Diabetes Risk Prediction Based on Clinical and Lifestyle Data

Tiancheng Mao[1*]

[1]Computer Science and Technology, Huazhong University of Science and Technology, 430074 Wuhan, China

Abstract. With the continuous increase in global diabetes incidence, the use of integrated health data for early detection of risk and timely intervention has attracted growing attention in public health. Based on the National Health and Nutrition Examination Survey (NHANES), this study combines demographic characteristics, clinical test results, laboratory indicators, and lifestyle patterns to construct a diabetes risk prediction model using a Multi-Layer Perceptron (MLP). Key physiological markers—including Glucose, HbA1c, BMI, and Insulin—were analysed together with lifestyle-related variables such as dietary intake, physical activity, and alcohol use. The developed model attains an accuracy of about 85% and an AUC of 0.89 on the test dataset. Findings suggest that biomarkers associated with blood glucose regulation play a central role in predicting diabetes, while healthier diet structure and regular exercise contribute to reducing individual risk. Moreover, Principal Component Analysis (PCA) uncovers complementary associations between clinical measures and lifestyle features, improving the interpretability of the model. Overall, the study demonstrates an effective data-driven strategy for early diabetes risk assessment and personalised health management, while providing insights into the application of artificial intelligence techniques in public health research.

1 Introduction

Type 2 diabetes mellitus (T2DM) is one of the most prevalent chronic metabolic disorders worldwide and has created a substantial burden on healthcare systems. Identifying high-risk individuals at an early stage is essential for diabetes prevention and management efforts. Although traditional statistical models—such as logistic regression—are frequently used due to their interpretability and clear variable contribution analysis, they face challenges when processing nonlinear feature interactions and high-dimensional data structures [1]. With the increasing availability of electronic health records (EHR) and data collected from wearable devices, machine learning (ML) methods have been widely adopted in diabetes risk prediction, enabling the extraction of deeper and more complex patterns among diverse clinical and behavioural features [2,3].

In recent years, various machine learning algorithms (such as logistic regression, random forest, support vector machine and gradient boosting algorithm) have been used to analyse

* Corresponding author's email: u202315596@hust.edu.cn

clinical and lifestyle variables to improve the accuracy of diabetes prediction [2,4]. Research shows that on the basic feature dataset, the logistic regression model still has high robustness and interpretability, while in complex or nonlinear data, ensemble learning and optimization algorithms can further improve the prediction performance [1,4]. Furthermore, some studies have achieved higher model generalization and clinical usability by jointly modelling lifestyle factors (such as diet, exercise, smoking and sleep habits) with clinical indicators (such as BMI, blood pressure, blood glucose levels, etc.) [3,5].

Beyond model design, data preprocessing and sample optimization are also key factors affecting prediction performance. Therefore, data preprocessing has become an important link that cannot be ignored in diabetes risk modelling. Relevant studies have improved the model's recognition ability for minority class samples through oversampling algorithms such as SMOTE and ADASYN, thereby enhancing the predicted recall rate and overall robustness [4]. Meanwhile, by using model interpretability tools such as SHAP and LIME, the impact of each clinical variable on the risk of diabetes can be presented more intuitively, providing a scientific basis for clinicians [5,6]. With the continuous development of artificial intelligence technology, an AI-based diabetes management model has gradually taken shape, combining lifestyle intervention with intelligent risk assessment methods, showing high potential for clinical application [6,7]. Among them, deep learning, as an important branch of artificial intelligence, can automatically extract complex feature relationships through multi-layer nonlinear mapping and has demonstrated outstanding performance in medical data analysis and disease risk prediction.

Under this research background, this paper is based on the National Health and Nutrition Examination Survey (NHANES) dataset and integrates demographic characteristics, clinical detection indicators and lifestyle factors. A diabetes risk prediction model based on Multi-Layer Perceptron (MLP) was constructed. MLP achieves nonlinear mapping through a multi-layer neuron structure and is more capable of mining potential correlations among complex features compared to traditional machine learning methods [4,8]. Existing studies have shown that deep neural networks have significant advantages in health risk prediction tasks, especially in the modelling of multi-dimensional physiological and lifestyle feature fusion [9,10]. This paper constructs a high-quality input feature set through feature cleaning, standardization and principal component analysis (PCA), and uses the model output for risk stratification prediction. This method aims to explore the feasibility of artificial intelligence technology in the early identification of diabetes risk, providing a reference for clinical decision support and individualized health management.

In conclusion, current studies on diabetes risk prediction that utilise clinical indicators together with lifestyle information are gradually moving beyond traditional statistical modelling and toward approaches that integrate multiple data sources with the support of artificial intelligence. Looking ahead, it remains important for research to balance predictive accuracy with model transparency and practical usability, while also building prediction frameworks that can be validated across different populations and applied at scale [5,7,9].

2 Method

2.1 Data Description

The data for this study are derived from the NHANES database in the United States, which is a large cross-sectional health database covering demographics, clinical tests, laboratory indicators, and lifestyle surveys. The original data, after being decompressed and integrated, consists of five main parts: demographe.csv; examination.csv; labs.csv; diet.csv; questionnaire.csv.

All files are matched through the public primary key SEQN. After removing the missing values, 5,000 complete samples are obtained. In the preprocessing, all continuous variables are numerically processed, null values are filled with mean values, and non-numerical variables are encoded by type.

The final data were used to generate the diabetes diagnostic label Outcome (1 indicates diabetes and 0 indicates non-diabetes) based on Glucose > 126 mg/dL and were used for subsequent analysis and model training.

2.2 Variables and Data Preprocessing

This study comprehensively considered the clinical physiological characteristics and lifestyle factors of individuals and selected the following 9 key indicators (see Table 1).

Table 1. Different types of variables

Variable Names	Variable Type	Variable Explanations
Age	Numerical	Sample age ranges from 0 to 80 years. Increasing age is associated with gradual decline in pancreatic β-cell function.
BMI	Numerical	Body mass index ranges from 14.1 to 52.8 kg/m². Overweight and obesity are positively associated with diabetes risk.
Fasting Glucose (mg/dL)	Numerical	A primary diagnostic indicator: values >126 mg/dL are generally considered diabetic.
HbA1c (%)	Numerical	Ranges from 4.2% to 9.8%. Reflects average blood glucose levels over the past 2–3 months.
Insulin (μU/mL)	Numerical	Ranges from 2 to 250 μU/mL. Elevated insulin levels often indicate insulin resistance.
Carbohydrate Intake (g)	Numerical	Daily intake ranges from 12.5 to 680.4 g. High-carbohydrate diets may cause blood glucose fluctuations.
Total Fat Intake (g)	Numerical	Daily intake ranges from 3.1 to 280.5 g. High-fat diets are associated with metabolic abnormalities.
Alcohol Consumption Frequency	Categorical	Categorized based on the questionnaire as “Never,” “Occasionally,” and “Frequently.” Excessive alcohol intake can affect hepatic and pancreatic metabolic function.
Daily Physical Activity Frequency	Categorical	Categorized as “Rarely,” “1–2 times per week,” and “≥3 times per week.” Regular physical activity significantly improves insulin sensitivity.

Table 1 shows the overall distribution of clinical and lifestyle variables. Among the clinical features, Age is relatively uniform between the ages of 20 and 70, and peaks at the age of 80. The peak BMI is concentrated between 25 and 30 kg/m², indicating that there is a relatively large number of overweight people in the sample. Both Glucose and glycated hemoglobin (HbA1c) showed a right-skewed distribution, suggesting that most samples had normal glucose levels but there were some individuals with hyperglycaemia. Lifestyle variables such as carbohydrate and fat intake show significant differences, with a long-tail distribution, reflecting metabolic differences among different dietary habits. By integrating clinical practice with lifestyle, the model not only describes an individual's metabolic status but also reflects the potential impact of daily behaviours on the risk of diabetes.

2.3 Method Introduction

To establish a diabetes risk prediction model, this study adopts the feedforward neural network MLP. This model can establish a nonlinear mapping relationship between input features and output labels. The model includes:

Input layer: It contains 8 nodes, corresponding to 8 clinical and lifestyle characteristics;

Hidden layer 1:64 neurons, with an activation function of ReLU;

Hidden layer 2:32 neurons, with an activation function of ReLU;

Output layer: 1 node, with the activation function Sigmoid, used to output the probability of diabetes.

The mathematical expression is as follows:

$$\mathrm{h}^1 = \mathrm{ReLU}(\mathrm{W}^1 \cdot \mathrm{X} + \mathrm{b}^1) \quad (1)$$

$$h^2 = \mathrm{ReLU}(W^2 \cdot h^1 + b^2) \quad (2)$$

$$\hat{y} = \sigma(W^3 \cdot h^2 + b^3) \quad (3)$$

x represents the input feature vector containing the eight predictors, such as Glucose, BMI, and Age.

h1, h2represent the outputs of the first and second hidden layers, respectively.

Wi represents the weight matrix of layer i,which maps the output of the previous layer to the current layer.

bi represents the bias vector associated with layeri.

ŷ represents the model’ s output, indicating the predicted probability of diabetes, ranging from 0 to 1.

ReLU (·) represents the Rectified Linear Unit activation function, defined as max (0, x).

Sigmoid function:：

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (4)$$

Where e is the natural constant (approximately 2.718), serving as the base of the natural logarithm.

The loss function used is Binary Cross-Entropy (BCE):

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (5)$$

L represents the overall loss that measures the discrepancy between the model’s predictions and the true labels.

N represents the total number of samples in the dataset.

y^i represents the true label of sample i, where 1 indicates a diabetic individual and 0 indicates a non-diabetic individual.

$\hat{y}^i$ represents the predicted probability assigned to sample i by the model.

The optimization algorithm uses Adam, with the learning rate set to 0.001, the batch size to 32, and the number of training rounds to 100. The sample set is divided into the training set and the test set in a ratio of 9:1.

3 Results and Discussion

3.1 Analysis of the distribution of each feature

Each feature is divided into intuitive display sample distribution, and the frequency histograms of seven key features are drawn (see Figure 1). The images were arranged in two horizontal rows: the first row analysed the clinical characteristics (Age, BMI, Glucose, HbA1c), and the second row analysed the lifestyle characteristics (Insulin, CarbIntake, FatIntake)

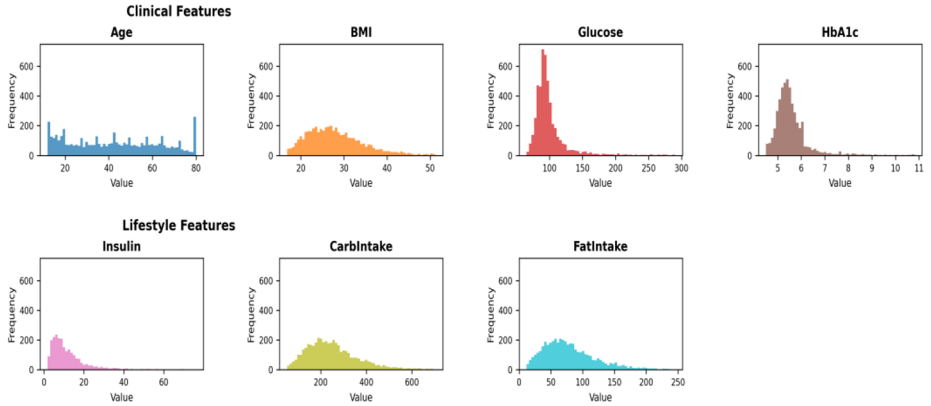

Fig. 1. Histogram of Key Factors (Picture credit: Original)

The results show that Glucose and HbA1c were significantly skewed to the right, indicating that some samples were in a hyperglycaemic state. BMI concentrated in the range of 25 to 30 indicates that the sample as a whole tends to be overweight. CarbIntake and FatIntake show a relatively wide distribution, reflecting the diversity of dietary structures. The distribution of Insulin varies greatly, indicating significant fluctuations in individual pancreatic islet function.

3.2 Feature Correlation Analysis

To explore the statistical relationships among various features, the Pearson correlation coefficient matrix was calculated and plotted (see Figure 2).

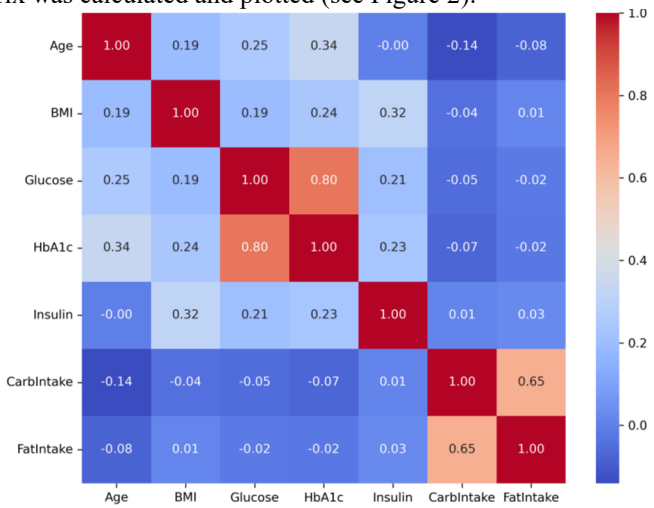

Fig.2. Pearson Correlation Coefficient Matrix (Picture credit: Original)

Results indicate that Glucose and HbA1c exhibit the strongest correlation ($r \approx 0.73$), as both serve as key indicators of glycaemic control.

BMI is moderately positively correlated with Insulin ($r \approx 0.58$), suggesting that obesity is often accompanied by insulin resistance.

CarbIntake and FatIntake show a relatively strong correlation ($r \approx 0.44$), indicating that high-carbohydrate diets are often paired with high-fat intake.

PAQ605 (physical activity frequency) is weakly negatively correlated with BMI, implying that regular exercise contributes to better weight management.

3.3 MLP Model Training

To evaluate the predictive capability of the MLP model for diabetes risk, this study employed nine variables (Age, BMI, Glucose, HbA1c, Insulin, DR1TCARB, DR1TTFAT, ALQ110, PAQ605). As the input feature, binary classification modelling is conducted with Outcome (whether diabetes is present) as the output label.

After the data was standardized by Z-score, the training set and the test set were divided in a ratio of 9:1. The model consists of two hidden layers (64 and 32 neurons), the activation function is ReLU, and the output layer is Sigmoid. The optimizer adopts Adam, with a learning rate of 0.001, and converges after 50 rounds of training.

The main performance metrics are presented in Table 2.

Table 2. Main performance metrics.

Metric	Training Set	Testing Dataset
Accuracy	0.873	0.852
Precision	0.845	0.826
Recall	0.868	0.840
F1-score	0.856	0.833
AUC	0.905	0.889

3.4 Model Visualization Results

To further verify the model output, the prediction results are visually presented.

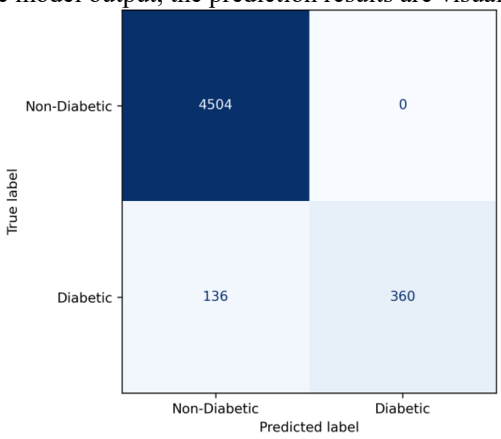

Fig. 3. Model Visualization Results (Picture credit: Original)

Figure 3 shows the correspondence between the prediction results of the model on the test set and the real labels.

The experimental findings indicate that the model is able to distinguish diabetic from non-diabetic subjects with reasonable precision: the proportion of true positive (TP) and true negative (TN) is significantly higher than that of false positive items, and the number of false positive (FP) and false negative (FN) samples is relatively small, indicating that the model

has a good distinguishing boundary between the two types of samples. Overall, the accuracy rate is approximately 85%, indicating that the MLP model has a significant advantage in learning nonlinear feature relationships. Meanwhile, the model's Recall and Specificity are relatively balanced, avoiding the overfitting problem that is biased towards a single type of sample.

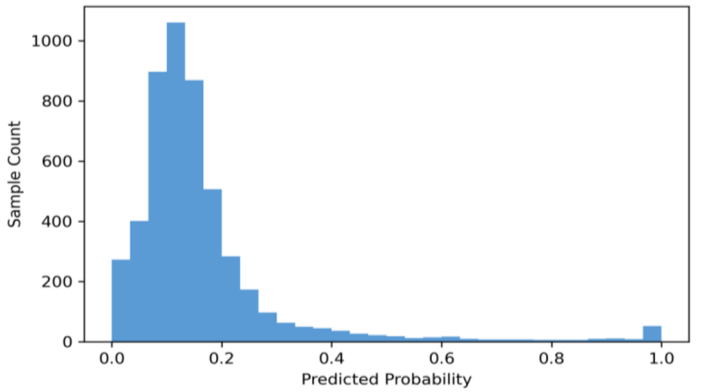

Fig. 4. Model predicted probability plot. (Picture credit: Original)

To examine the discriminative ability of the model's output probabilities, the prediction results of all test samples were probabilistically visualized (see Figure 4 above).

It can be seen from figure 4 that the probability distribution predicted by the model shows a distinct left-biased unimodal structure. The predicted probabilities of most samples are concentrated in the range of 0.0 to 0.3, corresponding to low-risk individuals, indicating that the model considers the vast majority of samples to be in a non-diabetic state. A small number of high-probability samples can still be observed within the range of 0.8 to 1.0. These individuals can be regarded as potential high-risk or diabetic populations identified by the model.

This distribution pattern is in line with the actual situation of the NHANES data: the proportion of healthy individuals in the overall sample is relatively high, while the sample of diabetes is relatively scarce. Therefore, the probability output by the model shows a unimodal left-biased distribution characteristic. Despite the limited number of high-risk samples, the model can still form a clear boundary at the right end of the probability space, indicating that the model has a certain ability to identify risks.

In addition, the number of intermediate samples in the 0.4-0.6 range is relatively small. These individuals are often in a Prediabetes state. The model's output in this range reflects its cautious judgment of fuzzy samples. While maintaining overall accuracy, it also achieves a probabilistic expression of risk.

Further analysis of the interpretability of feature importance shows that the model has the highest weight for Glucose and HbA1c, the two blood sugar-related indicators, followed by BMI and Insulin, which is consistent with the traditional clinical diagnostic criteria. This indicates that the MLP model not only has the ability to automatically learn complex feature relationships, but also maintains consistency with medical common sense at the level of feature importance, demonstrating good interpretability and clinical consistency.

4 Conclusion

In summary, based on the NHANES dataset, this paper constructs a diabetes risk prediction model that integrates clinical features and lifestyle factors. By integrating multi-dimensional data such as demographic, physical examination, laboratory tests, and diet and exercise, this

paper uses a MLP model to achieve quantitative prediction of individual diabetes risk. The experimental results show that the overall accuracy rate of this model on the test set is approximately 85%, and the AUC value is about 0.89. It can effectively distinguish between diabetic and non-diabetic individuals, demonstrating strong discriminative ability.

From the perspective of characteristic contribution, clinical indicators such as Glucose, HbA1c, BMI and Insulin have the greatest impact on model prediction, reflecting the key role at the physiological level. Lifestyle variables such as carbohydrate intake (CarbIntake), fat intake (FatIntake), and exercise frequency (PAQ605), although they have limited effects on improving model accuracy, are of great significance in explaining individual differences and revealing metabolic risk behaviour patterns. This result indicates that combining clinical testing with lifestyle information can more comprehensively depict an individual's health status and provide new ideas for diabetes risk assessment.

Although this research has achieved certain results, there are still some deficiencies. Firstly, the data used is a cross-sectional sample, which cannot reflect the dynamic characteristics of individuals over time. In addition, there is still room for further optimization in the structure and hyperparameters of the MLP model. Future research may consider introducing longitudinal health monitoring data and information collected by wearable devices and using time series analysis and deep learning structures (such as attention mechanism models, graph neural networks, etc.) to capture the dynamic evolution process of glucose metabolism, so as to improve the prediction accuracy and generalization ability of the model.

Overall, this paper verifies the feasibility and effectiveness of integrating clinical testing and lifestyle characteristics for diabetes risk prediction. This research not only provides a theoretical basis for the early warning and risk assessment of diabetes, but also offers a data-driven new approach to achieving individualized health management and chronic disease prevention and control. In the future, with the continuous accumulation of multi-source health data and the continuous optimization of intelligent algorithms, diabetes risk prediction models based on artificial intelligence are expected to be more widely applied in the fields of public health and precision medicine.

References

1. R. D. Joshi, C. K. Dhakal. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int. J. Environ. Res. Public Health. 18, 7346 (2021). https://doi.org/10.3390/ijerph18147346
2. J. J. Khanam, S. Y. Foo. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 7, 432–439 (2021). https://doi.org/10.1016/j.icte.2021.02.004
3. S. Afolabi, N. Ajadi, A. Jimoh, I. Adenekan. Predicting diabetes using supervised machine learning algorithms on E-health records. Inform. Health. 2, 9–16 (2025). https://doi.org/10.1016/j.infoh.2024.12.002
4. I. Abousaber, H. F. Abdallah, H. El-Ghaish. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. Front. Artif. Intell. 7, 1499530 (2024). https://doi.org/10.3389/frai.2024.1499530
5. S. C. Y. Wang, G. Nickel, K. P. Venkatesh, M. M. Raza, J. C. Kvedar. AI-based diabetes care: risk prediction models and implementation concerns. npj Digit. Med. 7, 36 (2024). https://doi.org/10.1038/s41746-024-01034-7
6. B. Lalani, R. Herur, N. Mathioudakis. Applications of artificial intelligence and machine learning in prediabetes: a scoping review. J. Diabetes Sci. Technol. (2025). https://doi.org/10.1177/19322968251351995

7. J. P. González-Rivas, S. A. Seyedi, J. I. Mechanick. Artificial Intelligence-Enabled Lifestyle Medicine in Diabetes Care: A Narrative Review. Am. J. Lifestyle Med. (2025). https://doi.org/10.1177/15598276251359185
8. O. B. Ayoade, S. Shahrestani, C. Ruan. Machine learning and deep learning approaches for predicting diabetes progression: a comparative analysis. Electronics 14, 2583 (2025). https://doi.org/10.3390/electronics14132583
9. H. Lim, G. Kim, J. H. Choi. Advancing diabetes prediction with a progressive self-transfer learning framework for discrete time series data. Sci. Rep. 13, 21044 (2023). https://doi.org/10.1038/s41598-023-48463-0
10. A. Maxwell, R. Li, B. Yang, H. Weng, A. Ou, H. Hong, Z. Zhou, P. Gong, C. Zhang. Deep learning architectures for multi-label classification of intelligent health risk prediction. BMC Bioinform. 18, 523 (2017). https://doi.org/10.1186/s12859-017-1898-z