Mathematical Biosciences and Engineering

MBE, 19 (4): 3597–3608.
DOI: 10.3934/mbe.2022166
Received: 10 December 2021
Revised: 20 January 2022
Accepted: 20 January 2022
Published: 07 January 2022

http://www.aimspress.com/journal/MBE

Research article

Risk prediction of diabetes and pre-diabetes based on physical examination data

Yu-Mei Han[1,#], Hui Yang[2,#], Qin-Lai Huang[2], Zi-Jie Sun[2], Ming-Liang Li[1], Jing-Bo Zhang[1], Ke-Jun Deng[2], Shuo Chen[1,*] and Hao Lin[2,*]

[1] Beijing Physical Examination Center, Beijing, China.
[2] School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.

* **Correspondence:** Email: cs@bjtjzx.com (Shuo Chen), hlin@uestc.edu.cn (Hao Lin).

[#] The authors contribute to the paper equally.

Abstract: Diabetes is a metabolic disorder caused by insufficient insulin secretion and insulin secretion disorders. From health to diabetes, there are generally three stages: health, pre-diabetes and type 2 diabetes. Early diagnosis of diabetes is the most effective way to prevent and control diabetes and its complications. In this work, we collected the physical examination data from Beijing Physical Examination Center from January 2006 to December 2017, and divided the population into three groups according to the WHO (1999) Diabetes Diagnostic Standards: normal fasting plasma glucose (NFG) (FPG < 6.1 mmol/L), mildly impaired fasting plasma glucose (IFG) (6.1 mmol/L ⩽ FPG <7.0 mmol/L) and type 2 diabetes (T2DM) (FPG> 7.0 mmol/L). Finally, we obtained1,221,598 NFG samples, 285,965 IFG samples and 387,076 T2DM samples, with a total of 15 physical examination indexes. Furthermore, taking eXtreme Gradient Boosting (XGBoost), random forest (RF), Logistic Regression (LR), and Fully connected neural network (FCN) as classifiers, four models were constructed to distinguish NFG, IFG and T2DM. The comparison results show that XGBoost has the best performance, with AUC (macro) of 0.7874 and AUC (micro) of 0.8633. In addition, based on the XGBoost classifier, three binary classification models were also established to discriminate NFG from IFG, NFG from T2DM, IFG from T2DM. On the independent dataset, the AUCs were 0.7808, 0.8687, 0.7067, respectively. Finally, we analyzed the importance of the features and identified the risk factors associated with diabetes.

Keywords: diabetes; fasting plasma glucose; physical examination; XGBoost

1. Introduction

Diabetes is a metabolic disorder disease caused by insufficient insulin secretion and insulin secretion disorders [1].The main manifestation of diabetes is hyperglycemia. Long-term exposure of organs to hyperglycemia will cause the damage of physiological system, then leading to chronic progressive lesions and failure of tissues and organs, such as eyes, kidneys, nerves, heart and blood vessels [2]. At present, diabetes Mellitus can be divided into type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM), among which T2DM is the most common type of diabetes, accounting for about 95% of diabetic patients [3]. The main factors leading to T2DM are environmental factors and bad living habits. In addition, age, overnutrition and insufficient exercise are all the triggers of diabetes [4]. From health to T2DM, the development usually goes through three stages: health, pre-diabetes, type 2 diabetes [5]. When T2DM is diagnosed, the blood glucose level of patients will continue to rise, and drug treatment is difficult to reverse [6,7]. However, patients in pre-diabetes can maintain blood glucose stability and even restore health through artificial intervention. Many studies have shown that early diagnosis and treatment of T2DM is the most effective way to prevent and control T2DM. Therefore, early detection and timely adjustment of lifestyle is the key to the treatment of T2DM [8].

With the development of economy and culture, people pay more and more attention to physical examination[9,10]. Finding valuable information related to diabetes from physical examination data and finding out the changing pattern of diabetes at all stages is of great importance to the prevention and treatment of diabetes.

In recent years, many algorithms have been used to predict diabetes. For example, Zou et al. used principal component analysis (PCA) and minimum redundant maximum (mRMR) correlation to screen risk factors, and utilized decision tree (DT), RF and neural network (NN) to predict diabetes [11]. By using mutual information (MI) and Gini impurity (GI) to screen diabetes-related risk factors in physical examination data, Yang et al. established a cascade diabetes risk prediction system [12]. The invasive risk assessment model HCL predicted diabetes by using invasive characteristics and referring to Harvard Cancer Risk Index [13].

Machine learning algorithms have been widely used in the field of medicine because of their powerful performance [14–17]. Therefore, based on physical examination data in real world, this study used XGBoost, RF, LR, and FCN to predict diabetes, and analyze the impact of these indicators at each stage of T2DM.

2. Materials and methods

2.1. Benchmark Dataset

The physical examination data were collected from Beijing Physical Examination Center from January 2006 to December 2017. In this study, fasting plasma glucose (FPG) index in the physical examination data was used as the standard to classify the sample types of the dataset. FPG can reflect the function of islet B cells, and generally indicate the secretion function of basal insulin, which is the most commonly used indicator for diabetes [18]. Clinical application of FPG is more conducive to the

early diagnosis and prevention of T2DM. According to WHO (1999) diagnostic criteria for diabetes, the population was divided into three groups: normal FPG (NFG, FPG < 6.1 mmol/L), slightly impaired FPG (IFG, 6.1 mmol/L ⩽ FPG < 7.0 mmol/L), and T2DM (T2DM, FPG > 7.0 mmol/L) [19]. Finally, the benchmark data included 1,221,598 NFG samples, 285,965 IFG samples, and 387,076 T2DM samples.

There are 14 initial features in the physical examination data, including waistline, age, systolic pressure (SP), gender, blood uric acid (BUA), serum creatinine (SC), triglyceride, diastolic pressure (DP), glutamic oxalacetic transaminase (GOT), hipline, high-density lipoprotein (HDL), glutamic-pyruvic transaminase (GPT), height, blood urea nitrogen(BUN), weight, total cholesterol (TC), and low density lipoprotein (LDL). Height and Waist circumference cannot directly evaluate a person's obesity, so we added waist height ratio (WHtR) to reflect whether a person has visceral fat accumulation. As a result, total of 15 features were used to perform further analysis and model construction.

To facilitate the performance evaluation of the model, we divided the data set into training set and test set according to the ratio of 7:3. Thus, the benchmark dataset can be formulated as

$$\begin{cases} S^{train} = S_1^{train} \cup S_2^{train} \cup S_3^{train} \\ S^{test} = S_1^{test} \cup S_2^{test} \cup S_3^{test} \end{cases} \tag{1}$$

where the symbol 1, 2 and 3 represent the NFG, IFG and T2DM, respectively. The "*train*" and "*test*" denotes the training data and test data, respectively.

2.2. Machine learning methods

In this study, eXtreme Gradient Boosting (XGBoost), random forest (RF), logistic regression (LR), and fully connected neural network (FCN) algorithm were used as the classifier. The details are as follows.

2.1.1. eXtreme Gradient Boosting (XGBoost)

XGBoost is based on the gradient boosting algorithm [20–22]. In the modeling process, features are spitted through continuous adding trees. In each time, a tree is added to learn a new function to fit the residual of the last prediction. After the training, a gradient boosting model of K trees is obtained. The ultimate goal of XGBoost is to make the predicted value of the tree group as close to the true value as possible, and to have as large a generalization range as possible.

The objective function of XGBoost is:

$$L(\emptyset) = \sum_i l(y_i' - y_i) + \sum_k \Omega(f_t) \tag{2}$$

where y_i' is the output of the entire cumulative model, and the regularization term $\sum_k \Omega(f_t)$ is a function representing the complexity of the tree. The smaller the value, the lower the complexity and the stronger the generalization ability of the model.

In this study, Gini impurity (GI) is used to evaluate the contribution of features to the model. In the tree model, better decision-making conditions can be selected by comparing the value of GI. Each division of tree nodes should try to make the GI as low as possible. GI is mainly used to solve the problem of high computational complexity. It is defined as:

$$Gini(t) = 1 - \sum_{i-0}^{c-1} p(i|t)^2 \quad (3)$$

where t represents a given node, i represents any category of label, and $p(i|t)$ represents the proportion of label category i on node t.

2.1.2. Random Forest (RF)

RF is also a tree-based ensemble classifier which is a representative model of the bagging method. The core idea of the bagging method is to construct multiple independent evaluators, and then the prediction results are determined by the principle of average or majority voting [23,24].

2.1.3. Logistic Regression (LR)

LR is a generalized linear regression analysis algorithm, and is often used in the field of disease diagnosis [25,26]. It is a variation of linear regression, and an algorithm widely used in the field of regression and classification. LR is to construct a mapping from X to $\hat{y}$ and calculates the parameters of the model formulated as.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \quad (4)$$

The process is calculated as follows. Firstly, a loss function is defined, and then the parameter vector is solved by minimizing the loss function. Finally, the LR uses the Sigmoid function to control the output between 0 and 1:

$$g(z) = \frac{1}{1+e^{-z}} \quad (5)$$

The Sigmoid function distributes the value of $g(z)$ between 0 and 1. When $g(z)$ approaches 0, the label of the sample is category 0, and when $g(z)$ is close to 1, the label of the sample is category 1. In this way, a classification model can be obtained.

2.1.4. Fully connected neural network (FCN)

FCN generally consists of three parts, an input layer, a hidden layer and an output layer [27,28]. Each layer uses the output of the previous layer as input, and then outputs to the next level. The most basic unit in a neural network is a neuron. Each neuron receives multiple inputs and produces an output. Multiple neurons are connected to each other to form a neural network. Fully connected neural network (FCN) generates nonlinear output through activation functions. The commonly used activation functions are ReLU, Sigmoid, and Tanh. FCN training is divided into two processes: forward propagation and backward propagation. The forward propagation fits the features, and then uses the loss function to calculate the gap between the model output value and the target value. Backpropagation uses the gradient descent method to update the parameters of each layer according to the loss function value generated by the forward propagation, thereby optimizing and updating parameter.

We established a three-layer fully connected neural network, the input layer has 18 neurons. The first layer has 7 neurons and the second layer has 4 neurons respectively, the activation function is 'ReLU', the optimization function is 'RMSprop'. The output layer has three neurons, the activation

function is 'Softmax'.

2.2. Performance measurement

In this study, accuracy, precision, recall, F1 and AUC were used to evaluate the performance of proposed models [29], which were calculated as follows:

$$\begin{cases} \text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} & 0 \le Sn \le 1 \\ \text{Precision} = \frac{TP}{TP+FP} & 0 \le Sp \le 1 \\ \text{Recall} = TPR = \frac{TP}{TP+FN} & 0 \le Acc \le 1 \\ \text{FPR} = \frac{FP}{FP+TN} & 0 \le Acc \le 1 \\ \frac{2}{F1} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} & 0 \le Acc \le 1 \end{cases} \tag{6}$$

where *TP* represents true positives, describing the number of correctly predicted positive samples; *FP* denotes false positives, representing the number of negative samples predicted as positive; *FN* indicates false negatives, representing the number of positive samples classified as negative; *TN* denotes true negatives, representing the number of samples correctly predicted as negative. Accuracy is the ratio of the number of all predicted correct samples divided by the total number of samples.

The receiver operating characteristic (ROC) curve is often used to measure the predictive power of the current method across the entire range of algorithm decision value [30]. The ROC can reveal the relationship between true positive rate (TPR) and false positive rate (FPR). We used the area under the ROC curve, referred to as area under curve (AUC), to evaluate the performance of the model.

2.3. Model validation

Generally, there are three methods for model verification: Holdout test, K-Fold cross-validation test and Leave-One-Out (LOO) test [31,32].

Holdout test divides the sample into two mutually exclusive parts, one part is used as the training set and the other part is used as the test set. The model is trained on the training set and examined on the test set. All evaluation indexes were calculated on the test set. K-Fold cross-validation divides the data set into K mutually exclusive data subsets. Each time, one data subset is used as the test set, and all other subsets are used as the training set. Traverse these K subsets in turn. Finally, the average values of the evaluation indexed are used as the final evaluation indexes. The stability of K-Fold cross-validation is closely related to the value of K. If the K value is too small, the experimental stability is not enough. If the K value is too large, the modeling cost may increase. Generally, the K value is 5 or 10. LOO is a special K-Fold cross-validation, where k is equal to the number of sample in the data set. The results obtained by this method are the same as the training entire test. The expected value of the set is the closest, but the cost is too large.

In this article, we use Holdout test for model verification.

3. Results and discussion

In this study, four kinds of machine learning methods that are XGBoost, RF, LR and FCN were

used as the classifier. The following two experiments were performed as follows.

3.1. Prediction of NFG, IFG and T2DM

In the first experiments, based on the above four methods, four-classification models were established to distinguish NFG, IFG and T2DM.We used S_1^{train}, S_2^{train} and S_3^{train} to train the four machine learning methods for constructing models. The S_1^{test}, S_2^{test} and S_3^{test} were utilized to investigate the performance of models for the prediction of NFG, IFG and T2DM. The results were recorded in Table 1 and shown in Figure 1. Table 2 displays the six evaluation indexes of four models on test data. From the table, we noticed that XGBoost could produce the best results with the AUC (macro) of 0.7874 and the AUC (micro) of 0.8633. It is worth noting that the prediction result of FCN is the worst, suggesting that FCN is not suitable for health data analysis. This is consistent with the fact that neural network is not suitable for the analysis of less characteristic samples. Figure 1 shows the ROC curves of four different classifiers on test set. For each algorithm, we draw the micro-average ROC curve, macro-Average ROC Curve and any two kinds of ROC curves. According to Figures 1 (a), we can also see that the AUCs of XGBoost identifying NFG, IFG, and T2DM from the entire population are 0.79, 0.70, and 0.84, respectively.

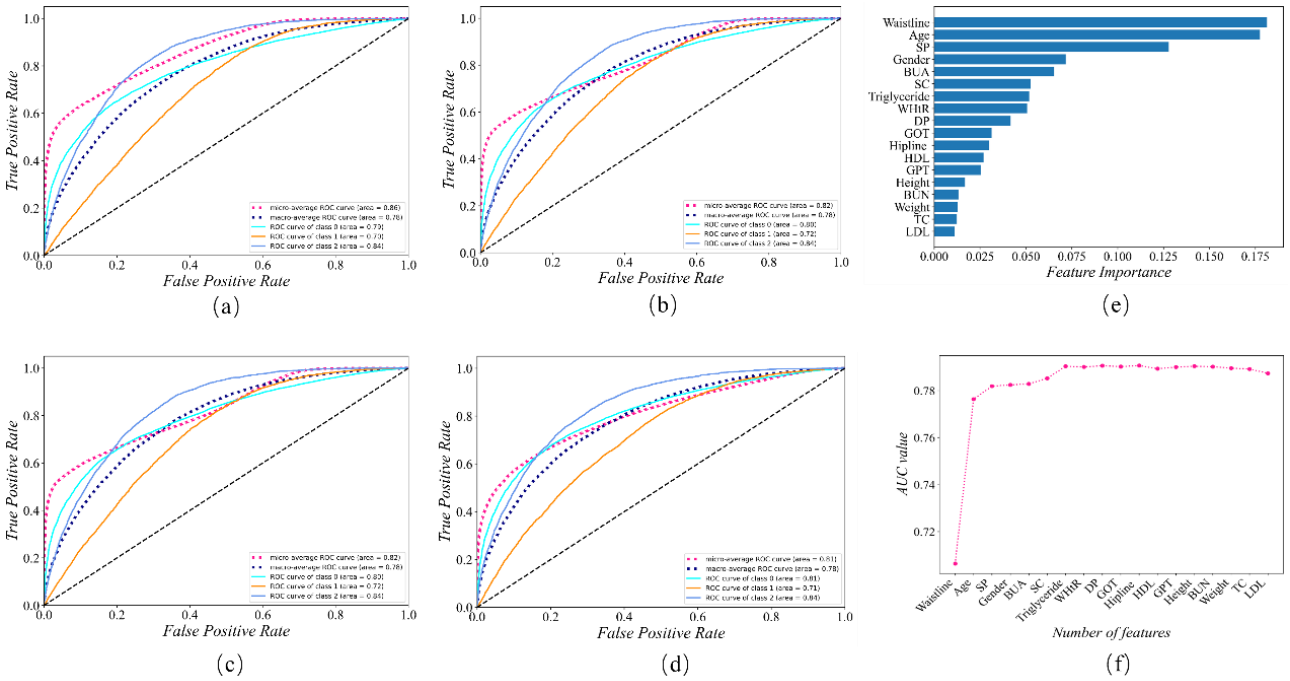

Figure 1. The results for the prediction of NFG, IFG and T2DM. (a) The ROC curves of the algorithm XGBoost, (b) The ROC curves of the algorithm RF, (c) The ROC curves of the algorithm LRs, (d) The ROC curves of the algorithm FCNs, (e) The feature importance using GI, (f) The IFS curve for feature importance using XGBoost.

Subsequently, we performed feature analysis and showed the results in Figure 2. shows the feature importance of XGBoost based on dataset 1. Waist circumference ranked first respectively, indicating that obesity is the most important risk factor for diabetes, and age ranked second. The older the age, the greater the risk of diabetes. Figure 3 shows the incremental feature selection strategy (IFS) curve,

it can be seen that when the first 7 features (Waistline, Age, SP, Gender, BUA, SC, Triglyceride) are used for modeling, the model achieves the highest AUC, and the addition of features does not improve the overall results of the model. We believe that these 7 features are important risk factors for distinguishing NFG, IFG and T2DM.

Table 1. The results for the prediction of NFG, IFG and T2DM.

Algorithm	Accuracy	Precision (weighted)	F1-score (weighted)	Recall (weighted)	AUC (micro)	AUC (macro)
XGBoost	0.6871	0.8192	0.7367	0.6871	0.8633	0.7874
RF	0.6590	0.8260	0.7185	0.6590	0.8233	0.7842
LR	0.6540	0.8334	0.7159	0.6540	0.8068	0.7841
FCN	0.5593	0.5601	0.5560	0.5593	0.7607	0.7472

3.2. Discrimination between any two classes

On the basis of benchmark dataset, three binary models were established to distinguish NFG and IFG, NFG and T2DM, as well as IFG and T2DM. The importance of features in each model was assessed using GI, and incremental feature selection (IFS) was used to find the optimal feature subset. Due to good performance and wide usage in healthy data, we only used XGBoost to construct the three models. Results have been recorded in Table 2.

Table 2. The results for the discrimination between any two classes by using XGBoost.

Dataset	Recall	Accuracy	Precision	F1-score	AUC
NFG *vs* IFG	0.6732	0.7220	0.2047	0.3140	0.7808
NFG *vs* T2DM	0.7611	0.8039	0.2194	0.3409	0.8687
IFG *vs* T2DM	0.7960	0.5891	0.4983	0.6129	0.7067

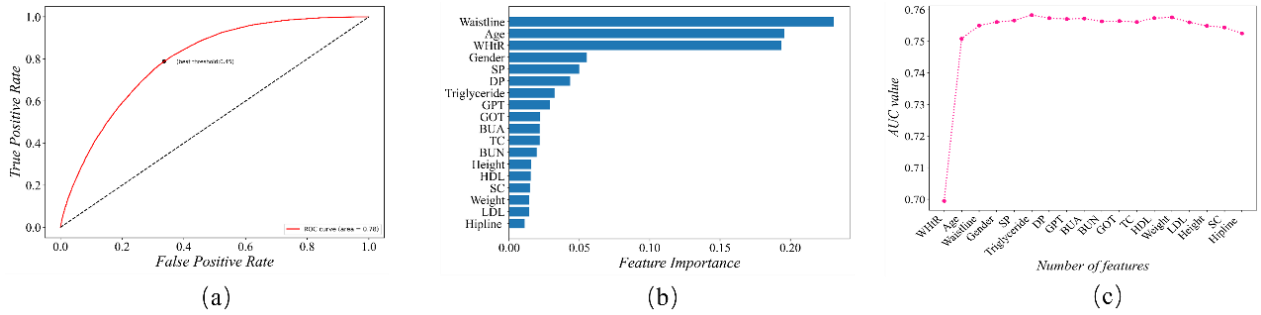

Figure 2. The results for discriminating NFG from IFG. (a) ROC curve, (b) The feature importance, (c) The IFS curve for feature selection.

At first, we built a model for discriminating between NFG from IFG. ROC curve and feature rank of the model were drawn in Figure 2. Results show that the AUC is 0.7808. There is little difference between NFG and IFG. Although blood sugar is elevated in the pre-diabetes stage, the pancreatic islets have not been completely impaired. It will not cause irreversible damage to the body. From Figure 2b

and c, it can be observed that the features with the most importance characteristics at this case are waistline, Age, WHtR, Gender and SP, indicating that the risk factors for the early population are obesity, age and hypertension.

Subsequently, we focused on the discrimination between NFG vs T2DM. From Table 2 and Figure 3a, the XGBoost-based model could produce the AUC of 0.8687. The model established by physical examination indicators can more accurately distinguish normal people from diabetic people. The order of feature importance is Age, Waistline, Triglyceride, WHtR, SP, Gender and SC (Figure 3b). In the identification of diabetic patients, some molecular markers, such as triglycerides, play an important role, which reflects the physiological level of diabetic patients. At present, the diagnosis rate of diabetes in China is less than 50%. It is of great significance to diagnose diabetic patients through physical examination indicators, especially in rural China's free physical examination.

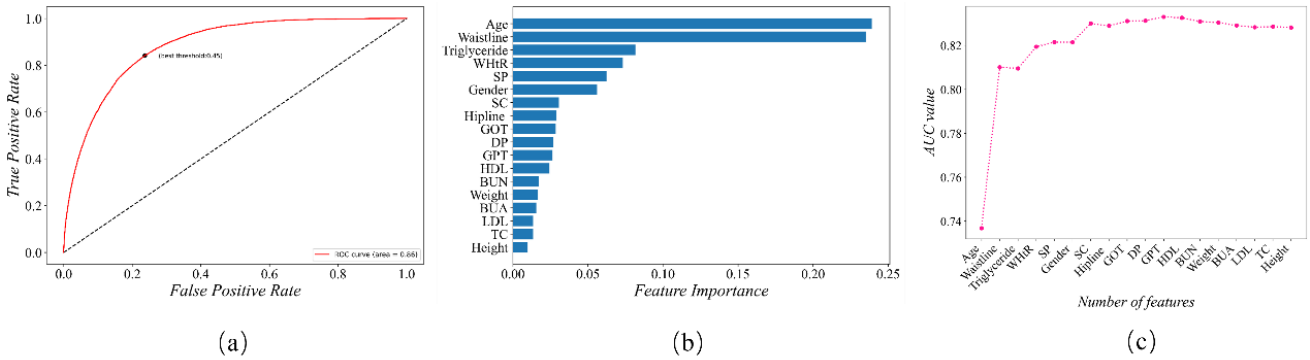

Figure 3. The results for discriminating NFG from T2DM. (a) ROC curve, (b) The feature importance, (c) The IFS curve for feature selection.

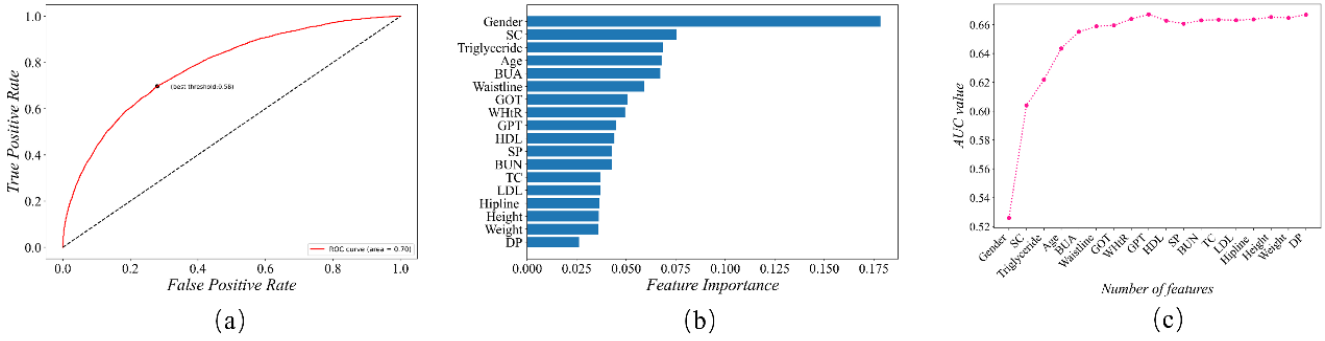

Figure 4. The results for discriminating IFG from T2DM. (a) ROC curve of XGBoost, (b) The feature importance, (c) The IFS curve for feature selection.

The third binary model was built for distinguishing IFG from T2DM based on XGBoost. Based on the results in Table 2 and Figure 4a, we may notice that the model could achieve the AUC of 0.7067 on test dataset. This prediction accuracy is the lowest among the three two classification models. This is mainly due to the fact that many physical indicators of pre diabetes and diabetes are very similar. Patients with pre diabetes are not easily controlled and treated, and are easily converted to diabetic patients. In this classification problem, both IFG population and T2DM population are exposed to hyperglycemia and have an impact on various physical indicators. Figure 4b and c conclude that the

most important features are Gender, SC, Triglyceride, Age, BUA, Waistline, GOT, WHtR, GPT. Some special features, such as SC and GOT, may indicate that renal and liver function of T2DM population may be impaired compared with IFG population.

4. Conclusions

Diabetes is a metabolic disease. From health to diabetes, there are generally three stages: health, pre-diabetes and type 2 diabetes. It is worth studying how to use machine learning methods to early predict and diagnose the disease. In the three-classification experiment of distinguishing NFG, IFG and T2DM, by comparing the results of the four classifiers: XGBoost, RF, LR, and FCN, we can find that there is little difference between them. XGBoost is slightly better than other classifiers, with AUC (macro) of 0.7874 and AUC (micro) of 0.8633. Then, we chose XGBoost as the basic classifier, and constructed three binary classification models to distinguish between NFG and IFG, NFG and T2DM, IFG and T2DM. The AUCs of these models on test dataset are 0.7808, 0.8687 and 0.7067, respectively. We used GI index to evaluate the importance of features, sort the features according to their importance, and mine relevant risk factors by combining with IFS strategy. Overall, Age, Triglyceride, WHtR, and SP are important risk factors. In particular, it was found that T2DM patients may have liver and kidney damage.

Through this work, we hope to explore the possibility of early prediction of diabetes with physical examination data. And we hope to dig out valuable information related to diabetes from the physical examination data and other omics data [33], and discover the changes in the each stage of diabetes, so as to provide clues for early prevention and treatment of diabetes. In the future, we hope to clarify the causal relationship between various risk factors and diabetes through cohort studies and Mendelian randomization studies, and explore some effective intervention schemes on this basis.

Acknowledgments

The study was supported by grants from the National Key R&D Program of China (2020YFC2003403), Capital's Funds for Health Improvement and Research (2018-2-2242) and the National Natural Science Foundation of China (82130112).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. J. M. Lachin, D. M. Nathan, D. E. R. Group, Understanding metabolic memory: The prolonged influence of glycemia during the Diabetes Control and Complications Trial (DCCT) on future risks of complications during the study of the Epidemiology of Diabetes Interventions and Complications (EDIC), *Diabetes Care*, (2021), Online ahead of print, https://doi.org/10.2337/dc20-3097
2. G. Triplett, S. Eichold, Concurrent diabetes mellitus and sickle cell disease, *Diabetes Care*, **2** (1979), 327–328. https://doi.org/10.2337/diacare.2.3.327a
3. C. Greenhill, Diabetes: How does leptin decrease hyperglycaemia in T1DM and T2DM? *Nat. Rev.*

Endocrinol., **10** (2014), 511. https://doi.org/10.1038/nrendo.2014.104
4. D. Holmes, Diabetes: New marker to predict risk of T2DM, *Nat. Rev. Endocrinol.*, **13** (2017), 625. https://doi.org/10.1038/nrendo.2017.128
5. M. Kaare, K. Mikheim, K. Lillevali, K. Kilk, T. Jagomae, E. Leidmaa, et al., High-fat diet induces pre-diabetes and distinct sex-specific metabolic alterations in Negr1-deficient mice, *Biomedicines*, **9** (2021), 1148. https://doi.org/10.3390/biomedicines9091148.
6. Correction: Prevalence of diabetes, pre-diabetes and associated risk factors: Second National Diabetes Survey of Pakistan (NDSP), 2016-2017, *BMJ Open*, **8** (2019), e020961corr1. https://doi.org/10.1136/bmjopen-2017-020961corr1
7. C. Ao, L. Yu, Q. Zou, Prediction of bio-sequence modifications and the associations with diseases, *Brief Funct. Genom.*, **20** (2021), 1–18. https://doi.org/10.1093/bfgp/elaa023
8. M .D. Campbell, T. Sathish, P. Z. Zimmet, K. R. Thankappan, B. Oldenburg, D. R. Owens, et al., Benefit of lifestyle-based T2DM prevention is influenced by prediabetes phenotype, *Nat. Rev. Endocrinol.*, **16** (2020), 395–400. https://doi.org/10.1038/s41574-019-0316-1
9. A. O. Amuta, W. Jacobs, A. E. Barry, An examination of family, healthcare professionals, and peer advice on physical activity behaviors among adolescents at high risk for Type 2 diabetes, *Health Commun.*, **32** (2017), 857– 863. https://doi.org/10.1080/10410236.2016.1177907
10. J.P. Wei, T. Luo, Y. Wang, W. Lu, Screening differential hub genes related with the hypoglycemic effect of quercetin through data mining, *Curr. Bioinform.*, **16** (2021), 1152–1160. https://doi.org/10.2174/1574893616666210617110314
11. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting diabetes mellitus with machine learning techniques, *Front. Genet.*, **9** (2018), 515. https://doi.org/10.3389/fgene.2018.00515
12. Z. Li, C. Zhao, Q. Fu, J. Ye, L. Su, X. Ge, et al., Neodymium (3+)-Coordinated black phosphorus quantum dots with retrievable NIR/X-Ray optoelectronic switching effect for anti-glioblastoma, *Small*, (2021), Online ahead of print. https://doi.org/10.1002/smll.202105160
13. A. B. Goldfine, V. A. Fonseca, The use of colesevelam HCl in patients with type 2 diabetes mellitus: Combining glucose- and lipid-lowering effects, *Postgrad. Med.*, **121** (2009), 13–18. https://doi.org/10.3810/pgm.2009.05.suppl53.288
14. Q. Zhu, Y. Fan, X. Pan, Fusing multiple biological networks to effectively predict miRNA-disease associations, *Curr. Bioinform.*, **16** (2021), 371–384. https://doi.org/10.2174/1574893615999200715165335
15. L. Wei, W. He, A. Malik, R. Su, L. Cui, B. Manavalan, Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework, *Brief. Bioinform.*, **22** (2021), bbaa275. https://doi.org/10.1093/bib/bbaa275
16. M. M. Hasan, M. A. Alam, W. Shoombuatong, H. W. Deng, B. Manavalan, H. Kurata, NeuroPred-FRL: An interpretable prediction model for identifying neuropeptide using feature representation learning, *Brief. Bioinform.*, **22** (2021), bbab167. https://doi.org/10.1093/bib/bbab167
17. M. M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, B. Manavalan, HLPpred-Fuse: Improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation, *Bioinformatics*, **36** (2020), 3350–3356. https://doi.org/10.1093/bioinformatics/btaa160
18. H. Jun, J. Lee, H. A. Lee, S. E. Kim, K. N. Shim, H. K. Jung, et al., Fasting blood glucose variability and unfavorable trajectory patterns are associated with the risk of colorectal cancer, *Gut. Liver*, (2021), Online ahead of print. https://doi.org/10.5009/gnl210048

19. The Expert Committee on the Diagnosis, Classification of Diabetes Mellitus, Report of the expert committee on the diagnosis and classification of diabetes mellitus, *Diabetes Care*, **26** (2003), S5–S20. https://doi.org/10.2337/diacare.26.2007.s5
20. A. Ogunleye, Q. G. Wang, X. G. Boost, Model for chronic kidney disease diagnosis, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17** (2020), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071
21. P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong, H. Zheng, Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer, *IEEE Trans. Biomed. Eng.*, **68** (2021), 148–160. https://doi.org/10.1109/TBME.2020.2993278
22. F. Ahmad, A. Farooq, M. U. G. Khan, Deep learning model for pathogen classification using feature fusion and data augmentation, *Curr. Bioinform.*, **16** (2021), 466–483. https://doi.org/10.2174/1574893615999200707143535
23. S. Jiao, Q. Zou, H. Guo, L. Shi, iTTCA-RF: A random forest predictor for tumor T cell antigens, *J. Transl. Med.*, **19** (2021), 449. https://doi.org/10.1186/s12967-021-03084-x
24. Y. M. Dong, J. H. Bi, Q. E. He, K. Song, ESDA: An improved approach to accurately identify human snoRNAs for precision cancer therapy, *Curr. Bioinfor.*, **15** (2020), 34–40. https://doi.org/10.2174/1574893614666190424162230
25. X. Song, X. Liu, F. Liu, C. Wang, Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis, *Int. J. Med. Inform.*, **151** (2021), 104484. https://doi.org/10.1016/j.ijmedinf.2021.104484
26. L. Zhang, Y. He, H. Song, X. Wang, N. Lu, L. Sun, et al., Elastic net regularized softmax regression methods for multi-subtype classification in cancer, *Curr. Bioinform.,* **15** (2020), 212–224. https://doi.org/10.2174/1574893613666181112141724
27. Y. Wang, R. Zhang, M. Pi, J. Xu, M. Qiu, T. Wen, Correlation between TCM Syndromes and Type 2 diabetic comorbidities based on fully connected neural network prediction model, *Evid. Based Complement Alternat. Med.*, **2021** (2021), 6095476. https://doi.org/10.1155/2021/6095476
28. M. Awais, W. Hussain, N. Rasool, Y. D. Khan, iTSP-PseAAC: Identifying tumor suppressor proteins by using fully connected neural network and PseAAC, *Curr. Bioinform.*, **16** (2021), 700–709. https://doi.org/10.2174/1574893615666210108094431
29. J. Phillips, S. K. Poon, D. Yu, M. Lam, M. Hines, M. Brunner, et al., A conceptual measurement model for ehealth readiness: A team based perspective, *AMIA Annu. Symp. Proc.*, **2017** (2017), 1382–1391.
30. M. Kottas, O. Kuss, A. Zapf, A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies, *BMC Med. Res. Methodol.*, **14** (2014), 26. https://doi.org/10.1186/1471-2288-14-26
31. M. T. Rouabah, A. Tounsi, N. E. Belaloui, Genetic algorithm with cross-validation-based epidemic model and application to the early diffusion of COVID-19 in Algeria, *Sci. Afr.*, **14** (2021), e01050. https://doi.org/10.1016/j.sciaf.2021.e01050
32. L. Zhu, G. Duan, C. Yan, J. Wang, Prediction of microbe-drug associations based on chemical structures and the KATZ measure, *Curr. Bioinform.*, **16** (2021), 807–819. https://doi.org/10.2174/1574893616666210204144721
33. J. Long, H. Yang, Z. Yang, Q. Jia, L. Liu, L. Kong, et al., Integrated biomarker profiling of the metabolome associated with impaired fasting glucose and type 2 diabetes mellitus in large-scale Chinese patients, *Clin. Transl. Med.*, **11** (2021), e432. https://doi.org/10.1002/ctm2.432

AIMS Press