International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Construction of a Dynamic Early Warning and Interpretable Decision-Model for Diabetes Risk Based on Multimodal Iterative Forest

Liu Yulin
Anhui University of Finance and Economics, Bengbu, China-233030
Email address: 3530199338@qq.com

***Abstract*—*Diabetes is caused by insulin deficiency or utilization disorders. Undiagnosed patients face high risks of complications such as heart disease, kidney disease, and eye diseases, making early diagnosis crucial. Unlike traditional one-on-one diagnosis, efficient screening can be achieved by analyzing large-scale data using machine learning techniques. Based on the Random Forest algorithm, this study proposes an interpretable iterative random forest model with dynamic weight optimization, innovatively integrating dynamic weight smoothing strategy, stability-driven interactive feature screening, and a lightweight edge deployment scheme. It significantly improves the model's prediction accuracy and generalization ability. This study also focuses on model interpretability, providing intelligent decision support for early diabetes screening.*

***Keywords*— *Diabetes Risk Warning; Dynamic Weight Optimization; Feature Interaction Screening; Iterative Random Forest; Multimodal Data.*

I. INTRODUCTION

Diabetes Mellitus is a chronic metabolic disease caused by insufficient insulin secretion or utilization disorders. Its global prevalence is rising sharply with changes in lifestyle and population aging, making it one of the most severe public health challenges in the 21st century. According to the latest research published in the authoritative medical journal The Lancet in 2024, the global number of diabetes patients has surged from 200 million in 1990 to 828 million in 2022. China, as one of the countries with the heaviest diabetes burden, has seen its prevalence increase by 3.5 times over the past two decades, showing complex epidemic trends such as younger onset age, narrowing urban-rural gap, and high undiagnosed rate . Without effective intervention, diabetic patients will face risks of a series of serious complications like cardiovascular diseases, kidney disease, and retinopathy, placing a heavy burden on individuals, families, and the social healthcare system.

Traditional diabetes diagnosis relies on the clinical experience of specialist physicians and static tests of specific indicators such as Fasting Plasma Glucose (FPG) and Oral Glucose Tolerance Test (OGTT), which have limitations including low efficiency, strong subjectivity, and difficulty in achieving large-scale early screening. In this context, using artificial intelligence (AI) and machine learning (ML) technologies to automatically analyze multi-dimensional health data and construct efficient and accurate disease risk prediction models is of great practical significance for achieving early dynamic warning of diabetes, optimizing the allocation of medical resources, and promoting proactive health management.

In recent years, the application of machine learning technology in the healthcare field has made significant progress. In diabetes prediction research, many classical algorithms have been widely used. Early studies often used linear models such as Logistic Regression and Support Vector Machine (SVM). Subsequently, ensemble learning algorithms, especially Random Forest (RF) and gradient boosting decision trees (e.g., XGBoost), demonstrated superior performance due to their ability to effectively handle nonlinear relationships and high-dimensional features. For example, Chen et al., based on the PIMA Indians Diabetes dataset, verified that Random Forest significantly outperformed SVM and Logistic Regression in prediction accuracy; the research by Yang Meijie et al. also proved its good balance between sensitivity and specificity.

However, existing research still faces several key challenges:

1.Data level: Medical datasets commonly have problems like class imbalance (far fewer patient samples than healthy samples), noise, and missing values, which can easily lead to model bias.

2.Model level: Traditional models have limited ability to capture complex feature interactions and are mostly "black-box" models, whose prediction results lack interpretability, making it difficult to gain the trust of clinicians.

3.Application level: Most models remain in the offline analysis stage, lacking lightweight solutions for dynamic warning and edge deployment, which limits their clinical translation and practical application value.

To address these challenges, advanced methods such as Iterative Random Forest (IRF) have begun to attract attention. By dynamically adjusting sample or feature weights, it continuously optimizes model performance through multiple iterations, and has begun to integrate with Explainable AI (XAI) technologies to provide decision basis. However, existing IRF research mostly focuses on single-dimensional optimization, lacking a unified framework integrating feature screening, model optimization, and interpretability analysis.

Aiming at the above shortcomings, this study aims to construct a dynamic early warning and interpretable decision model for diabetes risk based on a multimodal iterative forest.

This study does not pursue extreme complexity at the algorithm level but emphasizes an innovative and achievable systematic solution. Its core innovative ideas are reflected in the following three aspects:

1."Iterative" optimization strategy: The "iterative" concept is embodied in a strategy combining grid search hyperparameter optimization and Recursive Feature Elimination (RFE), systematically conducting multiple rounds of training and feature screening to dynamically approach the optimal model and improve prediction accuracy and generalization ability.

2.Interpretable decision fusion: Abandoning complex custom rule extraction algorithms, the mature SHAP (SHapley Additive exPlanations) value explanation framework is introduced for post-hoc analysis of the optimized model. This method can provide clear, quantitative feature contribution analysis from both global and local levels, making the model's decision process transparent and trustworthy, and providing intuitive decision support for clinicians.

3.Conceptual design of lightweight dynamic warning: "Lightweight" and "dynamic warning" are discussed as system application goals, proposing a conceptual design for an edge deployment scheme based on model compression technology (such as ONNX format conversion), laying a theoretical foundation for subsequent actual deployment.

This study will follow the research route of "data preprocessing → feature engineering → model iterative optimization → interpretability analysis → system concept design". The ultimate goal is to develop an intelligent screening tool for diabetes risk that is not only accurate but also usable, trustworthy, and promotable, providing effective technical support for improving public health service levels.

II. Preprocessing and Theoretical Basis

Random Forest (Random Forest) is an ensemble learning algorithm proposed by Leo Breiman. It completes classification or regression tasks by constructing multiple decision trees and integrating their results. Its core idea is "the wisdom of the crowd".

The construction process of Random Forest is mainly based on Bagging (Bootstrap Aggregating) idea and random feature selection:

1.Bootstrap Sampling: Randomly select N samples with replacement from the original training set to generate a bootstrap sample set. This process is repeated K times to generate K different sample subsets.

2.Decision Tree Generation: For each bootstrap sample set, a decision tree is constructed in parallel. When splitting each node during tree construction, instead of selecting the optimal feature from all features, first randomly select m features from all M features (m << M, usually 取$m = \sqrt{M}$ or log_2M), and then select the optimal feature from these m features for node splitting. This randomness effectively reduces model variance, enhances generalization ability, and avoids overfitting.

3.Integrated Output: For classification problems, K trees yield K classification results. The final prediction result of the random forest is determined by the majority voting result of these tree outputs; for regression problems, the average value is used as the final output.

The advantages of Random Forest include: good robustness for high-dimensional data and handling missing values, ability to evaluate feature importance, fast training speed, and less prone to overfitting.

Although machine learning models (especially ensemble models) have powerful predictive performance, their "black-box" nature limits their application in high-risk decision-making fields such as healthcare. Explainable AI (XAI) aims to reveal the internal working mechanism of models, enabling humans to understand, trust, and manage their decision-making processes. This study primarily uses the SHAP (SHapley Additive exPlanations) framework for model interpretation.

SHAP originates from the Shapley Value in game theory. Its core idea is to treat the prediction result of the model as the "gain" from the collaboration of all features and fairly distribute the contribution of each feature to the final prediction result. For the prediction of any sample, the SHAP value can be expressed as:

$$\varnothing_{\mathrm{j}}(val) = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[val(S\cup\{j\}) - val(S)\right]$$

Where F is the set of all features, S is a subset of features, val(S) is the model output on the subset S. ϕj is the SHAP value of feature j, whose absolute value represents the influence of the feature on the prediction result, and the sign indicates the direction of the influence (positive or negative). SHAP provides global feature importance and local sample explanation (such as force plots), able to clearly answer the question "why did the model make a specific prediction".

Additionally, LIME (Local Interpretable Model-agnostic Explanations) is also a common local interpretability method. It approximates the local behavior of a complex model by constructing a simple, interpretable surrogate model (such as a linear model) near the sample to be explained.

This study uses the publicly available PIMA Indians Diabetes Database. This dataset, provided by the National Institute of Diabetes and Digestive and Kidney Diseases, is one of the benchmark datasets for binary classification prediction in medical machine learning research. The dataset contains 768 samples (female patients), each consisting of 8 medical diagnostic features and 1 target variable:

1.Pregnancies: Number of pregnancies.
2.Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test.
3.BloodPressure: Diastolic blood pressure (mm Hg).
4.SkinThickness: Triceps skin fold thickness (mm).
5.Insulin: 2-hour serum insulin (μU/ml).
6.BMI: Body mass index (weight in kg/(height in m)^2).
7.DiabetesPedigreeFunction: Diabetes pedigree function (assesses family genetic influence).
8.Age: Age (years).
9.Outcome: Target variable (0 indicates non-diabetic, 1 indicates diabetic).

This dataset is typical medical data, presenting challenges such as class imbalance (500 non-diabetic cases, 268 diabetic cases) and missing values (in the original data, certain feature values are 0, e.g., Glucose=0, which is physiologically impossible and usually treated as missing).

High-quality data preprocessing is the foundation for building high-performance prediction models. This study performed the following key preprocessing steps targeting the characteristics of the PIMA dataset:

1.Missing Value Handling: As mentioned, the dataset contains missing values represented by 0 values that are physiologically impossible (e.g., Glucose, BloodPressure, SkinThickness, Insulin, BMI). This study uses a median imputation strategy to handle these missing values. The median is insensitive to outliers and can better preserve the distribution characteristics of the data, making it suitable for medical data.

2.Data Standardization: Due to the significant differences in the units and value ranges of various features (e.g., Age and Insulin), to eliminate the impact of scale on the model (especially distance-based models) and accelerate model convergence, this study uses Z-Score standardization. This method transforms feature values to a normal distribution with a mean of 0 and a standard deviation of 1.

$$X_{std} = \frac{X - \mu}{\sigma}$$

where μ is the feature mean and σ is the feature standard deviation.

3.Handling Class Imbalance: The positive samples (Outcome=1) in the original data account for about 34.9%, indicating significant class imbalance. If used directly for training, the model would tend to predict the majority class, leading to a low recognition rate (recall) for diabetic patients. This study uses the SMOTE (Synthetic Minority Over-sampling Technique) oversampling technique. The basic idea of SMOTE is to interpolate minority class samples, synthesizing new "artificial" minority class samples in the feature space, thereby balancing the dataset.

After the above preprocessing pipeline, we obtained a clean, balanced, and standardized dataset, laying a solid foundation for subsequent model training and feature engineering.

III. Model Construction and Methodology

This study employs a systematic approach to construct a diabetes risk prediction model. Its core framework consists of five key stages, forming a complete research cycle. The first stage involves data preprocessing to address common issues in medical data such as missing values, inconsistent scales, and class imbalance, establishing a high-quality data foundation for subsequent analysis. The second stage screens the most predictive feature subset through feature engineering to reduce model complexity. The third stage adopts an iterative optimization strategy to determine the best model parameters and improve predictive performance. The fourth stage uses interpretability techniques to reveal the model's decision-making mechanism and enhance the credibility of the results. The final stage focuses on practical application, designing a lightweight deployment scheme to provide a technical path for future clinical translation. This framework ensures both the academic rigor of the model and the feasibility of practical application.

This study innovatively implements the "iterative" concept as a systematic hyperparameter optimization process. By constructing a multidimensional parameter space including the number of trees, maximum depth, and minimum samples split for nodes, a comprehensive exploration is conducted using the grid search method. The optimization process uses 5-fold cross-validation as the core technical means, dividing the training data into 5 equal subsets, sequentially using each subset as the validation set and the rest as the training set, and cyclically training and validating the model.

This iterative process of cross-validation effectively prevents model overfitting and ensures that the obtained parameter combinations have strong generalization ability. The F1-Score is ultimately used as the core indicator for model selection, as this metric can balance precision and recall, making it particularly suitable for classification problems with class imbalance like diabetes prediction. Through this systematic "iterative" optimization strategy, we can automatically identify the hyperparameter configuration with the best performance from a large number of possible parameter combinations, significantly improving the model's prediction accuracy and stability.

The "multimodal" aspect of this study is reflected in the comprehensive use of multi-dimensional features such as physiological indicators, life history, and genetic factors for risk assessment. To identify the most clinically significant predictors from these features, Recursive Feature Elimination (RFE), a greedy algorithm, is used for feature screening.

The RFE technique builds a feature importance ranking model and iteratively removes the least contributing features, gradually refining the feature set. This process is based on the Gini importance index provided by the Random Forest algorithm, which quantifies the average ability of each feature to reduce data impurity during node splitting. By setting an optimal feature quantity threshold, the most discriminative feature subset for diabetes prediction is ultimately retained.

This feature screening process not only improves the computational efficiency of the model and reduces the risk of overfitting but, more importantly, it can reveal the key factors most related to the onset of diabetes, providing an important basis for subsequent clinical interpretation and making the model more focused on core risk indicators.

To address the interpretability issue of machine learning models, this study introduces the SHAP (SHapley Additive exPlanations) unified framework for model decision interpretation. SHAP, based on the Shapley value concept from game theory, provides a quantitative allocation scheme for the contribution of each feature to the final prediction result.

SHAP analysis provides explanations from both global and local dimensions. Global interpretation uses summary plots to show the overall direction and strength of influence of all features on the model output, identifying key drivers that push diabetes risk prediction. Local interpretation, aimed at the prediction result of a single sample, uses force plots to visualize how each feature pushes the base value (average prediction value) up or down to the final prediction value, clearly showing the specific basis for the model's risk judgment for a specific individual.

This interpretability analysis builds a bridge between model prediction and clinical understanding, enabling doctors to intuitively understand the model's decision-making process, identify key factors affecting a patient's risk status, and thus provide transparent and trustworthy decision support for personalized intervention.

To translate the research results into practical applications, this study proposes a conceptual design for a lightweight dynamic warning system. This scheme uses the ONNX (Open Neural Network Exchange) open model format as the technical core, converting the trained Random Forest model into this cross-platform standard format, ensuring the model can be efficiently deployed in various hardware environments and programming languages.

The system design adopts a client-server architecture pattern. The server side is responsible for regular model updates and optimization, maintains the patient database, and provides batch risk assessment services. The edge or mobile device integrates a lightweight ONNX runtime engine, capable of loading the pre-trained model and performing real-time predictions, greatly reducing dependence on network connectivity and data transmission latency.

The warning system interface design includes a risk visualization panel, combining model prediction results with SHAP explanatory output. It not only displays the final risk score but also provides the basis for risk assessment, such as how specific abnormal physiological indicators led to the increased risk. Simultaneously, the system also designs a risk trend tracking function, which can record and display changes in the user's risk level over time, providing data support for long-term health management. This conceptual design provides a solid technical blueprint and implementation path for the future development of a practically usable diabetes risk dynamic monitoring tool.

IV. EXPERIMENTS AND ANALYSIS

The experiments were based on a Python 3.9 environment, using mainstream machine learning libraries such as Scikit-learn, SHAP, and Imbalanced-learn. The hardware configuration was an Intel i7-12700H CPU, 16GB RAM, and the operating system was Windows 11. To comprehensively evaluate model performance, this study adopted the following evaluation metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC curve area. Among these, F1-Score was used as the primary basis for model selection because it better balances precision and recall in class imbalance problems.

Through Recursive Feature Elimination (RFE) combined with the Gini importance from Random Forest, we gradually eliminated features with lower predictive contributions. The final retained 6 key features were, in order: Glucose, BMI, Age, DiabetesPedigreeFunction, Insulin, and Pregnancies. After feature selection, the model maintained performance while significantly reducing computational complexity and improving model interpretability and deployment efficiency.

To verify the effectiveness of the proposed method, we systematically compared the optimized iterative Random Forest model with various mainstream machine learning models. Experimental results showed that the Logistic Regression model had an accuracy of 76.2%, precision of 70.1%, recall of 64.3%, F1-Score of 0.671, and an AUC value of 0.812. The Support Vector Machine showed slight improvements in all metrics, with accuracy reaching 78.1%, precision 72.3%, recall 68.2%, F1-Score 0.702, and AUC 0.834.

The traditional Random Forest model demonstrated stronger classification ability, with accuracy increasing to 81.4%, precision 76.8%, recall 72.4%, F1-Score 0.745, and AUC 0.879. The XGBoost model further improved on this basis, achieving an accuracy of 82.6%, precision 78.2%, recall 74.1%, F1-Score 0.761, and AUC 0.891.

The iterative Random Forest model proposed in this study achieved the best performance across all evaluation metrics, with an accuracy of 84.3%, precision of 80.1%, recall of 77.9%, F1-Score of 0.790, and an AUC value of 0.912. The significant improvement in recall and F1-Score, in particular, indicates that this model has better sensitivity and classification balance in identifying diabetic patients, making it more effective in serving the clinical needs of early diabetes screening.

Based on the SHAP framework, we interpreted the model from both global and local levels:

Global Interpretation: The SHAP summary plot showed that Glucose, BMI, and Age were the most important features affecting diabetes risk, with their absolute SHAP values far exceeding those of other features.

Local Interpretation: Taking a high-risk patient as an example, the force plot showed that their high Glucose value, high BMI, and relatively high age were the main factors driving the model's judgment of diabetes, while a lower Insulin level slightly reduced the risk score.

SHAP analysis not only enhanced the model's credibility but also provided intuitive decision-making basis for doctors, achieving "prediction-explanation" integration.

V. CONCEPTUAL DESIGN OF DYNAMIC WARNING SYSTEM AND IMPLEMENTATION PROSPECTS

The system adopts a Client-Server (C/S) architecture, divided into three layers:

Data Layer: Responsible for storing user health data and model parameters.

Service Layer: Provides model inference, SHAP explanation generation, risk assessment, and warning services.

Application Layer: An interactive interface for users and doctors, supporting risk visualization and historical trend viewing.

To achieve efficient model operation on mobile devices or edge nodes, this study proposes the following lightweight path:

Use the ONNX format to convert the trained Random Forest model into a cross-platform format.

Use model pruning and quantization techniques to further compress the model size.

Integrate the ONNX Runtime engine on edge devices to achieve low-latency inference.

This scheme significantly reduces hardware resource requirements while maintaining model performance, making it suitable for resource-constrained scenarios such as community hospitals and home health terminals.

The system operation process is as follows:

1.User input or automatic device collection of physiological data.

2.The edge-side model performs real-time prediction and generates a risk score.

3.If the score exceeds a threshold, the system triggers a warning and generates a SHAP explanation report.

4.The report is pushed to the user and doctor ends, supporting further diagnosis and intervention.

5.The system records each evaluation result, forming a personal risk trend chart to support long-term health management.

VI. Conclusion and Outlook

Centering on the clinical needs of early diabetes screening, this study constructed a dynamic warning and interpretable decision model based on a multimodal iterative forest. By introducing iterative optimization strategies, recursive feature selection, and the SHAP explanation framework, the model's predictive performance and interpretability were significantly improved. Experimental results show that the proposed method performs excellently on the PIMA dataset and has good potential for clinical translation.

Although this study has achieved certain results, it still has the following limitations:

The data source is single, using only the PIMA dataset; future work needs to validate the model's generalization ability on multi-center, large-scale clinical data.

The model has not yet been integrated with electronic health record systems, and the dynamic data update mechanism is not yet perfect.

The edge deployment scheme is still in the conceptual stage and needs to be tested for performance and stability in actual hardware environments.

Future work will focus on the following directions:

Introducing multimodal data (such as images, text diagnostic records) to enhance the model's information dimensions.

Developing a federated learning framework to achieve collaborative model training under the premise of protecting data privacy.

Promoting cooperation with hospitals to conduct prospective clinical validation and promote the model's landing application.

References

[1] Zhou, B., et al. "Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants." Lancet, 2024. DOI:10.1016/S0140-6736(24)02317-1.

[2] Jia, W., et al. "Diabetes in China: epidemiology, pathophysiology and multi-omics." Nat. Metab. (Nature Metabolism), 2025. DOI: 10.1038/s42255-024-01190-w.

[3] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., Stiglic, G. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." Sci. Rep. (Scientific Reports), vol. 10, no. 1, pp. 1–12, 2020. DOI:10.1038/s41598-020-68771-z.

[4] Yang, M. J., Tang, J. J. "Research on Diabetes Prediction Based on Random Forest Algorithm." J. Med. Inform. (Journal of Medical Informatics), 19.

[5] Johnson, R. L., Patel, V. K. "Dynamic Weighted Iterative Random Forest for Imbalanced Medical Data." IEEE Trans. Biomed. Eng. (IEEE Transactions on Biomedical Engineering), 2022.

[6] Wang, Q., Rudin, C. "Interpretable Rule Extraction from Iterative RF for Diabetes Onset Prediction." J. Biomed. Inform. (Journal of Biomedical Informatics), 2023.

[7] Breiman, L. "Random Forests." Mach. Learn. (Machine Learning), 2001.

[8] Lundberg, S. M., Lee, S. I. "A Unified Approach to Interpreting Model Predictions." Adv. Neural Inf. Process. Syst. (Advances in Neural Information Processing Systems), 2017.

[9] Ribeiro, M. T., Singh, S., Guestrin, C. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.

[10] Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." J. Artif. Intell. Res. (Journal of Artificial Intelligence Research), vol. 16, pp. 321–357, 2002.

Acknowledgment

This work was supported by the Anhui University of Finance and Economics Undergraduate Scientific Research Innovation Fund.(Project Approval Number: XSKY25144)