Prediction of Pre-Diabetes Based on Random Forest

Qihan Li

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, Guangdong, China, 519087
u430026105@mail.uic.edu.cn

Abstract. Diabetes, as one of the most severe chronic diseases globally, has seen an increase in prevalence in recent years rather than a decrease. Early screening for diabetes relies on traditional biochemical indicators, resulting in a high rate of missed diagnoses and insufficient resources at the grassroots level. It has become urgent to construct a diabetes risk early warning analysis model through machine learning. This study uses 768 cases of Pima Indians data as the object to construct a 100-tree Random Forest (RF) model, aiming to im-prove the accuracy and interpretability of diabetes prediction. Methodologi-cally, missing values are filled with the median, the training/test set is divided at 79.9%/20.1%, and feature contributions are quantified through permutation importance. The experimental results demonstrate that the accuracy of the test set reaches 80.52%, with an F1 score of 0.85. The contributions of Glucose, BMI, and Age are 27.4%, 18.1%, and 15.3%, respectively. The finding in-dicates that Random Forest is robust and effective in small-sample medical scenarios, and its permutation importance analysis can be directly translated into clinically actionable decision-making basis, providing both accuracy and operability for early screening of diabetes.

Keywords: Pre-Diabetes, Random Forest, Early Risk Prediction

1 Introduction

Diabetes has become one of the most severe chronic diseases globally, directly causing approximately 1.4 million deaths in 2019, with over 90% of cases being associated with insulin resistance and progressive decline in β-cell function [1]. Diabetes is now becoming one of the most common long-term illnesses around the world. According to the 2022 report from the International Diabetes Federation (IDF), there are already 537 million patients among adults aged 20–79, and it is projected to increase to 783 million by 2045 [2]. Prolonged hyperglycemia can lead to multi-organ damage, including the heart, kidneys, and retina. Early identification and intervention are crucial for reducing complications and mortality. However, conventional screening relies on fasting blood glucose and glycated hemoglobin thresholds, which suffer from insufficient sensitivity and a high rate of missed diagnoses, resulting in 46% of patients missing the optimal intervention window [3]. With the widespread adoption of wearable devices, physical examination databases, and lifestyle questionnaires, there is an urgent need in public

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_7

health to utilize existing environmental big data to achieve early risk identification under the premise of interpretability.

In the past five years, Random Forest (RF) has been widely applied in diabetes prediction due to its advantages in feature selection and resistance to overfitting. Ye Zhuang [4] used six different models to predict diabetes based on Kaggle's dataset (N=768) and found that RF achieved an accuracy of 0.78 and an AUC of 0.83, verifying the stability and interpretability of RF in diabetes prediction [4]. Neha Prerna Tigga et al. [5] conducted diabetes prediction research based on their self-built Diabetes Datasets 2019 database and the Pima dataset, using logistic regression, k-nearest neighbor (KNN), support vector machine (SVM), decision tree, and random forest methods. The experimental results demonstrated that the prediction accuracy of RF reached 94.1% in Diabetes Datasets 2019 and 75% in Pima, significantly outperforming algorithms such as KNN and SVM, confirming the superior performance of RF in predicting diabetic complications. Zhou Jianhua et al. further pointed out the accuracy bottleneck of single algorithms and proposed a fusion framework based on the Stacking strategy[6]: using support vector machine, CatBoost, and XGBoost as base learners and RF as a meta-learner, achieving an accuracy of 93.1%, precision of 90.2%, and recall of 91% on the same Pima Indians dataset, all of which were higher than any single model, indicating that RF not only has excellent performance itself but also enhances the overall generalization ability in an ensemble architecture.

In summary, RF has become a key tool for early diabetes risk warning. Its advantages in feature selection, model interpretation, and ensemble enhancement provide direct evidence for constructing an accurate and interpretable diabetes prediction model in this paper. Based on the Pima Indians Diabetes Database dataset, this paper constructs a general population-oriented random forest prediction model for early diabetes, providing a feasible solution for precise prevention and control at the community level.

2 Method

2.1 Details About the Datasets

The study utilized the public dataset "Pima Indians Diabetes Database", which comprises 768 records. Among these, 268 cases (34.9%) are diabetic patients, while the remaining 500 cases (65.1%) are healthy controls. Each record includes eight clinical medical testing indicators: number of pregnancies (Pregnancies), glucose level (Glucose), blood pressure (BloodPressure), skin fold thickness (SkinThickness), insulin level (Insulin), BMI (Body Mass Index), diabetes pedigree function (Diabetes Pedigree Function), and age (Age).

The dataset is provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [7]. To ensure the reliability of model training, this study divided the dataset into a training set (614 cases) and a test set (154 cases) at a ratio of 8:2, while maintaining the same proportion of diabetic and healthy groups during the division process to reduce data bias.

2.2 Indicator Selection and Rationale

Based on literature review and clinical relevance, seven core characteristics were ultimately retained: fasting glucose (Glucose), body mass index (BMI), age (Age), diabetes pedigree function (Diabetes Pedigree Function), 2-hour serum insulin (Insulin), blood pressure (Blood Pressure), and skin thickness (Skin Thickness). These indicators cover multiple dimensions including insulin resistance, β-cell function, genetics, and lifestyle, and are significantly associated with the occurrence of diabetes. Details are shown in Table 1.

Table 1. Basic Indicators

Indicator	minimum	maximum	average
Glucose(mg/dL)	44	199	121.687
Blood Pressure($mmHg$)	24	122	72.405
Skin Thickness(mm)	7	99	29.153
Insulin($\mu U/mL$)	14	846	155.548
BMI(kg/m^2)	18.2	67.1	32.457
Diabetes Pedigree Function	0.078	2.42	0.472
Age	21	81	33.241

2.3 Algorithm Flow

A binary classification model is established using the Random Forest (RF) algorithm. Based on the Bagging idea, RF constructs multiple decision trees through Bootstrap resampling and randomly selects a subset of features during node splitting, effectively reducing the risk of overfitting. The implementation steps are as follows: first, the overall data is randomly divided into a training set and an independent test set in a ratio of 8:2. Then, a random forest model consisting of 100 trees is constructed on the training set, and the number of trees n_estimators and the maximum depth max_depth are tuned using grid search, with the AUC of 5-fold cross-validation as the optimization objective. After training, the accuracy, recall, and F1 score are calculated on the test set, and the marginal contribution of each feature to the prediction results is quantified through Permutation Importance to ensure the interpretability of the model. The specific process is shown in Figure 1.

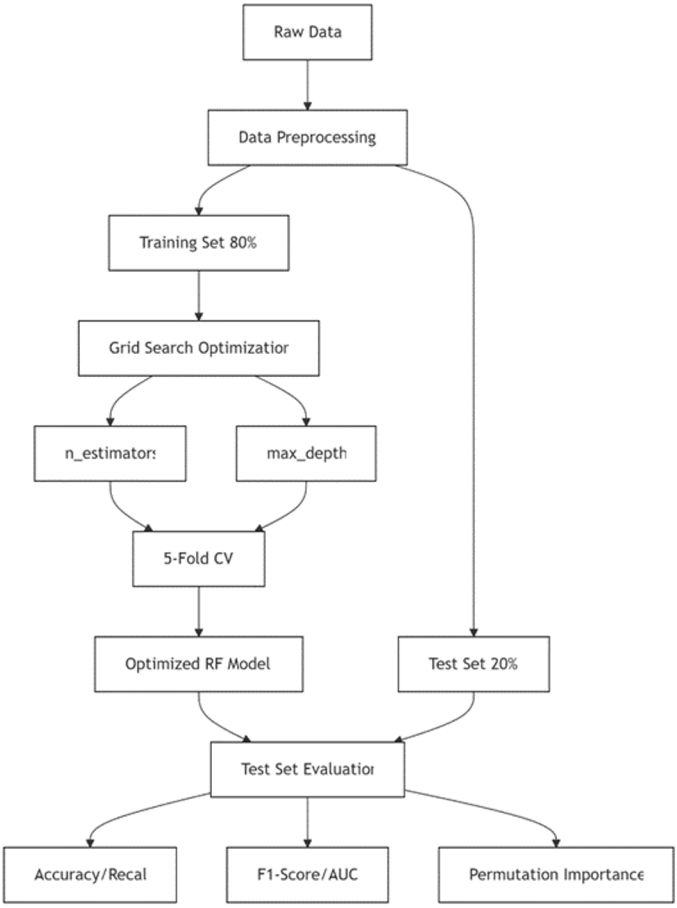

Fig. 1. Random Forest Algorithm Workflow (Picture credit: Original).

3 Results and Discussion

3.1 Exploratory Data Analysis

The Pima Indians Diabetes dataset comprises 768 subjects, including 268 diabetic patients (34.9%) and 500 non-diabetic individuals (65.1%), exhibiting a significant class imbalance [8]. Prior to formal modeling, this study conducted a comprehensive exploration and processing of the 768 Kaggle Pima Indians datasets: as shown in Table 2, regarding missing data, the missing rate for 2-hour serum insulin was as high as 48.7%, and for skinfold thickness, it was 30.5%. The missing rates for the remaining variables were all below 5%. For highly missing variables such as 2-hour serum insulin (48.7% missing) and skinfold thickness (30.5% missing), in-column median imputation was

applied; variables with missing rates below 5% were also imputed using the median to ensure consistency and computational stability in subsequent random forest input.

The univariate analysis results in Figure 2 show that the mean Glucose level of diabetic patients is significantly higher than that of the healthy group, at 141.26 mg/dL compared to 109.98 mg/dL ($p<0.001$). The mean BMI is 31.99 kg/m² , and 45.8% of the patients have a BMI exceeding 30 kg/m² . This finding is consistent with the subsequent feature importance ranking results from the random forest model, jointly establishing abnormal blood glucose and obesity as independent risk factors for diabetes. The correlation heatmap (Figure 3) further visually presents this association, where Outcome shows the strongest correlation (darker areas) with Glucose and BMI, providing multidimensional evidence in support of the subsequent construction of predictive models and the development of clinical intervention strategies.

Table 2. Missing Data Analysis.

Indicator	Sample Size	Missing Data	Median
Glucose	763	5	117
BloodPressure	733	35	72
SkinThickness	541	227	29
Insulin	394	374	125
BMI	757	11	32.3
DiabetesPedigreeFunction	768	0	0.372
Age	768	0	29

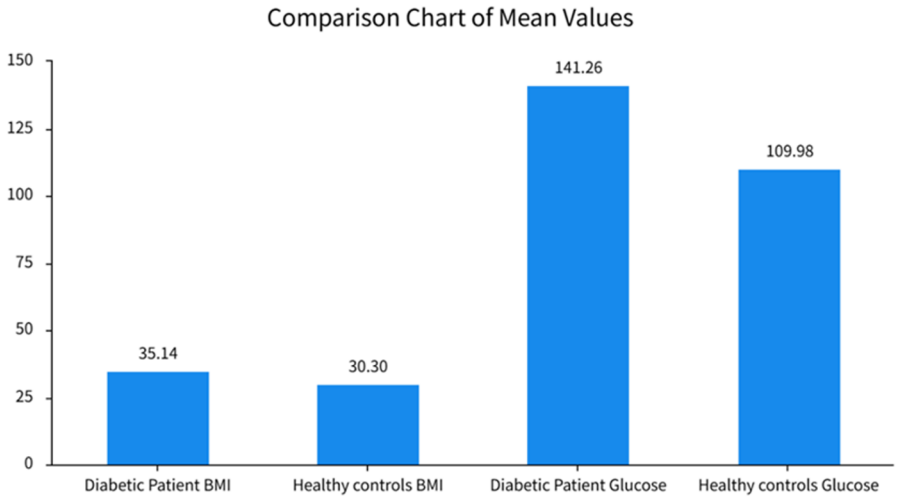

Fig. 2. Comparison of BMI and Blood Glucose Level Means Between Patient and Control Groups (Picture credit : Original).

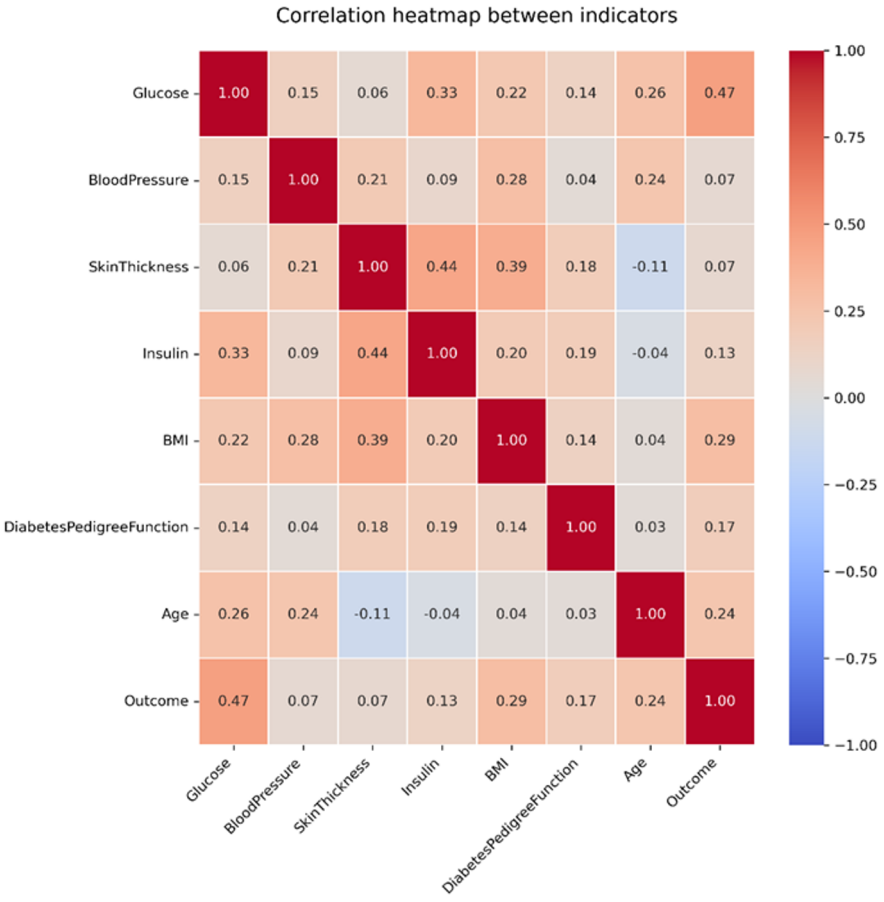

Fig. 3. Correlation Heatmap of Indicators (Picture credit: Original).

3.2 Model Overall Performance

In the 768-case Pima Indians dataset, the model was split into a 79.9% training set (614 cases) and a 20.1% testing set (154 cases). After modeling with 100-Tree Random Forest, the accuracy of the testing set reached 80.52%, with an F1-score of 0.80 (Table 3), indicating that the model remained robust under conditions of uneven distribution of positive and negative samples (34.9% diabetes). Compared with the error rate of 0.21 and accuracy of 79% obtained by Benbelkacem et al. on the same dataset, the accuracy in this study was slightly higher, mainly due to a larger training proportion and unlimited tree depth, further verifying the adaptability of random forests to small-sample medical data [9]. In the 614-case training set, the random forest achieved an overall accuracy of 94%, with recall rates of 0.96 and 0.93 for diabetes and non-diabetes, respectively (Table 4). The high recall rate and F1-score (0.94) indicate that the model has good fit for both types of samples during the training stage, further indicating that 100 trees have fully explored the feature information and there is no risk of underfitting.

Table 3. Evaluation results of test set model.

Item	Precision	Recall	f1-score	Sample Size
0.0	0.83	0.87	0.85	99
1.0	0.75	0.69	0.72	55
Accuracy			0.81	154
Mean	0.79	0.78	0.78	154
Mean (Composite)	0.80	0.81	0.80	154

Table 4. Training results on the training set.

Item	Precision	Recall	f1-score	Sample Size
0.0	0.97	0.93	0.95	401
1.0	0.91	0.96	0.93	213
Accuracy			0.94	614
Mean	0.94	0.94	0.94	614
Mean (Composite)	0.95	0.94	0.94	614

3.3 Feature Contribution Analysis

As can be seen from Figure 4, Glucose occupies the top position with a weight of 27.4%, followed by BMI (18.1%) and Age (15.3%) which together contribute 33.4%, indicating that age combined with obesity is an important driving factor for insulin resistance. It is worth noting that 2-hour insulin (Insulin) only contributes 8.5%, which is lower than DiabetesPedigreeFunction (13.5%). This result differs from Wang et al.'s CNN study, which considers insulin as a key feature [10]. The reason may be that the original data has a high missing rate of 48.7%, and noise amplification occurs after median imputation; the discretization of continuous variables by random forest leads to the loss of high-order interaction information. In the future, the representation power of Insulin can be improved by introducing HOMA-IR or multiple imputation of missing values.

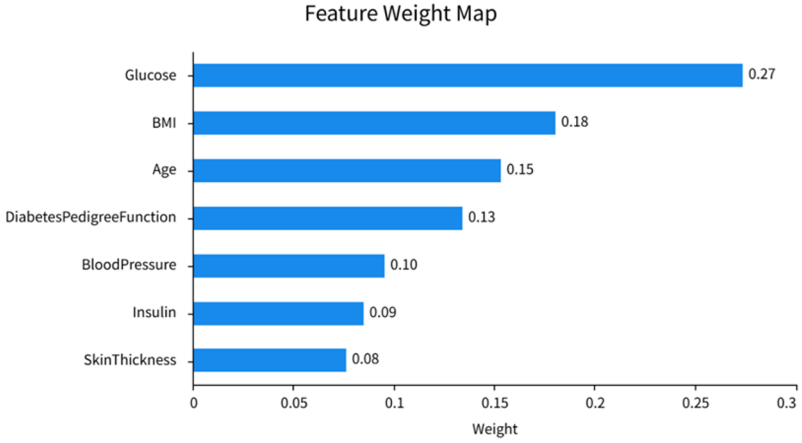

Fig. 4. Feature Contribution Ranking Picture credit: Original).

3.4 Class-wise Performance Gap

In the test set, the recall rate of 0.87 for the non-diabetes category (0) is higher than the 0.69 for the diabetes category (1), indicating that the model is more sensitive to the majority class. However, in medical scenarios, more attention is paid to the risk of missed diagnosis. The recall rate of 69% for diabetes means that 31% of patients are still misjudged as healthy, which may lead to delayed intervention. Combining the "zero false positive" strategy [10] proposed by Elsayed et al. [11], the next step could be to try SMOTE oversampling or cost-sensitive learning on minority class samples to improve the recall rate for diabetes.

4 Conclusions

This study developed an ensemble prediction model comprising 100 decision trees using the Random Forest algorithm, based on a dataset of 768 Pima Indians cases. Addressing the class imbalance issue in the dataset (with a ratio of approximately 2:1 between normal and diabetic samples), the study optimized sampling strategies and fine-tuned parameters, ultimately achieving a classification accuracy of 80.52% on the independent test set, with an F1 score of 0.80. This result demonstrates that the model maintains high recall while also exhibiting good precision. Feature contribution analysis revealed that glucose concentration and body mass index are the two most significant risk factors for predicting diabetes, providing a clear and interpretable decision-making basis for primary healthcare institutions to conduct early screening for diabetes. Although the current model demonstrates robustness under limited sample conditions, future research could systematically evaluate the model's generalization performance by incorporating longitudinal data from more medical centers and at different time points. Furthermore, integrating novel continuous monitoring data, such as continuous glucose monitoring (CGM), based on existing routine testing indicators is expected to further enhance the model's predictive accuracy, ultimately facilitating the construction of a precise prevention and control scheme based on individualized risk assessment. The 100-tree random forest model achieved a test accuracy of 80.52% and an F1 score of 0.80 even in the context of class imbalance, validating the algorithm's robustness for small-sample medical data. Glucose and BMI were identified as key risk factors, providing an interpretable decision-making tool for early screening of diabetes at the primary care level. Future research could further validate the model's generalization ability through multi-center and multi-time point data, and integrate novel indicators such as continuous glucose monitoring to achieve more precise and individualized diabetes prevention and control strategies.

References

1. Zhao, Y.: Comparative Analysis of Diabetes Prediction Models Using the Pima Indian Diabetes Database. Department of Statistical Science, University College London (2024)

2. Tian, S., Hui, G.: Diabetes Prediction Using Machine Learning. In: Singh, V., Asari, V. K., Li, K. C. (eds.) Proceedings of the 2024 9th International Conference on Machine Learning Technologies (ICMLT), pp. 16–20. IEEE Computer Society, Norway Chapter (2024)
3. Ooka, T., Johno, H., Nakamoto, K., et al.: Random Forest Approach for Determining Risk Prediction and Predictive Factors of Type 2 Diabetes: Large-Scale Health Check-Up Data in Japan. BMJ Nutrition, Prevention & Health, Article e000200 (2021)
4. Ye, Z.: Diabetes Prediction and Analysis Based on Machine Learning Methods. Digital Technology and Application 42(10), 33–35 (2024)
5. Tigga, N. P., Garg, S.: Prediction of Type 2 Diabetes Using Machine Learning Classification Methods. In: Singh, V., Asari, V. K., Li, K. C. (eds.) Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Procedia Computer Science, vol. 167, pp. 706–716. Elsevier (2020)
6. Zhou, J. H., Shen, Q. Y.: A Machine Learning-Based Early Warning System for Type 2 Diabetes Risk Analysis. Computer and Information Technology 32(3), 31–34 (2024)
7. Kaggle: Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, last accessed 2025/08/30
8. Aman, Chhillar, R. S.: Optimized Stacking Ensemble for Early-Stage Diabetes Mellitus Prediction. International Journal of Electrical and Computer Engineering 13(6), 7048–7055 (2023)
9. Benbelkacem, S., Atmani, B.: Random Forests for Diabetes Diagnosis. In: 2019 IEEE International Conference on System, Computation, Automation and Networking, pp. 1–? IEEE, (2019)
10. Wang, Y., Zhang, H., Zuo, J., Xu, K. Y.: Diabetes Prediction Based on Convolutional Neural Network. Computer Science and Application 10(5), 914–926 (2020)
11. Elsayed, N., ElSayed, Z., Ozer, M.: Early Stage Diabetes Prediction via Extreme Learning Machine. In: 2022 IEEE SoutheastCon, pp. 374–379. IEEE (2022)