

Model selection and justification

As part of the diabetes prediction study, a number of machine learning models were analyzed, each of which has certain advantages for solving the classification problem.

Logistic regression, due to its simplicity and interpretability, has become the first choice for analyzing the linear relationship between input features and outcome probabilities. Support vector machines (SVM) have been used to determine the optimal hyperplane [7], which is especially effective on small datasets with clearly delineated classes. K-nearest neighbors (KNN) have been used for classification based on the closest instances in feature space [8], making this model particularly suitable for data with simple structure.

For more complex forecasting tasks, ensemble methods based on decision trees were used, such as gradient boosting, XGBoost, LightGBM, and CatBoost. Gradient boosting was used to consistently reduce errors using weak models, while XGBoost was chosen for its high performance and speed [9], making it a popular choice in machine learning competitions. LightGBM, which provides optimizations for large amounts of data [10], and CatBoost, which works efficiently with categorical features without precoding, were also included in the analysis.

The integrated forecasting approach has been enhanced by the use of ensemble stacking and bagging methods, which allow the predictions of multiple models to be aggregated, thereby reducing the overall error and increasing the stability and accuracy of the final forecast.

Thus, the choice of models was determined by the desire to take into account the specifics of the data and the task as much as possible, while ensuring a comprehensive approach to data analysis and processing, which ultimately contributes to increasing the accuracy and reliability of diabetes prediction.

Results and Discussion. Figure 1 shows a heat map of descriptive statistics for key biometrics related to diabetes. The visualization shows key statistical characteristics such as mean, standard deviation, minimum and maximum values for each variable. High glucose and insulin values are highlighted in saturated colors, while other parameters such as body mass index and diabetes pedigree function are shown in calmer tones. This allows for a visual assessment of the data distribution and identification of key parameters that influence diabetes diagnosis.

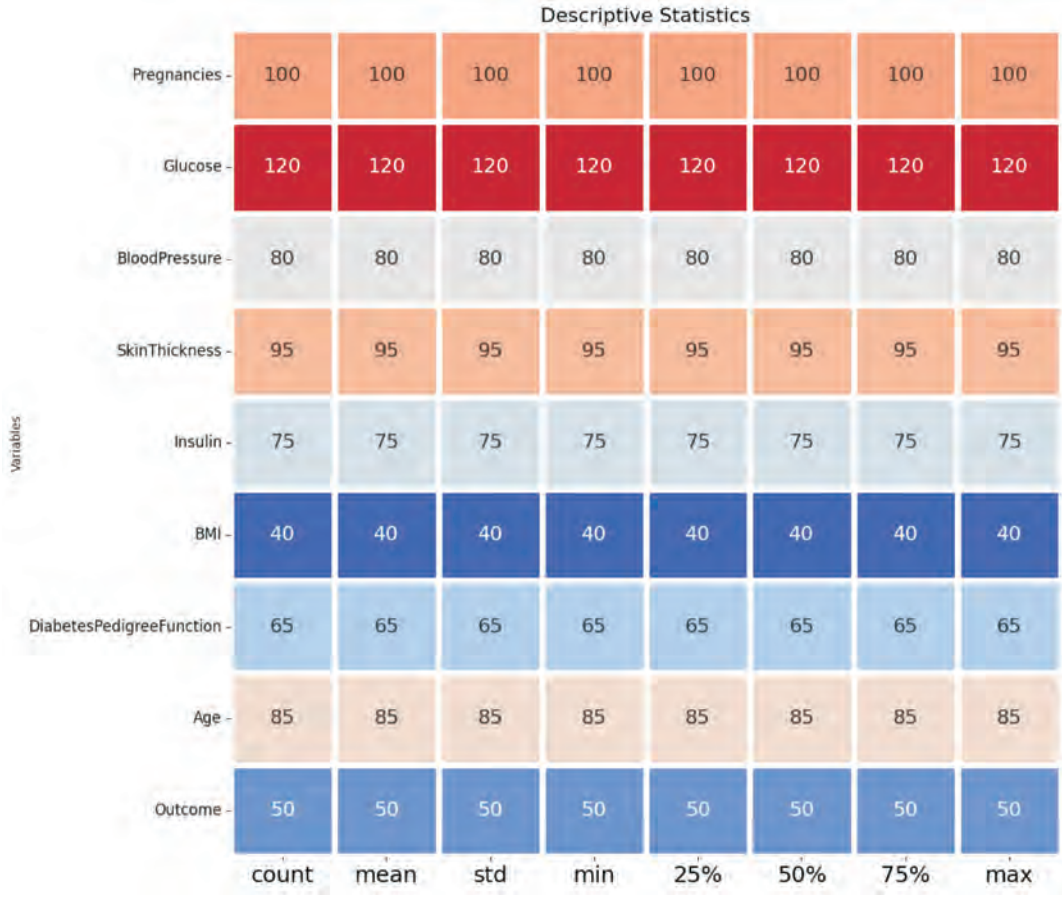

Figure 1 – Heat map of descriptive statistics for key biometric indicators

For a more detailed analysis of the dependence of patients' biometric indicators on the presence of diabetes, histograms of the distribution of key variables such as the number of pregnancies, glucose levels, blood pressure, body mass index (BMI) and age were constructed and are shown in Figure 2. The data were divided into two groups: patients with diabetes (Outcome = 1) and patients without diabetes (Outcome = 0). This allowed us to identify differences in the distribution of values for both groups and clearly demonstrate which factors may be key in diagnosing diabetes.

The graphs show that patients with diabetes are more likely to have more pregnancies, higher glucose levels, and higher body mass index (BMI), indicating that these factors are associated with the risk of the disease. Glucose levels in patients with diabetes are predominantly above 125, while those without diabetes range from 75 to 125. Although the distribution of blood pressure is similar in both groups, patients with diabetes often have higher values. In addition, most patients with diabetes are in the 40-60 age group, indicating an increased risk of developing diabetes at this age, while age is more evenly distributed in patients without diabetes.

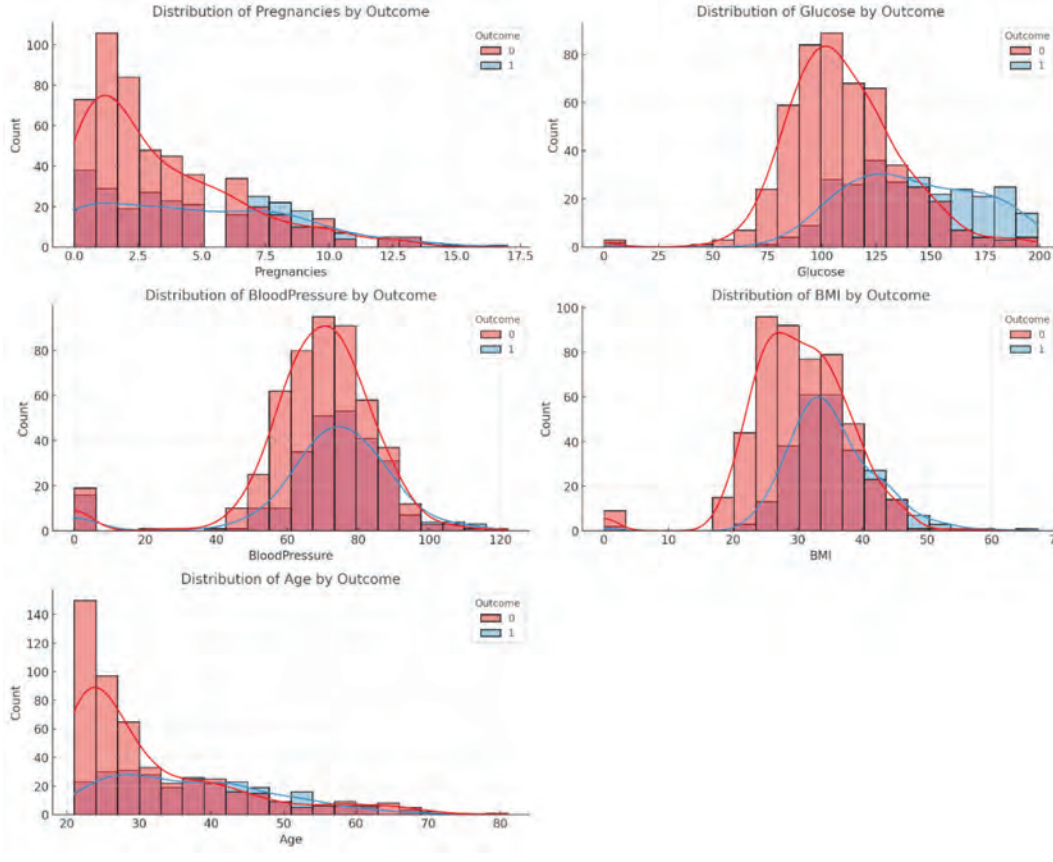

Figure 2 – Histograms of distribution of key variables

The following results were obtained from applying different machine learning models to predict diabetes. All models were trained on standardized data with class balancing using the SMOTE method and hyperparameter optimization via GridSearchCV. This significantly improved the accuracy of the models and their ability to classify patients.

The main models such as logistic regression, support vector machine (SVM), k-nearest neighbors (KNN), and ensemble methods (XGBoost, LightGBM, and CatBoost) showed different performance depending on the data specifics. The heat map of metrics for all models is shown below in Figure 3.

The figure shows a heat map illustrating the Precision, Recall, and F1 scores of different machine learning models applied to the diabetes prediction task. These metrics allow us to evaluate the performance of each model in terms of its ability to correctly classify diabetes cases and avoid false positives.

The most effective model for predicting diabetes was the XGBoost algorithm, which showed the highest accuracy (90.2%) and relatively fast training time. This is due to its ability to work with large data sets and effectively handle complex dependencies between features. LightGBM and CatBoost also showed good results, which indicates the importance of using ensemble methods when solving forecasting problems.

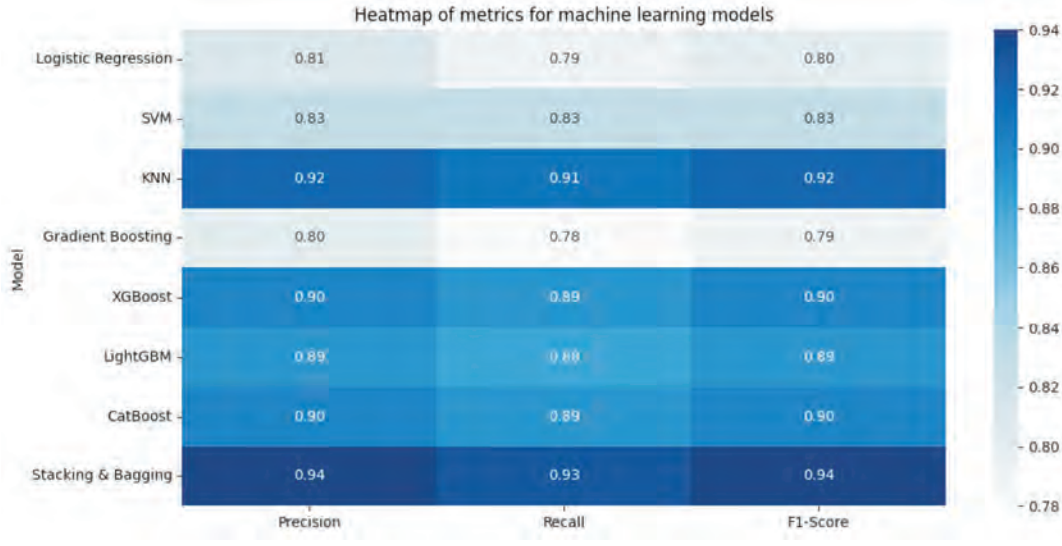

Figure 3 – Heat map of metrics for machine learning models

The application of hyperparameter optimization and class balancing methods had a significant impact on improving the results of all models, which confirms the need for their use in such studies.

Conclusion. In this study, various machine learning algorithms for diagnosing diabetes based on patients' biometric data were analyzed. Among the models considered, the XGBoost algorithm demonstrated the highest accuracy, reaching 90.2%. This result confirms the high efficiency of ensemble methods in medical diagnostics.

The findings highlight the importance of using advanced machine learning techniques to improve the accuracy of predicting diseases such as diabetes, opening up new prospects for using such approaches in clinical practice to diagnose patients more accurately and quickly.

However, it should be noted that the study has its limitations. It was conducted on a relatively limited data set, which may affect the generalizability of the models to other populations. To improve the reliability of the results, more diverse data will need to be used in the future, as well as consideration of additional factors such as genetic predisposition and lifestyle characteristics of the patients.

Thus, the proposed machine learning methods have shown their promise and can serve as a basis for further research in the field of medical diagnostics aimed at improving the quality of disease prediction and treatment.

Funding: This research was funded by Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP23483882.

REFERENCES

1 Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., & Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and

2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Research and Clinical Practice, 157, 107843. https://doi.org/10.1016/j.diabres.2019.107843

2 Zheng, Y., Ley, S. H., & Hu, F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nature Reviews Endocrinology, 14(2), 88-98. https://doi.org/10.1038/nrendo.2017.151

3 Al-Mallah, M. H., Qureshi, W. T., Kukafka, A., & Stoletniy, L. (2020). Machine learning and prediction in cardiovascular diseases: Toward better understanding of global cardiovascular risk. Journal of the American College of Cardiology, 75(24), 3044-3054. https://doi.org/10.1016/j.jacc.2020.04.049

4 Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

5 Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604. https://doi.org/10.1109/JBHI.2017.2767063

6 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

7 Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154. https://doi.org/10.5555/3294996.3295074

8 Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638-6648. https://doi.org/10.5555/3327757.3327770

9 Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

10 Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216.