Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 6
GEMFE 2026

A Multi-Modal Machine Learning Framework for Type 2 Diabetes Risk Stratification and Progression Time Prediction

Zizhong Yang [1, #], Kailun Li [1, *, #], Xiaoqi Lian [1], Ziqiu Yang [2]

[1] School of Electronics and Information Engineering, Guangdong Polytechnic Normal University, Guangzhou, China, 510450

[2] School of Mechanical and Electrical Engineering & Automation, Foshan University, Foshan, China, 528225

* Corresponding Author Email: lkl07_ef@163.com

[#]These authors contributed equally.

Abstract. The growing global burden of Type 2 Diabetes (T2D) urgently requires advanced predictive tools that go beyond traditional binary risk classification. This study presents an innovative machine learning framework that integrates multimodal data, including clinical parameters, lifestyle factors, Nuclear Magnetic Resonance (NMR) metabolomics data, and 12-lead Electrocardiogram (ECG) signals, to simultaneously predict T2D risk and disease progression time. This study employs a two-stage prediction framework: an AdaBoost classifier optimized by Optuna is used for risk stratification (AUROC=0.9468, F1=0.8857), combined with an Elastic Net regression model for progression time estimation (R^2=0.7649, MSE=15.1535). This framework achieves superior predictive performance compared to single-source models by synergistically integrating complementary data modalities, guides clinical intervention timing through quantitative estimation of individualized progression time, and enhances model interpretability through rigorous feature importance analysis, demonstrating key advantages. This method, which combines high accuracy with clinical utility, fills a critical gap in current T2D prediction methodologies and provides a practical tool for personalized prevention strategies and optimized healthcare resource allocation.

Keywords: Warwick Evans; Publishing; These keywords will also be used by the publisher to produce a keyword index.

1. Introduction

Over the past 50 years, the ageing population has accelerated and the incidence of Type 2 Diabetes (T2D) has continued to rise. Currently, nearly half of T2D patients are elderly people aged 26 to 55 [1]. Elderly T2D patients face serious challenges, and the disease imposes a heavy burden on society and the economy through microvascular/macrovascular complications, reduced quality of life, and high medical costs. Therefore, early identification of high-risk populations and prediction of disease progression are critical for implementing precise prevention, delaying onset, and optimizing the allocation of medical resources.

Existing T2D prediction studies have mainly focused on binary risk assessment, with traditional methods and machine learning models being widely used. Machine learning models have demonstrated advantages in capturing complex non-linear relationships: Henock M. Deberneh [2] and Intaek Kim constructed a T2D prediction model using multidimensional electronic health record features; Ram D. Joshi et al. [3] and Chandra K. Dhakal used logistic regression and decision trees to identify key risk factors; Nikos Fazakis et al. [4] employed a dual-objective genetic algorithm to optimize weights and developed a long-term risk prediction model based on the WeightedVotingLRRFs ensemble method, validating the superiority of ensemble methods; Alexandra Kautsky-Willer et al. [5] emphasized gender differences in risk factors, complications, and treatment responses for Type 2 Diabetes, highlighting the importance of gender-specific management strategies; Luis Fregoso-Aparicio et al. [6] systematically evaluated machine learning and deep learning models, found that tree-based algorithms performed best, and emphasized the value of data

balancing and feature selection; Leon Kopitar et al. [7] compared machine learning with traditional regression models and found no significant advantages for machine learning, but emphasized the importance of interpretability and stability; Md. Kamrul Hasan et al. [8] developed a predictive framework based on AUC-weighted AdaBoost and XGBoost ensemble learning, validating the effectiveness of ensemble methods; Rashi Rastogi and Mamta Bansal [9] developed a scheme that includes logistic regression (accuracy of 82.46%), support vector machines, random forests, and naive Bayes models, demonstrating the value of data mining. Isfafuzzaman Tasin et al. [10] combined XGBoost with explainable artificial intelligence (LIME, SHAP) to build a diabetes prediction system, proving its value in clinical practice.

Existing research has key gaps: most models only perform binary classification and cannot quantify the disease progression time of high-risk populations, and they often rely on single-modal data, failing to fully utilize the synergistic advantages of multi-source heterogeneous data. To address this issue, this study proposes an innovative multi-stage machine learning framework to achieve accurate T2D risk classification and introduces disease progression time prediction. The framework integrates individual baseline characteristics, lifestyle data, Nuclear Magnetic Resonance (NMR) data, and 12-lead Electrocardiogram (ECG) signals to construct a high-information-density feature system. In the classification stage, the AdaBoost model is optimized by Optuna to achieve excellent performance (AUROC = 0.9743, AUPRC = 0.9468, F1 = 0.8857). In the time prediction stage, an ElasticNet regression model with Bayesian parameter tuning achieved $R^2 = 0.7649$ and MSE = 15.1535. L1/L2 regularization ensured model robustness and improved result interpretability, providing reliable support for clinical decision-making. (Data sources include publicly available datasets from the Artificial Intelligence Crowdsourcing Collaboration Community. Modelling data can be downloaded via the following link: http://www.aisccc.cn/database/data-details?id=122&type=info).

2. Study on T2D Risk Prediction and Progression Time Analysis

2.1. Iterative Sample Reweighting Approach for T2D Classification Using AdaBoost

AdaBoost is an ensemble learning method that combines multiple weak classifiers to build a strong classifier. Its core idea is to iteratively adjust sample weights so that subsequent classifiers pay more attention to samples that were previously misclassified. The algorithm flow is as follows:

(1) The Initialise sample weights:

$$w_i^{(1)} = \frac{1}{\epsilon}, \forall i = 1,2, \dots, n., \tag{1}$$

(2) Iteratively train weak classifiers and calculate their weights:

First, train weak classifiers and calculate weighted errors

$$\epsilon_t = \sum_{i=1}^{N} w_i \, I(h_t(x_i) \neq y_i), \tag{2}$$

Second, calculate classifier weights

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right), \tag{3}$$

Third, update sample weights

$$w_i \leftarrow w_i \cdot \exp(\alpha_t I(h_t(x_i) \neq y_i)), \tag{4}$$

(3) Update sample weights and integrate weak classifiers:

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)), \quad (5)$$

2.2. ElasticNet Regression for T2D Progression Time Prediction

This study uses the ElasticNet model to predict the disease duration of patients diagnosed with T2D by the AdaBoost model. This model is a linear regression method that combines L1 (LASSO) and L2 (Ridge) regularization, and its objective function is expressed as.

$$\min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \left[(1-\alpha)^{\frac{1}{2}} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right], \quad (6)$$

In this equation, $y \in R^n$is the target vector, $X \in R^{n \times p}$is the feature matrix, $\beta \in R^p$is the regression coefficient vector, $\lambda \geq 0$ is the regularisation strength parameter, $\alpha \in [0,1]$ is the mixing coefficient, $\|y - X\beta\|_2^2$ is the squared loss term, $\|\beta\|_1$ is the L1 regularisation term, and $\|\beta\|_2^2$ is the L2 regularisation term.

The model is initialized with β as a zero vector or random value, and iterative optimisation is performed using the coordinate descent method:

Fix the remaining coefficients, based on the residual$r = y - X\beta$, and update the current coefficient β_j using a soft threshold function:

$$\beta_j \leftarrow S\left(X_j^T r, \lambda_1, \lambda_2\right), \quad (7)$$

3. Results

3.1. AdaBoost-Based T2D Risk Prediction: Optimization and Feature Importance

(1) Optuna hyperparameter optimization

Optuna employs a tree-based Bayesian optimisation algorithm that guides the search process through a probability model of the objective function. Compared to grid search, this method offers advantages in terms of memory and computational efficiency. After several iterations, Optuna can quickly approach the global optimum, thereby making the optimisation process more economical and efficient. Therefore, this study uses Optuna to perform hyperparameter optimisation on the AdaBoost model. The results are shown in Table 1.

Table 1. The parameter tuning results of the AdaBoost model

Name of parameter	Adjustment range	Adjustment results
N_estimators	50-200	137
Learning_rate	0.01-1.0	0.3757

As shown in Table 1, the results of AdaBoost model hyperparameter tuning. After parameter tuning based on the tree structure Bayesian optimisation algorithm in the Optuna framework, the optimal parameter combination for the model was finally determined as: N_estimators is 137, and Learning_rate is 0.3757. This set of parameters was obtained through probabilistic model-guided search and multiple iterations within the predefined parameter tuning range (N_estimators: 50–200, Learning_rate: 0.01–1.0). It represents a global approximate optimal solution that ensures model performance while significantly improving the efficiency and computational economy of the parameter tuning process.

(2) Confusion Matrix and Model Evaluation

To evaluate model performance, this study calculated Recall and False Positive Rate (FPR) based on the confusion matrix, and combined them with Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and F1 score for comprehensive evaluation. The confusion matrix is detailed in Table 2, and the specific evaluation results of the AdaBoost model are shown in Table 3.

Table 2. Confusion matrix of the AdaBoost model

Predicted value / True value	Healthy	Sicken
Healthy	374	18
Sicken	14	123

Table 3. AdaBoost model evaluation metrics

	Recall	FPR	AUROC value	AUPRC value	F1 value
AdaBoost	0.9541	0.1022	0.9743	0.9468	0.8857

As shown in Tables 2 and Table 3, the evaluation results of the AdaBoost model are as follows: among non-diabetic individuals, there were 374 true negatives and 18 false positives (FPR = 0.1022); among diabetic individuals, there were 123 true positives and 14 false negatives (Recall = 0.9541). The misclassification rate is significantly lower than the false negative rate, indicating superior performance in reducing misclassifications. Additionally, AUROC = 0.9743, AUPRC = 0.9468, and F1 = 0.8857, demonstrating excellent overall model performance.

(3) Individual baseline information, lifestyle habits, and NMR data characteristics

By training the model, feature importance scores were extracted to identify the most influential features of individual baseline information, lifestyle habits, and NMR data in diabetes prediction. Feature importance is shown in Figure 1.

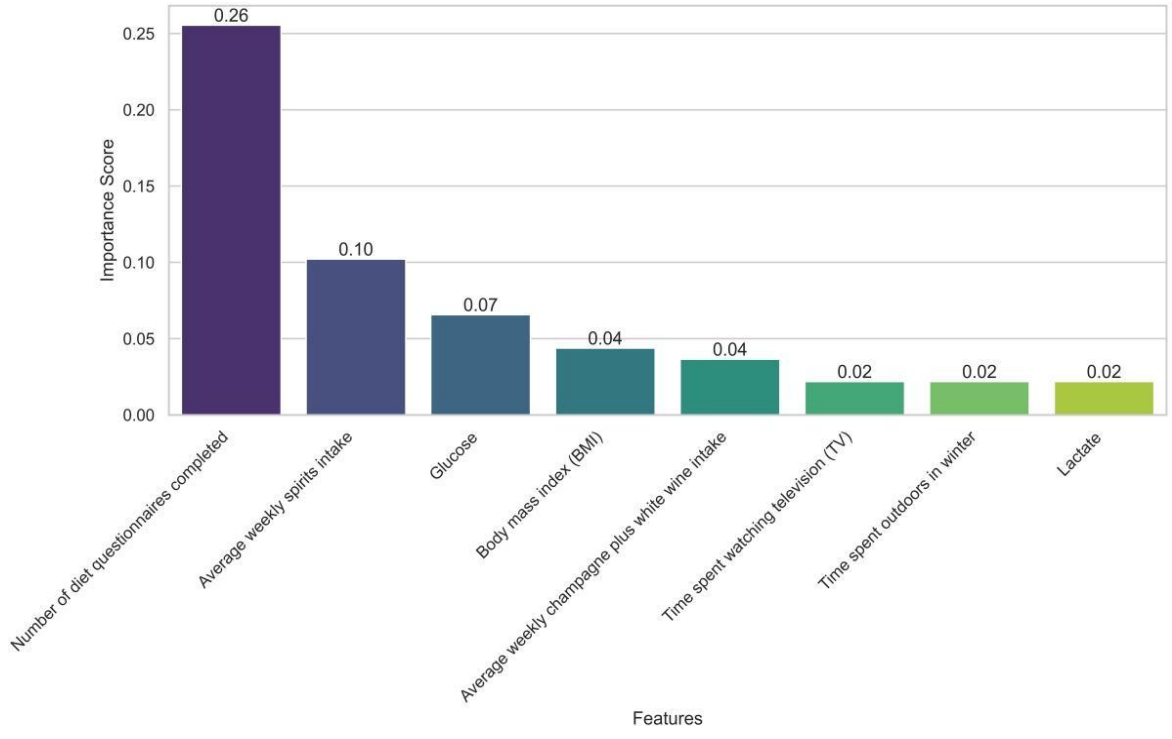

Figure 1. Top 8 Feature Importance from AdaBoost Model

As shown in Figure 1, the feature importance ranking obtained through the innovative AdaBoost model is presented, identifying Glucose, Number of Diet Questionnaires Completed, Average Weekly Spirits Intake, and BMI as the core predictive factors for T2D risk stratification. This study demonstrates significant methodological innovation, effectively capturing complex nonlinear relationships (such as the association between alcohol intake and diabetes risk) and seamlessly integrating multimodal clinical indicators, behavioral data, and NMR biomarkers. The model achieves high predictive accuracy while providing clinically interpretable decision pathways, validating its ability to extract biologically meaningful features from heterogeneous health data. These advancements address key limitations of current T2D prediction methods and provide an innovative framework for risk stratification.

(4) Twelve-lead ECG data characteristics

By training the model and extracting feature importance scores, the following are the most influential features of ECG data in T2D classification. Feature importance is shown in Figure 2.

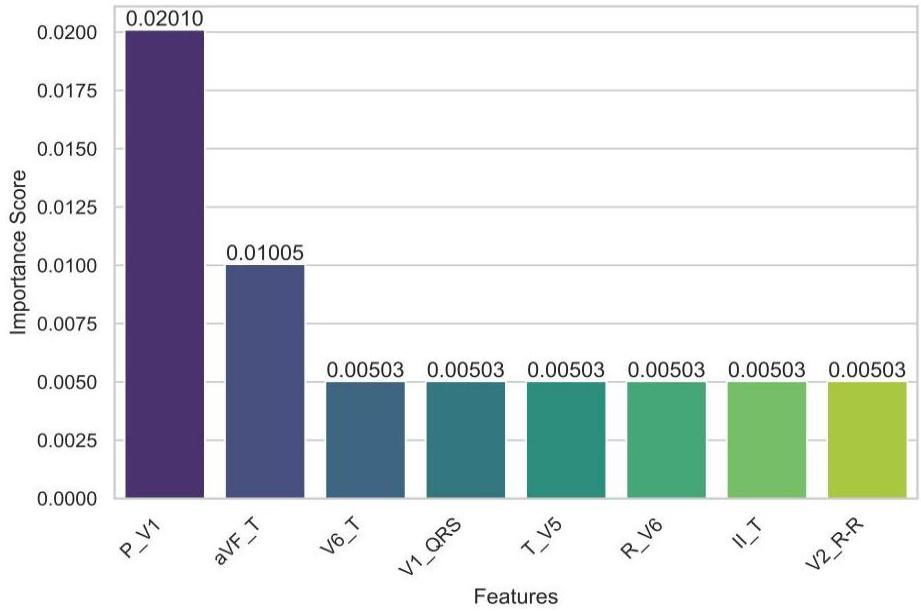

Figure 2. Top 8 Feature Importance from AdaBoost Model

As shown in Figure 2, based on the analysis of the importance of ECG data features, P_V1 is the most critical ECG indicator for predicting T2D, followed by aVF_T, with V6_T also being valuable. P_V1 reflects the P wave characteristics of the V1 lead, and its abnormal changes are associated with diabetes-induced atrial remodeling and neurological disorders, enabling the monitoring of early abnormalities in atrial electrical activity; aVF_T reflects the T wave characteristics of the aVF lead, and its changes are caused by metabolic disorders and neuropathy, aiding in the early screening and risk assessment of T2D and improving the monitoring of cardiac complications; V6_T is associated with the T wave of the V6 lead, and metabolic abnormalities affect ventricular repolarization, providing electrophysiological evidence of cardiac involvement. These features provide biomarkers for early cardiac electrophysiological changes in T2D from a multi-lead perspective, aiding disease warning and risk assessment. This has positive implications for improving the management of cardiovascular complications in T2D and establishing a cardiac abnormality monitoring system.

3.2. ElasticNet Regression Performance and Time-to-Onset Prediction

(1) Bayesian parameter tuning

Bayesian parameter tuning is an efficient global optimization method suitable for hyperparameter tuning in high-dimensional spaces and models with high computational costs. Compared with grid and random searches, it can intelligently select points, balance exploration and exploitation, and adapt

to various objective functions by utilizing historical results. Therefore, this study uses Bayesian optimization parameter tuning for the ElasticNet model, and the results are shown in Table 4.

Table 4. The parameter tuning results of the ElasticNet regression model

Name of parameter	Adjustment range	Adjustment results
Alpha	0.00001-1.0	0.2171
L1_ratio	0-1.0	0.9994
Max_iter	100-1000	665

As shown in Table 4, the results of hyperparameter tuning of the ElasticNet regression model using the Bayesian optimisation method are presented. Within the predefined parameter ranges (Alpha: 0.00001–1.0; L1_ratio: 0–1.0; Max_iter: 100–1000), through multiple rounds of iterative sampling and evaluation, the optimal hyperparameter combination was finally determined as: Alpha = 0.2171, L1_ratio = 0.9994, Max_iter = 665. This set of parameters effectively balances the L1 and L2 regularisation constraints of the model, improving prediction accuracy while significantly increasing the efficiency and convergence stability of parameter search.

(2) Model evaluation

To evaluate the performance of the ElasticNet model, this study comprehensively evaluated it using the coefficient of determination (R^2), mean square error (MSE), and Pearson correlation coefficient (PCC). The specific evaluation results of the ElasticNet model are shown in Table 5.

Table 5. ElasticNet model evaluation metrics

	R^2score	MSE	PCC
ElasticNet	0.7649	15.1535	0.8772

This study used the ElasticNet model to predict the onset time of T2D in individuals who may be at risk of developing the disease. Partial time prediction results are shown in Table 6.

Table 6. Partial results of ElasticNet model for predicting onset time

Patient Number	1302242.0	1305609.0	1306252.0	1308485.0	1312252.0	1322389.0
Predicted Disease Age	70	63	51	63	49	61

(3) Feature selection for time prediction

By training the model and extracting feature importance scores, we can identify the most influential features in predicting the onset time of T2D. The feature importance of the ElasticNet model is shown in Figure 3.

As shown in Figure 3, the ElasticNet model combines L1 and L2 regularisation to perform feature selection, enabling it to more effectively screen out features most relevant to the target variable. This highlights key features related to T2D, such as Free Cholesterol in IDL, Tyrosine, and Triglycerides in Large HDL. This indicates that the ElasticNet model focuses more on the underlying mechanisms that cause diabetes and is suitable for predicting the duration of diabetes.

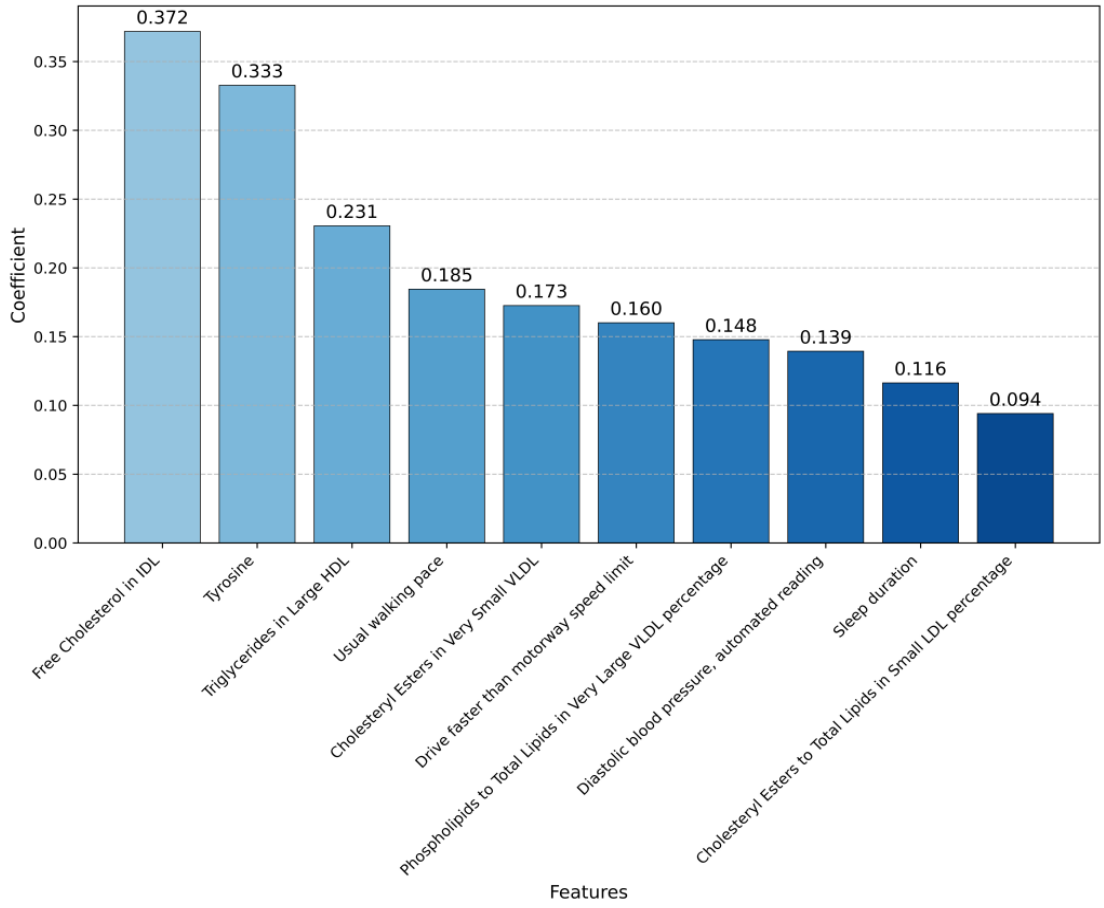

Figure 3. Importance of top 15 features of ElasticNet

(4) Deviation between the true values and the predicted values

The scatter plot of 'true values and predicted values' intuitively reflects the relationship between the model's predicted values and the true values, showing the systematic deviation between the predicted values and the true values. The results are shown in Figure 4.

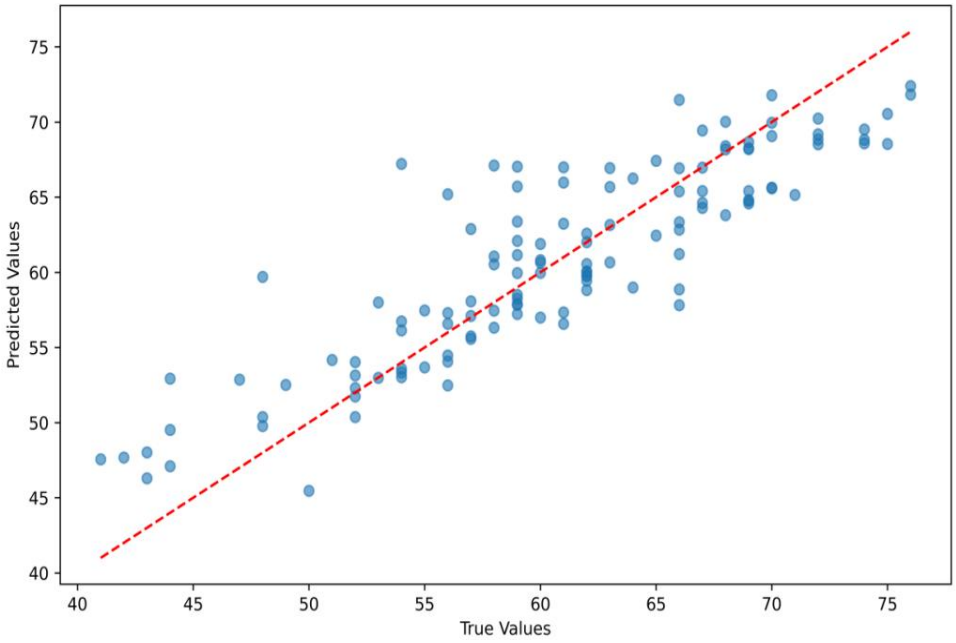

Figure 4. Scatter plot of the true values and predicted values of ElasticNet

As can be seen from Figure 4, the scatter plot of the ElasticNet model shows that the data points are concentrated, and the predicted values and true values are basically in perfect linear relationship,

indicating that the ElasticNet model performs well in fitting data and can more accurately capture potential linear patterns.

(5) Residual analysis

The residual scatter plot intuitively shows the distribution characteristics of model residuals and can effectively identify outliers or outliers, providing an important visual basis for evaluating model fit and stability. Through the analysis of the residual pattern, we can further determine whether the model assumptions are reasonable. The residual analysis results of the ElasticNet model are shown in Figure 5.

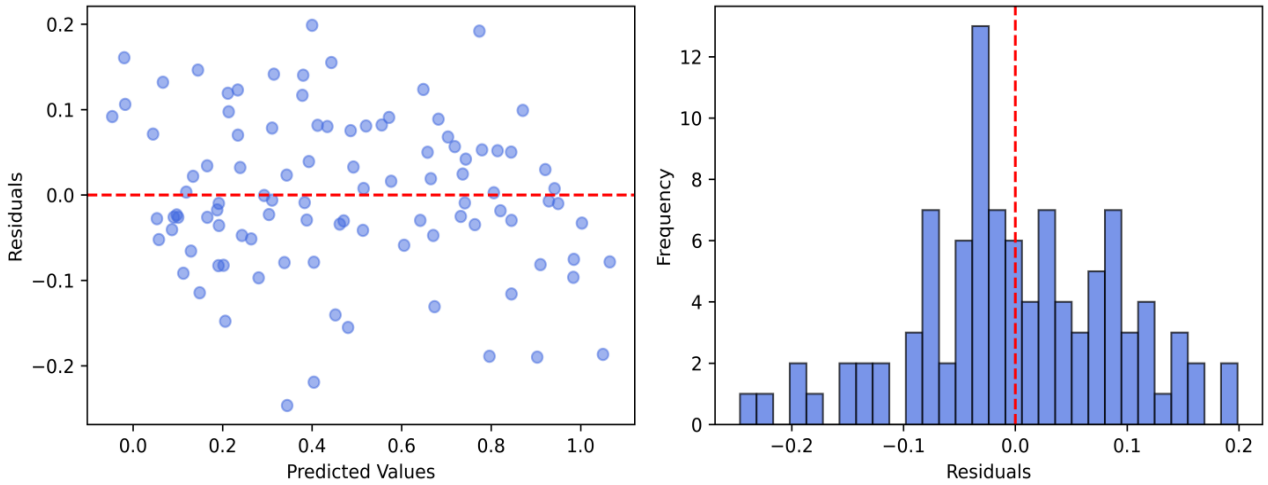

Figure 5. ElasticNet residual analysis dual view

As shown in Figure 5, the residual scatter plot of the ElasticNet model performs well, with data points distributed relatively closely and adjacent to the 0 axis, indicating that the model has excellent predictive performance. The residual histogram shows a concentrated and relatively symmetrical distribution, which is highly consistent with the actual observed values, indicating that the ElasticNet model can fit the data well. The model demonstrates significant application value in predicting the onset time of T2D. Its outstanding interpretability makes the results not only accurate but also easy to understand, effectively serving clinical decision-making. The stable predictive performance combined with good interpretability makes this model a valuable and reliable tool for personalized medical decision-making.

4. Conclusion

This study proposes an innovative machine learning framework for predicting Type 2 Diabetes (T2D) by integrating multimodal data, including clinical indicators, lifestyle data, Nuclear Magnetic Resonance (NMR) metabolomics features, and 12-lead Electrocardiogram (ECG), to achieve simultaneous calculation of disease risk classification and progression time prediction. The framework adopts a two-stage prediction architecture: an AdaBoost classifier optimized by Optuna (AUROC = 0.9743, F1 = 0.8857) is used for risk classification, combined with an ElasticNet regression model (R^2=0.7649, MSE=15.1535) for progress time estimation, whose performance significantly outperforms traditional single-task prediction methods.

This study overcomes the limitations of existing T2D prediction models by achieving a double breakthrough through an innovative two-stage prediction framework: First, an optimized AdaBoost model (Recall = 0.9541, FPR = 0.1022) was used to accurately identify high-risk individuals. Then, ElasticNet regression was employed to quantify their expected disease progression trajectories (R^2 = 0.7649, MSE = 15.1535), providing a basis for the precise timing of clinical interventions. The synergistic integration of multi-source data (clinical indicators, lifestyle factors, NMR biomarkers, and ECG signals) significantly improved prediction accuracy (AUROC = 0.9743, AUPRC = 0.9468). Feature importance analysis identified key indicators such as Glucose, Number of Diet

Questionnaires Completed, Average Weekly Spirits Intake, and BMI as contributors to early risk identification, with ECG-derived features proven to have significant value in T2D risk assessment. The model maintains high accuracy while offering good interpretability, enabling doctors to make clinical decisions based on transparent risk assessments.

This study addresses the critical issue of individualized prognosis assessment in T2D prediction by developing a two-stage model that combines risk classification and progression time prediction. Experimental results demonstrate that the AdaBoost classifier achieves high classification accuracy on the test set, while the ElasticNet regression model exhibits significant regression prediction performance, validating the framework's effectiveness in precise prediction. Future research will focus on multi-centre clinical validation and exploring the application of the model in a tiered healthcare system. The method proposed in this study provides technical support for diabetes risk stratification and individualized intervention timing selection, demonstrating potential for translation into clinical decision-making tools.

References

[1] Ahmad E, Lim S, Lamptey R, et al. Type 2 diabetes [J]. The Lancet, 2022, 400 (10365): 1803 - 1820.

[2] Deberneh H M, Kim I. Prediction of type 2 diabetes based on machine learning algorithm [J]. International journal of environmental research and public health, 2021, 18 (6): 3317.

[3] Joshi R D, Dhakal C K. Predicting type 2 diabetes using logistic regression and machine learning approaches [J]. International journal of environmental research and public health, 2021, 18 (14): 7346.

[4] Fazakis N, Kocsis O, Dritsas E, et al. Machine learning tools for long-term type 2 diabetes risk prediction [J]. ieee Access, 2021, 9: 103737 - 103757.

[5] Kautzky-Willer A, Leutner M, Harreiter J. Sex differences in type 2 diabetes [J]. Diabetology, 2023, 66 (6): 986 - 1002.

[6] Fregoso-Aparicio L, Noguez J, Montesinos L, et al. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review [J]. Diabetology & metabolic syndrome, 2021, 13 (1): 148.

[7] Kopitar L, Kocbek P, Cilar L, et al. Early detection of type 2 diabetes mellitus using machine learning-based prediction models [J]. Scientific reports, 2020, 10 (1): 11981.

[8] Hasan M K, Alam M A, Das D, et al. Diabetes prediction using ensembling of different machine learning classifiers [J]. IEEE Access, 2020, 8: 76516 - 76531.

[9] Rastogi R, Bansal M. Diabetes prediction model using data mining techniques [J]. Measurement: Sensors, 2023, 25: 100605.

[10] Tasin I, Nabil T U, Islam S, et al. Diabetes prediction using machine learning and explainable AI techniques [J]. Healthcare technology letters, 2023, 10 (1-2): 1 - 10.