Proceedings of the 4th International Conference on Functional Materials and Civil Engineering
DOI: 10.54254/2755-2721/2026.38393

Large-Scale Population Screening for Type 2 Diadetes Based on TabNet and BRFSS 2021 Data

Jie Gao

College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot, China
20241101898@mails.imnu.edu.cn

Abstract. Type 2 diabetes is a global health crisis. Early screening is of great significance for disease prevention and control. Based on the data from the CDC BRFSS 2021 large-scale real population survey (236378 cases, 21 characteristics), this study systematically evaluated the performance of TabNet, XGBoost and random forest in diabetes screening. 50% cross validation results showed that the recall rate of TabNet reached 77.81% ± 2.02%, which was significantly better than XGBoost (35.92% ± 0.51%) and random forest (32.64% ±0.39%), and the AUC reached 0.8043±0.0031. The age subgroup analysis showed that among the high-risk population aged 30-50, the TabNet recall rate was as high as 80.88%, 51.93% higher than XGBoost. SHAP interpretability analysis showed that self-rated health, hypertension and age were the core characteristics for predicting diabetes. This result shows that TabNet's attention mechanism can effectively capture complex feature interactions and significantly improve the recognition ability of patients on unbalanced data. This study shows that TabNet has significant advantages of high recall rate in large-scale real population diabetes screening, and has important clinical application value.

Keywords: Machine learning, TabNet, Type 2 diabetes, SHAP

1. Introduction

Type 2 diabetes is a global health crisis. It is estimated that the number of patients with diabetes in the world will be about 537million in 2021, and it is expected to reach 783million by 2045 [1]. The disease is usually asymptomatic in the early stage, and irreversible complications such as cardiovascular, renal and retinal complications often occur in the diagnosis. It is of great significance to develop effective early risk prediction tools [2].

In recent years, machine learning methods have been widely used in diabetes risk prediction. Roobini and Lakshmi use AdaBoost and other algorithms to achieve a classification accuracy of 83% on the Pima dataset [3]. Xu et al. Used random forest model to predict type 2 diabetes based on age, waist circumference and other easily available indicators, with an accuracy rate of 85% [4]. Sajid et al. Fused random forest, support vector machine and XGBoost to build a hybrid ensemble paradigm for diabetes prediction [5].

The above research verifies the effectiveness of traditional machine learning methods. However, the accuracy rate can not fully reflect the value of screening tools. In the screening scenario, the cost of missed diagnosis is much higher than that of misdiagnosis, and the recall rate is the more critical

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38393

indicator. When dealing with the data of patients who are far less than those who are not, the traditional model has limited recognition ability for patients and low recall rate, which is difficult to meet the needs of screening.

TabNet, as a deep neural network architecture designed specifically for tabular data, provides a new technical path for improving the screening recall rate [6]. The instance level feature selection is realized through the attention mechanism, and TabNet can dynamically identify the most critical features for each sample; Its multi-step decision-making architecture simulates the layer by layer decision-making process of the decision tree, while retaining the end-to-end learning ability of the neural network. Previous studies have preliminarily verified the potential of TabNet in the prediction of diabetes: Usha et al. Integrated SCAW net with TabNet and achieved an accuracy of 98.9% on multiple data sets [7]; Chokshi's research on 100000 health records shows that TabNet has an accuracy of 91.84% and an au-roc of 0.962, which is superior to eight comparative models such as XGBoost and CatBoost [8]; Narvaez et al. Verified the value of lifestyle factors in diabetes prediction based on BRFSS 2015 data set [9]. The above research has preliminarily verified the effectiveness of TabNet, but it is mostly based on small-scale data sets such as Pima (768 cases). The performance of TabNet in the recall rate of hundreds of thousands of real population data is still lack of systematic evaluation, and its actual effect in the screening scene still needs to be verified. The performance difference of the model in different age groups is also lack of analysis.

In view of the above gaps, this study systematically evaluated the performance of TabNet in large-scale real population diabetes screening based on the survey data (236378 cases) of CDC behavioral risk monitoring system in the United States in 2021 [10]. The random forest was used as the baseline model and compared with XGBoost. This paper uses the recall rate as the core index to verify whether TabNet maintains the advantage of high recall rate on large-scale data, evaluates the stability of the model in different populations through age subgroup analysis, and reveals the core features of the prediction with the help of SHAP analysis, so as to enhance the clinical interpretability of the model [11].

2. Method

2.1. Data source and variable description

This study uses the survey data of the US CDC behavioral risk factor surveillance system (BRFSS) in 2021. The dataset is one of the largest health survey databases in the United States and has been used by many academic studies.

The respondents were non institutionalized American adult civilians (aged≥18 years). In 2021, 438693 interviews were completed. After data cleaning, 236378 complete records were included in this study. Among them, 33568 (14.2%) patients with diabetes (including prediabetes) and 202810 (85.8%) patients without diabetes. The data set contains 21 predictive characteristics, mainly including hypertension, high cholesterol BMI, Smoking status, history of heart disease, physical activity, self-rated health, age, income, etc.

2.2. Data preprocessing

2.2.1. Unified data type

Some binary features in the original data are read as floating-point types and are uniformly converted to integer types to ensure the consistency of the model input format.

DOI: 10.54254/2755-2721/2026.38393

2.2.2. Standardization

In order to avoid the influence of feature dimensional difference on model training, all features are standardized by Z-score, and the formula is $x' = (x - \mu)/\sigma$, where μ is the mean value of the training set and σ is the standard deviation of the training set. Standardization calculates parameters on the training set of each cross validation, and then applies them to the corresponding validation set.

2.2.3. Category imbalance treatment

The positive samples in the data set accounted for only 14.2%, and there was an obvious category imbalance problem. In order to alleviate the impact of imbalance on model training, smote (synthetic minority over sampling technique) oversampling technique is used to balance each training set, so that the proportion of positive and negative samples can reach 1:1.

2.3. Model introduction

In this study, random forest was selected as the baseline model, and XGBoost and TabNet were introduced for performance comparison.

Random forest is an algorithm based on bagging integration, which constructs multiple decision trees through bootstrap sampling, and finally gets the prediction results through voting. Its core features include the ability to process high-dimensional data, good anti over fitting ability, and the ability to provide feature importance assessment. The number of trees in this study is set to 100, and other parameters remain default.

XGBoost is an integrated algorithm based on gradient lifting framework, which is used to fit residuals by iteratively adding decision trees. Its core features include introducing regularization term to control the complexity of the model, supporting parallel computing, and automatically processing missing values. This study set the learning rate as 0.1, the maximum tree depth as 6, and the number of trees as 100.

TabNet is a deep neural network architecture for tabular data proposed by Google research. Its core innovation is to use attention mechanism to realize instance level feature selection: each sample can have different feature importance weights. TabNet's multi-step decision-making architecture simulates the layer by layer decision-making process of the decision tree, while retaining the end-to-end learning ability of the neural network. This study sets n _d=n_a=64, n_steps=5, gamma=1.5, using Adam optimizer, the learning rate is 2e-2.

To ensure the fairness of the comparison, all models are trained on the same training set and evaluated on the same test set, using the same 50% cross validation and smote parameters. Once the superparameter is determined, it remains unchanged in all turns.

2.4. Evaluation method

In this paper, we use hierarchical 50% cross validation to ensure that the proportion of positive and negative samples in each compromise is consistent with the original data, and finally report the mean and standard deviation of the results for five times.

The samples were divided into three groups according to age:<30 years old, 30-50 years old and ≥ 50 years old. Within each age group, the training set and test set were divided by 70/30 ratio, and the recall rate of each model was calculated.

DOI: 10.54254/2755-2721/2026.38393

2.5. Evaluation index

This study takes recall rate as the core evaluation index, supplemented by AUC to reflect the discrimination ability of the model. In the screening scenario, the recall rate, that is, the ability to find patients, is more important than the accuracy rate - false positives can be eliminated by subsequent examination, while missed diagnosis means that patients miss the best opportunity for intervention.

2.6. Interpretability analysis

The interpretability of XGBoost model is analyzed by using the SHAP method [11], the marginal contribution of each feature to the model prediction is calculated, and the global feature importance ranking is provided.

3. Result

3.1. 50% discount cross validation

Table 1 shows the recall rate and AUC performance of the three models under the 50% discount cross validation. TabNet had the highest recall rate (77.81%), which was significantly better than XGBoost (35.92%) and random forest (32.64%), with a gap of 41.89% and 45.17% respectively. In terms of AUC, XGBoost was the highest (0.8145), TabNet was the second (0.8043), and random forest was the lowest (0.7877).

From the perspective of standard deviation, the standard deviation of the recall rate of TabNet (± 2.02%) is greater than XGBoost (± 0.51%), indicating that the performance of TabNet fluctuates slightly under different data divisions, but it remains at a high level as a whole.

Table 1. Performance comparison of each model (50% discount cross validation)

model	recall (%)	AUC
Random forest	32.64 ± 0.39	0.7877 ± 0.0019
XGBoost	35.92 ± 0.51	0.8145 ± 0.0024
TabNet	77.81 ± 2.02	0.8043 ± 0.0031

3.2. Age subgroup analysis

Table 2 shows the recall rate performance of the three models in different age groups. In all age groups, the recall rate of TabNet was significantly better than XGBoost and random forest. In the 30-50 age group, the TabNet recall rate reached 80.88%, 51.93% higher than XGBoost (28.95%), which was the most significant age group. In the≥50 year old group, the TabNet recall rate was 70.97%, 39.17% higher than XGBoost (31.80%). In the young group under 30 years old, due to the very low positive sample rate (1.96%), the recall rate of all models was low, but TabNet (30.43%) was still better than other models.

TabNet is the most prominent in the 30-50 age group. This age group is experiencing accelerated accumulation of lifestyle related risk factors (such as irregular diet, reduced physical activity, work pressure, etc.), and the correlation between these factors and blood glucose metabolism has strong nonlinear characteristics. TabNet's attention mechanism can dynamically capture the most critical

DOI: 10.54254/2755-2721/2026.38393

feature combinations in each sample. Compared with XGBoost and random forest's fixed feature segmentation method, TabNet has natural advantages in modeling such complex interactions. In addition, the difference in the recall rate of TabNet between the 30-50 year old group and the≥50 year old group (80.88% vs 70.97%) may reflect the change in the importance of age-related characteristics: age itself has become a strong one-way risk factor in the≥50 year old group, and the traditional tree model can better identify it through a simple threshold; The occurrence of diabetes in the 30-50 age group depends more on the complex interaction of multiple factors, which is exactly the advantage of TabNet over the traditional tree model. In the<30-year-old group, the recall rate of all models was low. The fundamental reason was that the prevalence rate in this age group was very low (only 1.96%), and even TabNet was difficult to effectively learn the pattern of minority samples under extremely unbalanced data conditions, suggesting that special sampling strategies may be needed for diabetes screening in young people.

Table 2. Comparison of model recall rates in different age groups

model	<30 years	30-50 years	≥50 years
Random forest	1.09%	25.57%	31.01%
XGBoost	6.52%	28.95%	31.80%
TabNet	30.43%	80.88%	70.97%

3.3. SHAP interpretability analysis

3.3.1. Feature importance analysis

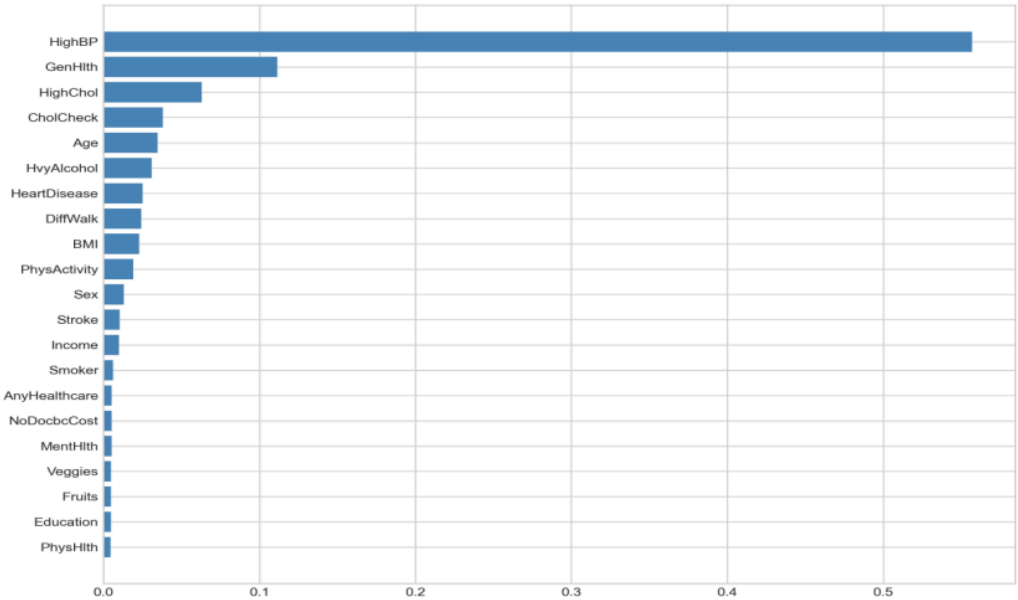

Figure 1. XGBoost feature importance feature ranking (picture credit:original)

Figure 1 shows the feature importance ranking of the XGBoost model. Hypertension (highbp) is the most important predictive feature, with an importance value of 0.52, which is much higher than other features. Followed by self-rated health (genhlth, 0.12) and high cholesterol (highchol, 0.08), age (age, 0.03) ranked fifth, BMI (0.01) ranked ninth. This ranking reflects the status of

DOI: 10.54254/2755-2721/2026.38393

hypertension as the strongest related factor of diabetes in this data set, which is consistent with the clinical epidemiological cognition; Self rated health, age and BMI were also verified as the core risk factors of diabetes.

3.3.2. SHAP profile

Figure 2 shows the influence direction and intensity of each feature on the model prediction. Key findings include: first, higher self-rated health scores (i.e., poor health status) are associated with higher diabetes risk; Second, patients with hypertension have a higher risk of diabetes; Third, age is positively correlated with the risk of diabetes.

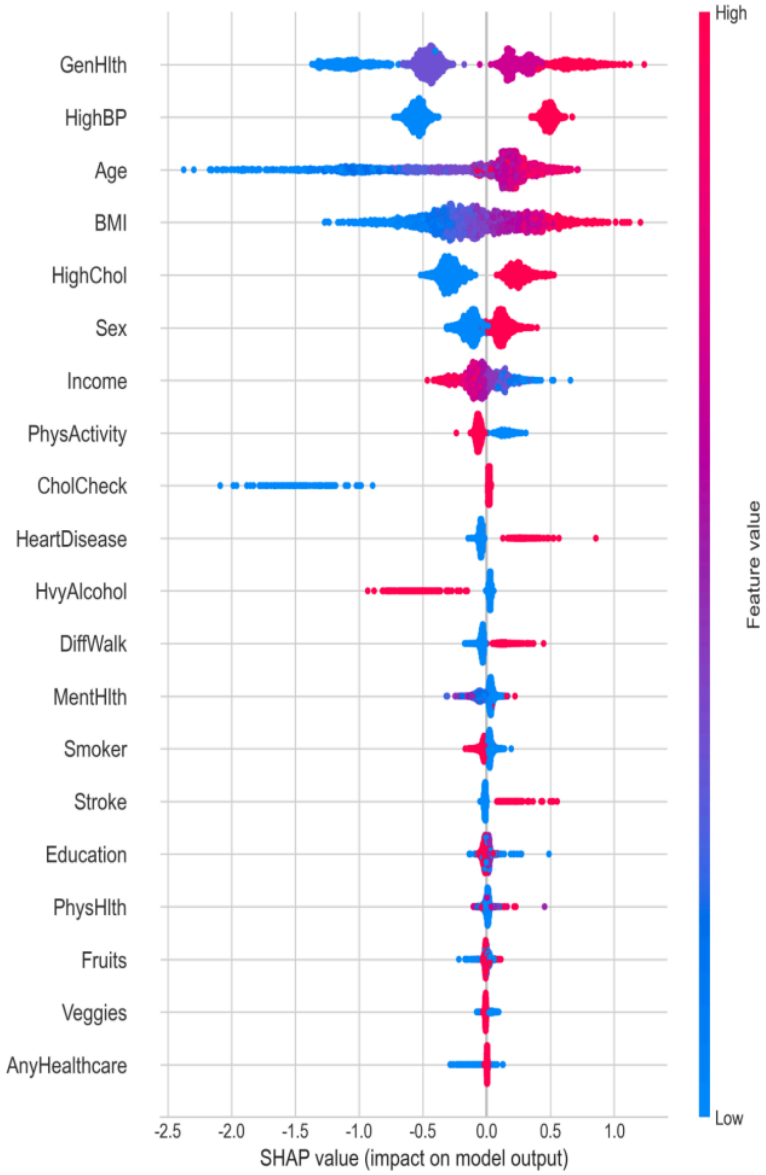

Figure 2. Summary diagram of XGBoost model SHAP (picture credit : original)

DOI: 10.54254/2755-2721/2026.38393

3.3.3. Age dependent analysis

Figure 3 shows the marginal effect of age characteristics on model prediction. With the increase of age, the SHAP value showed an upward trend, indicating that there was a stable monotonous positive correlation between age and diabetes risk, which verified the clinical consensus that age was the core risk factor of diabetes.

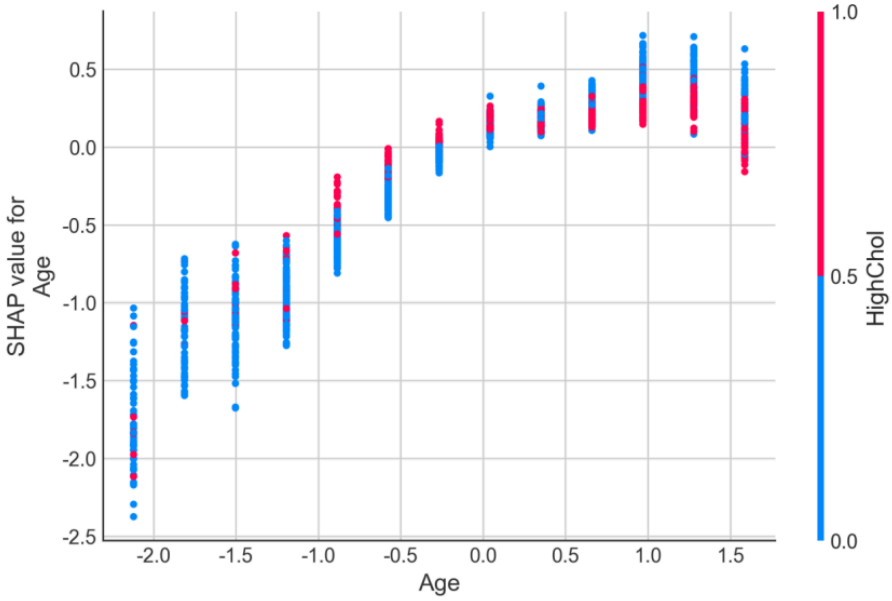

Figure 3. SHAP dependence graph of age (picture credit : original)

4. Discussion

4.1. Main findings and clinical significance

The core finding of this study is that TabNet has a significant advantage of high recall rate (77.81%) in diabetes screening, 41.89 percentage points higher than XGBoost. Among the high-risk population aged 30-50, the recall rate of TabNet is as high as 80.88%.

These findings have significant clinical implications. In the diabetes screening scenario, recall rate (that is, the ability to find real patients) is more important than accuracy rate. The high recall feature of TabNet makes it particularly suitable as a preliminary screening tool for large-scale populations. Compared with XGBoost, TabNet can find about 419 more diabetes patients per 1000 people screened, which has direct value for early intervention and disease prevention and control.

4.2. Comparison with existing studies

Compared with the study using Pima data set (768 cases), this study uses BRFSS 2021 large-scale real population data (236000 cases), and the research results are more universal and statistically effective. Different from the research with accuracy as the core index, this study focuses on screening scenarios and takes recall rate as the core evaluation index, which is more suitable for clinical application needs.

DOI: 10.54254/2755-2721/2026.38393

4.3. Limitations

This study has the following limitations. First, BRFSS data mainly reflect the health status of the American population, and the conclusion should be carefully verified when it is extended to other countries. Second, the training time of TabNet is significantly longer than that of the traditional tree model (about 5-10 times), and the computing resource limitation should be considered in the actual deployment. Third, the sample size of young people under 30 years old is small (15678 cases), and the results need to be further verified.

4.4. Future outlook

Future research can be conducted in the following directions: first, collect multi-source data (such as Chinese population data) for external verification; The second is to combine the questionnaire survey data with multimodal data such as electronic health records and genome data; The third is to carry out a prospective cohort study to verify the screening effect of the model in the actual clinical scene; Fourth, explore model lightweight technology to reduce deployment costs.

5. Conclusion

This study systematically evaluated the performance of TabNet in diabetes screening on BRFSS 2021 large-scale real population data, and reached the following conclusions:

First, TabNet has a significant advantage of high recall rate (77.81%), 41.89 percentage points higher than XGBoost, and 45.17 percentage points higher than random forest, which is suitable for large-scale population diabetes screening.

Secondly, the age subgroup analysis shows that among the high-risk population aged 30-50, the TabNet recall rate is as high as 80.88%, 51.93 percentage points higher than XGBoost, which has direct guiding significance for the screening of key population for diabetes prevention and control.

Thirdly, SHAP interpretability analysis shows that self-assessment of health, hypertension and age are the core characteristics of predicting diabetes, which are consistent with clinical cognition and enhance the clinical acceptability of the model.

To sum up, TabNet performs steadily in large-scale real population diabetes screening, especially in terms of recall rate. TabNet is a feasible and efficient screening tool in diabetes screening scenarios that require a high recall rate.

References

[1] Li, W., Peng, Y., & Peng, K. (2024). Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. *PLOS ONE, 19*, e0311222. https://doi.org/10.1371/journal.pone.0311222

[2] Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice, 87*, 4–14. https://doi.org/10.1016/j.diabres.2009.10.007

[3] Roobini, M. S., & Lakshmi, M. (2021). Predictive supervised learning model for classification of type 2 diabetes mellitus. *Research Square*. https://doi.org/10.21203/rs.3.rs-1009663/v1

[4] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017). Risk prediction of type II diabetes based on random forest model. In *Proceedings of the 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 1–5. https://doi.org/10.1109/AEEICB.2017.7972421

[5] Sajid, M., et al. (2025). Next-generation diabetes diagnosis and personalized diet-activity management: A hybrid ensemble paradigm. *PLOS ONE, 20*, e0307718. https://doi.org/10.1371/journal.pone.0307718

[6] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), pp. 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

DOI: 10.54254/2755-2721/2026.38393

[7] Usha, V., & Rajalakshmi, N. R. (2026). A novel framework for diabetic risk prediction using SCAW-Net integrated with TabNet architecture. *Computer Methods in Biomechanics and Biomedical Engineering, 29*, 412–430. https://doi.org/10.1080/10255842.2025.2566962

[8] Chokshi, V. (2025). An enhanced AI framework for predicting type II diabetes – Insights from 100000 health records. In *Proceedings of the IEEE ICINVENTS*. https://doi.org/10.1109/ICINVENT.2025.10001234

[9] Narvaez, C., et al. (2025). Diabetes prediction based on lifestyle factors using BRFSS 2015. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2503.04137

[10] Centers for Disease Control and Prevention. (2021). Behavioral risk factor surveillance system (BRFSS) 2021 survey data. CDC, Atlanta, GA, USA.

[11] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 30*, pp. 4765–4774. https://doi.org/10.1609/aaai.v35i8.16826