Proceedings of CONF-APMM 2025 Symposium: Simulation and Theory of Differential-Integral Equation in Applied Physics
DOI: 10.54254/2753-8818/2025.DL27358

Application and Challenges of Machine Learning in Prediction of Type 2 Diabetes: A Systematic Review

Daoji Bai[1], Jiajing Chen[1*]

[1]*School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai, China*

**Corresponding Author. Email: 221490423@stu.lixin.edu.cn*

Abstract. Type 2 diabetes mellitus (T2DM) represents a major global health challenge, necessitating robust strategies for early detection and intervention. Machine learning has emerged as a powerful approach for enhancing clinical prediction of T2DM, employing a spectrum of algorithms from traditional logistic regression and support vector machines to advanced ensemble methods and deep learning architectures. This systematic review explores the integration of diverse multimodal data—including clinical measures (e.g., blood glucose, BMI, blood pressure), electronic health records (EHRs), genomic and proteomic profiles, and lifestyle indicators—to improve predictive accuracy. Despite promising results, critical challenges remain, such as data quality issues (missing values, class imbalance, and privacy concerns), model interpretability, and limited generalizability across populations. Future research should prioritize the development of interpretable, fair, and clinically adaptable machine learning systems. Leveraging time-series data and novel AI techniques such as federated learning could further refine risk stratification and support the translation of predictive models into real-world clinical settings, ultimately contributing to personalized prevention and improved patient outcomes.

Keywords: Type 2 diabetes, Machine learning, Clinical prediction

1. Introduction

Type 2 Diabetes Mellitus (T2DM), a metabolic disorder characterized by impaired insulin secretion or insulin resistance, has emerged as a major global health challenge in the 21st century. According to the International Diabetes Federation (IDF), diabetes prevalence continues to rise worldwide, with over 90% of cases now diagnosed as T2DM, imposing significant social and economic burdens. This condition can lead to severe complications including cardiovascular diseases, kidney disorders, retinopathy, and neurological disorders, severely compromising patients' quality of life and contributing substantially to disability and mortality rates.

However, type 2 diabetes develops gradually, with a critical "prediabetes" phase. Numerous studies demonstrate that lifestyle modifications or medication during this period can significantly delay or even prevent the onset of T2DM. Therefore, achieving early accurate prediction of T2DM, identifying high-risk populations, and implementing targeted prevention and treatment measures

DOI: 10.54254/2753-8818/2025.DL27358

hold vital practical significance for reducing disease incidence and improving public health outcomes.

For decades, clinicians and researchers have primarily relied on statistical methods like Logistic regression and Cox proportional hazards models to assess patient risks. While these approaches demonstrate strong interpretability and operational simplicity, their predictive capabilities generally face inherent limitations. Current methodologies predominantly depend on clinical indicators, which struggle to characterize and quantify complex nonlinear interactions among risk factors. This fundamental limitation makes them ill-equipped to handle modern healthcare big data characterized by massive scale, high dimensionality, and multimodal nature.

In recent years, machine learning-driven artificial intelligence has demonstrated tremendous potential in healthcare. As a groundbreaking approach, machine learning automatically identifies complex patterns and uncovering hidden nonlinear correlations and latent features from massive multi-source data. This project integrates health checkup data, electronic health record (EHR) data, and multi-omics data including genomics and proteomics. Machine learning models such as random forests and XGBoost have already achieved remarkable results across multiple fields.

However, there remain significant challenges in making machine learning an effective clinical diagnostic tool. Key issues include: data quality and privacy concerns (missing data, imbalanced datasets, heterogeneity), model interpretability (black-box decisions failing to gain clinicians' trust), generalization limitations (models performing well on one data set may not apply to another population), and practical implementation feasibility.

Current research primarily focuses on machine learning-based prediction methods for type 2 diabetes, yet there remains a lack of systematic organization, comparison, and synthesis. Existing studies present fragmented findings with inconsistent results, and critical issues in this field remain inadequately addressed. Therefore, there is an urgent need to conduct comprehensive analysis of existing research achievements, clarify their developmental patterns, evaluate the strengths and weaknesses of various methodologies, and provide clear guidance for future research directions.

2. Research status

Diabetes mellitus (DM) is a chronic metabolic disorder primarily characterized by pancreatic β-cell dysfunction. In 2022, global estimates indicated approximately 828 million people with diabetes, representing a 210 million increase from 1990. Notably, 445 million adults aged 30 and above remained untreated—a threefold increase compared to 1990. Prediabetes, a metabolic disorder, involves blood glucose levels within the normal range but not meeting diagnostic criteria. Projections suggest over 470 million new prediabetes cases worldwide by 2030. Research demonstrates that high blood sugar begins harming health even before diabetes diagnosis [1]. As an irreversible condition, early intervention can restore normal glucose levels and delay progression. Therefore, early screening and detection are crucial for effective management.

Common chronic diseases include diabetes, hypertension, cardiovascular disorders, and chronic respiratory conditions. These illnesses often impose significant psychological burdens on patients' physical and mental well-being, making early diagnosis of chronic diseases clinically crucial. In traditional medical practice, diagnostic methods typically involve molecular biology and imaging techniques to monitor various physiological indicators. The primary symptom of diabetes is persistent hyperglycemia. Without timely treatment, this condition can lead to chronic organ damage and various complications. Given the complex etiology of diabetes, diagnostic results may vary widely depending on individual cases.

DOI: 10.54254/2753-8818/2025.DL27358

Current clinical testing methods such as blood glucose measurement, glycated hemoglobin (HbA1c) testing, and oral glucose tolerance tests have been widely adopted, yet they exhibit limitations. Fasting blood glucose measurements are prone to undetected hypoglycemia, while HbA1c tests may yield false positives or negatives, compromising their sensitivity and specificity. Furthermore, factors like dietary habits, stress levels, and timing of testing make single-measurement results unreliable for long-term glycemic status monitoring. Current screening approaches primarily focus on blood glucose levels, with insufficient research on other influencing factors, making it challenging to accurately assess individualized risks. Existing diabetes screening methods demonstrate limitations in specificity, sensitivity, and personalized risk evaluation, making precise early detection of diabetes difficult. Recent advancements in medical technology and machine learning have significantly improved our understanding of diabetes pathogenesis.

Consequently, several simple traditional prediabetes prediction models have emerged. However, these models exhibit significant limitations when handling complex data structures, handling complex data, and performing personalized predictions. This has resulted in traditional diabetes early prediction models failing to meet the demands of sophisticated medical data. Therefore, emerging multi-factor predictive methods such as machine learning hold promise for providing more accurate and effective approaches for diabetes early screening.

In recent years, the rapid development of artificial intelligence (AI) technology in the medical field has demonstrated broad application prospects in areas such as medical image recognition, precise diagnosis of retinal lesions, personalized treatment based on patients 'lifestyle habits, reduction of medical resource waste, and disease prediction. Currently, multiple machine learning models both domestically and internationally have been employed to predict disease probabilities. Building on this foundation, single or integrated learning methods are utilized to provide early warnings for prediabetes conditions, enabling timely detection of pre-diabetic lesions. Explainable AI, a novel research approach emerging in recent years, effectively interprets complex decision-making processes in machine learning and deep learning systems. By integrating these insights with machine learning, it not only enhances understanding of disease occurrence and progression but also improves clinicians' comprehension of pathological mechanisms.

In artificial intelligence research, intelligent analysis of massive medical big data helps uncover correlations between diseases to improve diagnostic accuracy. Current diabetes research primarily focuses on traditional statistical models, machine learning models, deep learning models, and ensemble learning models.

In recent years, significant advancements have been made in disease classification and prediction. Wu developed an integrated model for type 2 diabetes (T2DM) risk assessment by combining preprocessing techniques, an enhanced K-means algorithm, and Logistic regression [2]. Experimental results demonstrated a 3.04% improvement in prediction accuracy compared to control methods while maintaining dataset quality. Samant and colleagues created an iris-based T2DM diagnostic model by integrating iris recognition technology with modern computational methods [3]. Their study of 338 participants analyzed bilateral eye infrared images, conducting experiments with a random forest classifier that achieved 89.63% optimal classification accuracy. Zhou redefined diabetes prediction as a classification task, employing hidden layers in deep neural networks for feature extraction and dropout regularization to prevent overfitting [4]. They established a Deep Learning Predictive Disease (DLPD) model through parameter tuning and a binary cross-entropy loss function. Mishra optimized clinical symptoms using an Adaptive Genetic Algorithm (EAGA) and combined it with Multilayer Perceptrons (MLPs) to develop the EAGA-MLP model [5]. Validation through seven datasets achieved a peak prediction accuracy of 97.76%.

DOI: 10.54254/2753-8818/2025.DL27358

Mallika combined CrowSearch and BGWO algorithms to optimize the SVM model and compared it with existing SVM methods [6]. Experiments showed that the improved CS-BGWO-SVM algorithm outperformed conventional SVM in multiple aspects. Mahesh proposed hybrid ensemble learning (EL) for diabetes prediction and optimized classification [7]. Building on this, they developed a statistical least squares method combining Bayesian networks with radial basis functions, which was compared with five common machine learning methods. Experimental results demonstrated that this approach achieved the best prediction accuracy for diabetes, reaching 97.11%. Gu ndog du proposed integrating multiple linear regression, random forest, and XGBoost [8]. Using MLR-RF for feature selection and XGBoost for image classification, they achieved 99.2% accuracy on the Westheart Diabetes Database. Addressing the increasing complication burden in type 2 diabetes (T2DM) patients, Luo utilized association rules from Apriori algorithm to identify high-risk complications and built a risk prediction model using Logistic regression [9]. Existing studies indicate that diabetic lower limb lesions, diabetic foot, and retinopathy are high-risk factors for type 2 diabetes mellitus. Zhou developed a diabetes prediction framework integrating Boruta feature selection with K-means++ clustering and stacked ensemble learning, validated using the Indian Diabetes Patient Database [10]. Experimental results demonstrated over 98% prediction accuracy, surpassing all existing models. Pan applied deep learning to diabetic retinopathy (DR) binary classification, focusing on convolutional neural networks while conducting comprehensive comparative analyses of various algorithms [11]. Work showed that ensemble models achieved 85% accuracy in diabetes prediction [12]. Building on this foundation, researchers employed information gain to optimize feature subsets, enhancing model performance. Shao addressed low prediction accuracy in imbalanced datasets by combining sparse over-sampling (SMOTE) with random under-sampling (RUS), optimizing hyperparameters of LightGBM using Optuna [13]. Modak integrated multiple machine learning methods including Logistic regression and support vector machines, combining LightGBM, XGBoost, CatBoost, AdaBoost, and Bagging to improve prediction accuracy, with the CatBoost model achieving 95.4% precision [14].

3. Diabetes and its types

Diabetes is a chronic condition that disrupts the body's ability to convert food into energy. Normally, the body breaks down nutrients into glucose or sugars before circulating them through the bloodstream. A hormone secreted by pancreatic islets helps deliver glucose to cells for energy production. Type 1 diabetes is an autoimmune disorder where the immune system attacks insulin-producing beta cells in the pancreas. While it can occur at any age, it most commonly affects children, adolescents, and young adults. Early symptoms include vision problems, extreme fatigue, unexplained weight loss, intense thirst, and frequent urination. Type 2 diabetes develops when the body becomes insulin-resistant or the pancreas fails to produce enough insulin to maintain blood sugar levels. Key risk factors include obesity, lack of exercise, and poor dietary habits. This type typically appears in people over 40, with incidence rates rising with obesity severity. Symptoms develop gradually, ranging from persistent fatigue and thirst to blurred vision, slow wound healing, and recurrent infections. Gestational diabetes manifests as elevated blood sugar during pregnancy due to insufficient insulin production, which increases the body's need for insulin. Scientists believe this occurs because hormonal changes during pregnancy reduce insulin sensitivity. The condition usually develops between weeks 24-28 of pregnancy and resolves spontaneously after delivery. While most cases show no obvious symptoms, some expectant mothers may experience intense thirst, frequent urination, and persistent fatigue.

DOI: 10.54254/2753-8818/2025.DL27358

When blood glucose levels are within the normal range but haven't yet reached the threshold for type 2 diabetes, this condition is known as prediabetes. This precursor to diabetes typically emerges before full-on onset, meaning it could progress to diabetes if left untreated. While these early warning signs may persist for years without noticeable symptoms, similar to type 2 diabetes, prediabetes usually shows no apparent symptoms. However, medical tests can detect significantly elevated glucose concentrations in the bloodstream during this phase.

4. Deep learning methods

With the continuous expansion of data scale and advancements in computing power, Deep Learning (DL), as a subfield of machine learning, has demonstrated tremendous potential in type 2 diabetes prediction through its powerful representation learning capabilities. Unlike traditional machine learning models that require tedious manual feature engineering, DL can automatically extract high-level, complex nonlinear features from raw or preprocessed data. This makes it particularly suitable for processing multimodal, high-dimensional data such as electronic health records (EHR), medical imaging, and time-series physiological signals. This section will systematically review the current application status of Deep Learning in this field.

4.1. Mainstream deep learning models and their applications

4.1.1. Deep Neural Network (DNN) and Multilayer Perceptron (MLP)

Deep Neural Networks (DNNs), the most fundamental deep learning models, consist of multiple hidden layers. In type 2 diabetes (T2DM) prediction, DNNs are commonly employed to process structured tabular data such as clinical indicators and physical examination records. Researchers input a series of risk factors—including age, BMI, blood glucose levels, and blood pressure—as features, enabling deep networks to learn complex mapping relationships between characteristic combinations and disease endpoints. Multiple studies have demonstrated that even basic DNNs often outperform traditional linear models like logistic regression [15].

4.1.2. Recurrent Neural Networks (RNN) and their variants (LSTM, GRU)

The RNN family is crucial due to its powerful temporal modeling capabilities. EHR data inherently represents time-varying sequences (such as multiple outpatient records and annual health checkup results). RNNs—particularly Long Short-Term Memory networks (LSTM) and gated recurrent units (GRU)—effectively capture long-term temporal dependencies in clinical indicators, enabling more accurate assessment of disease progression risks.

Application scenario: Predict the risk of developing T2DM at a future time point by using patients' historical multiple medical records. For example, Choi et al. demonstrated the feasibility of using LSTM model to process EHR time series data for diabetes prediction earlier [16].

4.1.3. Convolutional Neural Network (CNN)

Although CNN was originally designed for image processing, its ability to extract local features through convolution kernels has also been successfully applied to tabular data and time series signals.

Application scenarios: One-dimensional Convolutional Neural Networks (1D-CNN): These models can process time-series data from single variables (e.g., Continuous Glucose Monitoring

DOI: 10.54254/2753-8818/2025.DL27358

[CGM] data) or multi-variable clinical records. By automatically learning local fluctuation patterns within datasets, CNNs can identify critical risk indicators that may be crucial for clinical decision-making.

Feature crossover: CNN can also be used to automatically learn the interaction between different features in table data, instead of the traditional method of manually constructing interaction terms.

4.1.4. Autoencoder (AE) and its variants

An autoencoder is an unsupervised learning model designed to learn a low-dimensional dense representation (encoding) of data. It has two main uses in T2DM prediction:

Feature dimensionality reduction and denoising: Processing high-dimensional and sparse EHR data (such as diagnosis codes and drug codes), learning their low-dimensional and meaningful representations, and then using these representations as input to other prediction models (such as DNN) can effectively improve model performance and alleviate overfitting.

Anomaly detection: By learning the distribution of data from healthy people, the autoencoder can identify individuals that deviate from this distribution, thereby identifying potential prediabetics or patients.

4.1.5. Hybrid models

Current research trends tend to combine the advantages of multiple deep learning models to build hybrid architectures to maximize prediction performance.

CNN-LSTM Hybrid Model: The architecture first extracts local features from clinical data in each time window using the CNN layer, then feeds these feature sequences into the LSTM layer to capture long-term temporal dependencies. This framework effectively captures both horizontal (between features) and vertical (over time) complex patterns simultaneously.

Graph Neural Networks (GNNs): This represents an emerging research direction. By treating patients as nodes in a graph and their similarities (e.g., based on genomic data or living environments) or relationships between medical entities (diseases, medications) as edges, we can construct medical knowledge graphs. GNNs enable reasoning within this graph structure, providing richer contextual information for predictive modeling [17].

4.2. Data modalities and feature learning

Deep learning not only improves prediction accuracy, but more importantly, it expands the available data modes:

EHR timing data: The DL model can take full advantage of the rich longitudinal information in the EHR.

Medical imaging: Some studies have begun to explore the use of retinal fundus images to directly predict the risk of T2DM or its complications (such as diabetic retinopathy) through CNN, realizing the prediction of "imaging omics".

Continuous glucose monitoring (CGM) data: The DL model is well suited to analyze the high frequency time series data generated by CGM to detect subtle patterns of blood glucose fluctuations.

Multimodal data fusion: Deep learning architecture (such as multi-input model) provides the technical possibility to integrate multimodal information such as clinical data, imaging data and genomic data, which is the key to achieve truly personalized accurate prediction.

DOI: 10.54254/2753-8818/2025.DL27358

4.3. Advantages and challenges

Superiority :Automatic feature engineering: manual feature screening and construction do not need to rely too much on domain knowledge. Processing complex data: good at processing time series, high dimension and multi-mode data. High performance potential: With sufficient data, it can usually achieve the highest prediction performance available. throw down the gauntlet :Data hunger: Deep learning models often require a large amount of annotated data to train robust models and avoid overfitting. Medical data is expensive and scarce to obtain. This remains the primary clinical translation challenge for deep learning models. Their decision-making processes function like "black boxes" that are difficult for physicians to comprehend and trust. While post-interpretation tools such as SHAP and LIME exist, endogenous interpretability continues to pose significant research challenges. Training large deep learning models requires powerful computing resources (such as GPU) and longer time.

Requirements for data preprocessing: How to deal with the problems of large amounts of missing values, irregular sampling and time alignment in medical data also poses challenges to DL models.

4.4. Summary and prospect

Deep learning has brought a paradigm shift to T2DM prediction, from relying on manual features to automatic representation learning. Although models such as LSTM and CNN have shown excellent performance, future research should focus more on:

Develop lightweight and efficient models to cope with the reality of relatively small medical data volume.

Improve the interpretability of the model, develop an endogenous interpretable DL architecture or combine it with a knowledge graph to make its decision-making process transparent and gain clinical trust.

Explore more effective multimodal data fusion framework. Transfer learning and federated learning are used to train a more generalized model with multi-center data under the premise of protecting data privacy.

5. Predictive model framework based on machine learning

Developing machine learning-based predictive models for type 2 diabetes is a systematic endeavor that extends far beyond algorithmic screening. The process encompasses data collection and preprocessing, feature extraction, model selection and training, evaluation and interpretation of results, culminating in practical applications and monitoring. This chapter will detail this framework and explore its specific medical implementations.

5.1. Data acquisition and preprocessing

Source: Public data sources such as the U.S. CDC National Health and Nutrition Examination Survey, UK Biobank, and MIMIC-III/IV. The collected data has been systematically categorized to facilitate academic research. Electronic Health Records (EHR): These store extensive clinical diagnostic, laboratory test, medication, and disease progression data within hospitals, though they face challenges like complex structures and data noise. Targeted Cohort Studies: High-quality, well-defined research materials with specific objectives, but involve substantial collection costs. Multimodal Data: Includes genomic, proteomic, and continuous physiological data from wearable devices (e.g., heart rate, step counts).

DOI: 10.54254/2753-8818/2025.DL27358

Data Cleaning: Addressing data entry errors and significant outliers. Missing Value Handling: In medical datasets, missing values are extremely common. This project plans to employ three approaches: direct deletion (for minor missing values), statistical imputation (mean/median), and model-based imputation (K nearest neighbors, multiple interpolation methods) [18]. Caution must be exercised in method selection to avoid biased results. Category Imbalance: Type 2 Diabetes (T2DM) patients typically represent a small proportion of the population, resulting in severe imbalance between positive and negative samples in datasets. Common solutions include over-sampling (e.g., SMOTE), under-sampling, and class weighting adjustments in loss functions. Time Series Alignment: For electronic health record (EHR) time series data with irregular sampling patterns, operations such as window segmentation, alignment, and fusion should be performed, followed by conversion into rule-based formats that can be processed by machine learning models.

5.2. Feature engineering

Feature engineering is a crucial step in enhancing model performance, aiming to effectively extract existing data to better reflect the characteristics of the target subject. Feature construction: Generating new features based on clinical knowledge. The triglyceride-glucose index (TyG index) = fasting blood glucose (mg/dL)/2 can serve as a simple and clinically applicable diagnostic indicator. Through statistical analysis of multiple EHR datasets, we obtained historical statistics including maximum values, minimum values, mean values, and trend slopes. Feature transformation: Standardizing continuous features to meet the assumptions of machine learning models like support vector machines and neural networks, thereby accelerating model convergence.

Feature Selection: The process of identifying the most predictive and relevant samples from multiple datasets to mitigate overfitting risks and enhance model interpretability. Common screening methods include: 1) Screening Method: Rapid ranking based on statistical parameters (e.g., chi-square test, correlation coefficients). 2) Bagging Method: Techniques like recursive feature elimination (RFE) that select features through model performance feedback, though computationally intensive. 3) Embedding Methods: Automated feature selection using approaches such as LASSO regression, decision trees, and tree-based feature importance analysis.

5.3. Model selection and training

Model Selection: Select appropriate algorithms based on data characteristics, prediction tasks (classification/regression), and requirements (accuracy/explanability). This project will first establish baselines using classical modeling methods such as logistic regression and SVM, then conduct in-depth research through ensemble learning approaches (e.g., random forests, XGBoost, LightGBM) and deep learning techniques. Currently, there is no "one-size-fits-all" optimal model, requiring experimental comparisons. Hyperparameter Optimization: Model parameters (including number of trees, learning rate, etc.) significantly impact system performance. Systematic optimization methods include:

1. Grid search: Exhaust all combinations in a specified parameter space.
2. Random search (RS): Random sampling of parameter space is carried out to improve the efficiency of search.
3. Bayesian optimality: A more advanced adjustment strategy that intelligently selects the next set of parameters based on past evaluations.

Among these, k-fold cross-validation (k-fold CV) is a crucial technique for evaluating model generalization capabilities and preventing overfitting. This method divides the dataset into k groups, sequentially selecting one group at a time. Each iteration uses the first group as the base sample, then selects k samples from this group to form the next sample set, and finally conducts cluster analysis using the remaining k samples from the initial group.

5.4. Model evaluation and interpretation

Evaluation Criteria: In medical research, selecting appropriate evaluation criteria is crucial. Accuracy alone should not be the sole focus, particularly when dealing with imbalanced data. Common metrics include: area under the curve (AUC), accuracy, recall, and F1 score. Clinical Relevance Index: Calibration curves may also be used to verify whether a model's predicted risk levels match actual outcomes (e.g., whether 20% of a population with 20% risk actually develops the disease). Model Explainability: Black-box models must be interpretable and clinically applicable. Physicians need to understand why models make specific predictions to determine their credibility. Overall Explainability: Analyzing which features most influence the entire model (e.g., ranking them by importance). Partial Explainability: Explaining individual predictions. Common methods include: SHAP (SHAP): A game-theoretic approach that assigns weights to attribute predictions, currently the most popular and robust interpretation method [19]. LIME (Local Interchangeability Model-Diagnostic Model): Using locally interpretable models (e.g., linear models) to approximate complex models through sample collection.

5.5. Deployment and monitoring

For a model to achieve true clinical value, it must first demonstrate effectiveness in real-world applications. By integrating medical big data with web-based and mobile applications, this technology enables patients to conduct self-screening. However, as disease prevalence rates, diagnostic criteria, and treatment methods evolve, the model's performance may gradually deteriorate (conceptual drift). A continuous monitoring protocol should be established to periodically assess model performance using updated data. When system performance shows significant decline, retraining or adjustments with new data are essential to maintain predictive accuracy.

6. Conclusion

This paper primarily focuses on machine learning-based prediction methods for type 2 diabetes. It clarifies their developmental patterns, evaluate the strengths and weaknesses of various methodologies, and provide clear guidance for future research directions. Key issues include: data quality and privacy concerns (missing data, imbalanced datasets, heterogeneity), model interpretability (black-box decisions failing to gain clinicians' trust), generalization limitations (models performing well on one data set may not apply to another population), and practical implementation feasibility.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

DOI: 10.54254/2753-8818/2025.DL27358

References

[1] Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice, 183, 109119.

[2] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100–107.

[3] Samant, P., & Agarwal, R. (2018). Machine learning techniques for medical diagnosis of diabetes using iris images. Computer Methods and Programs in Biomedicine, 157, 121–128.

[4] Zhou, H., Myrzashova, R., & Zheng, R. (2020). Diabetes prediction model based on an enhanced deep neural network. EURASIP Journal on Wireless Communications and Networking, 2020(1), 122.

[5] Mishra, S., Tripathy, H. K., Mallick, P. K., Bhoi, A. K., & Barsocchi, P. (2020). EAGA-MLP—An enhanced and adaptive hybrid classification model for diabetes diagnosis. Sensors, 20(14), 4036.

[6] Mallika, C., & Selvamuthukumaran, S. (2021). A hybrid crow search and grey wolf optimization technique for enhanced medical data classification in diabetes diagnosis system. International Journal of Computational Intelligence Systems, 14(1), 730–742.

[7] Mahesh, T. R., Dhilip Kumar, V., Kumar, V., Junaid Asghar, Mekcha Bazezew Banchigize, Natarajan, R., & Vivek, V. (2022). Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease. Computational Intelligence and Neuroscience, 2022, 4451792.

[8] Gündoğdu, S. (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. Multimedia Tools and Applications, 82(22), 34831–34849.

[9] Luo, X., Sun, J., Pan, H., Wang, D., Wu, J., & Zeng, D. (2023). Establishment and health management application of a prediction model for high-risk complication combination of type 2 diabetes mellitus based on data mining. PLOS ONE, 18(8), e0289749.

[10] Zhou, H., Xin, Y., & Li, S. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics, 24(1), 136.

[11] Pan, L., Sun, W., Wan, W., & Zeng, D. (2023). Research progress of diabetic disease prediction model in deep learning. Journal of Theory and Practice of Engineering Science, 3(12), 15–21.

[12] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, 104672.

[13] Shao, H. Z., Liu, X., Zong, D. S., Liu, Y., & Li, J. (2024). Optimization of diabetes prediction methods based on combinatorial balancing algorithm. Nutrition & Diabetes, 14(1), 63.

[14] Modak, S. K. S., & Jha, V. K. (2024). Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications, 83(13), 38523–38549.

[15] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515.

[16] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361–370.

[17] Shang, J., Ma, T., Xiao, C., & Sun, J. (2019). Pre-training of graph augmented transformers for medication recommendation. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 5953–5959). AAAI Press.

[18] Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ, 338, b2393.

[19] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (NIPS 2017)