APL Bioengineering ARTICLE pubs.aip.org/aip/apb

Stacking model framework reveals clinical biochemical data and dietary behavior features associated with type 2 diabetes: A retrospective cohort study

Cite as: APL Bioeng. **8**, 046111 (2024); doi: 10.1063/5.0207658
Submitted: 10 March 2024 · Accepted: 6 November 2024 ·
Published Online: 21 November 2024

Yong Fu,[1,2] Xinghuan Liang,[1] Xi Yang,[3] Li Li,[1] Liheng Meng,[1] Yuekun Wei,[2,4] Daizheng Huang,[2,a)] and Yingfen Qin[1,a)]

AFFILIATIONS

[1]Department of Endocrinology, The First Affiliated Hospital of Guangxi Medical University. Guangxi Endocrine Clinical Key Specialty, No. 6 Shuangyong Road, Nanning 530021, Guangxi, China
[2]Life Sciences Institute, Guangxi Medical University, No. 22 Shuangyong Road, Nanning 530021, Guangxi, China
[3]Department of Geriatric Endocrinology and Metabolism, The First Affiliated Hospital of Guangxi Medical University, No. 6 Shuangyong Road, Nanning 530021, Guangxi, China
[4]School of Information and Management, Guangxi Medical University, No. 22 Shuangyong Road, Nanning 530021, Guangxi, China

[a)]Authors to whom correspondence should be addressed: huangdaizheng@gxmu.edu.cn and qinyingfen@gxmu.edu.cn

ABSTRACT

Background: Type 2 diabetes mellitus (T2DM) is the most common type of diabetes, accounting for around 90% of all diabetes. Studies have found that dietary habits and biochemical metabolic changes are closely related to T2DM disease surveillance, but early surveillance tools are not specific and have lower accuracy. This paper aimed to provide a reliable artificial intelligence model with high accuracy for the clinical diagnosis of T2DM. **Methods:** A cross-sectional dataset comprising 8981 individuals from the First Affiliated Hospital of Guangxi Medical University was analyzed by a model fusion framework. The model includes four machine learning (ML) models, which used the stacking method. The ability to leverage the strengths of different algorithms to capture complex patterns in the data can effectively combine questionnaire data and blood test data to predict diabetes. **Results:** The experimental results show that the stacking model achieves significant prediction results in diabetes detection. Compared with the single machine learning algorithm, the stacking model has improved in the metrics of accuracy, recall, and F1-score. The test set accuracy is 0.90, and the precision, recall, F1-score, area under the curve, and average precision (AP) are 0.91, 0.90, 0.90, 0.90, and 0.85, respectively. Additionally, this study showed that HbA1c ($P < 0.001$,OR = 2.203), fasting blood glucose (FBG) ($P < 0.001$,OR = 1.586), Ph2BG ($P < 0.001$,OR = 1.190), age ($P < 0.001$,OR = 1.018), Han nationality ($P < 0.001$,OR = 1.484), and carbonate beverages ($P = 0.001$,OR = 1.347) were important predictors of T2DM. **Conclusion:** This study demonstrates that stacking models show great potential in diabetes detection, and by integrating multiple machine learning algorithms, stacking models can significantly improve the accuracy and stability of diabetes prediction and provide strong support for disease prevention, early diagnosis, and individualized treatment.

I. INTRODUCTION

About 422×10^6 people worldwide have diabetes, the majority living in low- and middle-income countries, and 1.5×10^6 deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades.[1] According to International Diabetes Federation statistics, China has become the country with the largest number of adults aged 20–79 years with diabetes. The huge diabetic population places a great burden on modern healthcare systems and a great economic burden on the family of patients with diabetes. Diabetes accounts for more than $966 billion in health spending, and this amount is still increasing.[2]

Type 2 diabetes is the most common type, a metabolic chronic disease characterized by insulin resistance and elevated blood glucose levels.[3] Type 2 diabetes not only imposes health burdens on the

patient, such as a significantly increased risk of complications, such as cardiovascular disease, neuropathy, and retinopathy, but it also puts significant socioeconomic pressures, such as rising healthcare costs, a reduced workforce, and an increased burden on families and social increased burden on family and social support systems. However, the symptoms of the disease are not obvious at the onset, making it difficult to detect and confirm the diagnosis.[4] Prevention and detection are essential means to cope with diabetes. Therefore, establishing effective prediction models to assess an individual's risk of diabetes is essential.

Machine learning (ML) is widely used in various fields, Shaukat *et al.*[5,6] explored the potential of machine learning techniques to improve the detection of unknown and polymorphic security attacks, and Alam[7,8] proposed a new framework to identify prognostic factors for malignant mesothelioma through association rule mining techniques. (All abbreviations in the paper are listed in Nomenclature). The application of machine learning and deep learning (DL) algorithms in healthcare has been proven.[9] Nowadays, many scholars use ML for disease detection, Kumar *et al.*[10] used ML and DL models for speech recognition task to achieve detection of dementia, Abdullah *et al.*[11] study proved the potential of deep learning convolutional neural networks (CNNs) and sequential CNNs for disease detection and classification, Srinivas *et al.*[12] proposed three migration learning based CNN models to localize brain tumors, and Alsubaie *et al.*[13] proposed a novel CNN architecture called ConvADD for detecting Alzheimer's disease. Shaukat *et al.*[14–16] applied machine learning and deep learning technology in the field of network security, and in the past decade, the application of machine learning technology in the field of network security has made remarkable progress. From static analysis to dynamic analysis, and then to use deep learning for malicious software detection, expanding the application scope of machine learning technology, combining DL and ML, eliminates the demand for intensive characteristic engineering tasks and field knowledge, the accuracy of malicious software reached 99.06%, and this fusion can combine the advantages of both, improve the performance of the model and robustness.

This work consists of the following contributions:

Propose a new stacking machine learning framework to analyze the 8981-case cross-sectional dataset from the First Affiliated Hospital of Guangxi Medical University. Data problems are solved by feature engineering and data preprocessing methods, hyperparameters are optimized by learning curves and grid search, and model performance is evaluated using cross-validation and medical statistical methods.

The proposed machine learning framework in this work is superior to any separately constructed machine learning methods and ensemble models. The effect of data preprocessing on the model was also examined using the PIMA database to demonstrate the stability of the model. Statistical analysis was performed to show that our proposed stacking model is capable of better detection efficiency. The reliability of the model was demonstrated using feature importance visualization, and the potential value of other features for diabetes diagnosis was explored.

II. RELATED WORKS

The application of machine learning (ML) technology in medical diagnostic systems has become mature. This technology has been proven to be accurate in diagnosis, successful in treatment, and cost effective.[17] In this research, we conducted a thorough search of the PubMed and Web of Science databases using the following search terms to identify recent and relevant studies: Type 2 diabetes mellitus (T2DM), type 2 diabetes, diabetes mellitus, machine learning, stacking model, fusion model, and ensemble model. Nineteen relevant studies were adopted. Many scholars have performed several research using the Pima Indians dataset to improve the accuracy of models in clinical prediction. Joshi *et al.*[18] achieved 78.26% accuracy on the Pima Indians dataset using logistic regression (LR) and decision tree (DT). Chang *et al.*[19] used random forest (RF) to improve accuracy to 79.57%. Furthermore, Adua *et al.*[20] recruited 219 patients with type 2 diabetes mellitus (T2DM) and 219 healthy individuals. Four ML classification algorithms, namely, Naïve–Bayes (NB), k-nearest neighbor (KNN), support vector machine (SVM), and DT, were used to predict T2DM. NB classifiers yielded 94% accuracy.

In processing small datasets, traditional ML methods can obtain satisfactory results. However, in terms of accuracy, the development of traditional ML models is near saturation. Many scholars use traditional ML algorithms for comparison with new models, and more scholars focus their attention on ensemble classifiers and neural network models. In the Pima Indians dataset, Khanam and Foo used a neural network with two hidden layers to increase accuracy to 88.6%.[21] Edeh *et al.*[22] presented a supportive diagnostic system based on the comparison of four predictive algorithm models for predicting diabetes in two different databases (Frankfurt Hospital diabetes dataset and Pima Indian dataset). Based on several performance assessment methods, such as accuracy, recall, and F1 score, the authors concluded that RF provides a more accurate prediction and a higher performance than other models. Xie *et al.*[23] constructed a ML prediction model using diabetes data from 138 146 participants, and the experimental results showed that the neural network model had the best model performance with an area under the curve (AUC) of 0.7949 and an accuracy of 82.4%. When using a single machine learning model, there may be challenges such as overfitting, underfitting, and lack of generalization ability. These difficulties can result in good performance on the training data but inadequate performance on new data.[24,25] To address these concerns, an ensemble machine learning approach can be utilized.

Medical data are usually unbalanced, which affects the performance of the model, and many scholars have used various methods in order to solve this problem. Khushi *et al.*[26] explored the performance effects of 23 class-imbalance methods and three classical classifiers using two datasets, the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial dataset and the National Lung Screening Trial (NLST) dataset (with imbalance ratios of 24.7 and 25.0, respectively); the results of the study showed that class-imbalance balanced learning can significantly improve the classification ability of the model. Talha Mahboob Alam *et al.* encountered the class imbalance problem in a number of studies,[27,28] and they solved the problem by synthetic minority over-sampling technique (SMOTE) technique, borderline SMOTE technique, data augmentation method, and other methods to solve the data imbalance problem, and the experimental results were significantly improved after class imbalance learning. Yang *et al.*[29] used extremely imbalanced dataset (imbalance ratio of 143.7) in their study and employed 23 class imbalance learning methods combined with machine learning models to improve the accuracy of early screening for ovarian cancer.

Ensemble ML methods is a powerful and widely used technique in the field of machine learning, which aims to improve the predictive performance and stability of a model by constructing and combining multiple learners to accomplish a learning task. Ali *et al.*[30] solved the problem of data imbalance by using the resampling algorithms such as random under-sampling, random over-sampling, and SMOTE, and four integrated machine learning algorithms were used to predict schistosomiasis and finally CatBoost model performed the best with 87.1% accuracy. Devnath *et al.*[31] applied deep integration learning technique for automatic detection of cxr pneumoconiosis in coal miners and showed that the integrated integration technique combining deep learning methods outperformed the other methods and achieved 91.50% accuracy in the automatic detection of pneumoconiosis. Integrated ML methods, such as LightGBM and CatBoost, were also applied to diabetes prediction. LightGBM and CatBoost are also applied to diabetes prediction. Liu *et al.* predicted T2DM in older Chinese adults by applying random under sampling to address category imbalance and Shapley additive interpretation to calculate and visualize feature importance. The XGBoost model with 21 features performed the best in predicting T2DM.[32] Yang *et al.*[33] also did related research using the 2011–2017 dataset of patients with diabetes in Luzhou City, China. The XGBoost model also performed excellently in this dataset. Xue *et al.* compared DT, RF, AdaBoost, and DT (AdaBoost), and extreme gradient enhanced DT (XGBoost) based on a cross-sectional study of 584 168 adult subjects from a national physical examination and showed that XGBoost was the best classifier with an AUC of 0.9680.[34] Dong *et al.*[35] compared the XGBoost and LightGBM models in the dataset of PLA General Hospital, and the results showed that LightGBM was better than XGBoost. Rufo *et al.* explored the application of diabetes data from Zewditu Memorial Hospital in Addis Ababa, Ethiopia, in the field of ML and constructed the LightGBM model, which was validated by comparing KNN, SVM, NB, bagging, RF, and XGBoost.[36] Rufo *et al.*[36] also used the LightGBM model, which yielded 98.1% accuracy. In the field of clinical prediction, machine learning models like LightGBM, CatBoost, and XGBoost have proven to be effective. However, there is still a concern with bias and variance. To address this issue, a stacking model that combines various models can reduce bias and variance, ultimately improving the overall performance and generalization of the model.

Ensemble techniques that stacking propose diversity, stability, and outstanding performance illustrated in many recent studies. Xiong *et al.*[37] used voting to combine five ML models to predict diabetes in the dataset from Nanjing Drum Tower Hospital with 91% accuracy and 0.97 AUC. Sumathi and Meganathan used voting to predict gestational DM, and studies showed that the fusion model is superior to the classical ML model and achieved good results with a precision of 94%, a recall of 94%, an accuracy of 94.24%, and a F1 score of 94%.[38] Deberneh and Kim used LR, RF, SVM, XGBoost, and model fusion methods (stacking and soft voting) to train and predict DM using electronic health records collected by a private medical institution and achieved effective results.[39] However, without deep exploration, the superiority of model fusion is not obvious.

Table I outlines the development of ML in diabetes diagnostics. From the table, fusion models can integrate the benefits of a single model to better predict outcomes. The potential of fusion models in disease diagnosis and prognosis remains to be explored. In this paper, we try to prove that the stacking model has more advantages in obtaining a higher prediction accuracy. The study flow chart is depicted in Fig. 1.

III. RESULTS

This section will be divided into five parts to fully demonstrate the results of each step of the experiments. Section III A describes the results of class imbalance learning and feature selection. Section III B shows the performance comparison of stacking with other models, and the results show that the stacked model is able to have better results relative to a single integrated learning model. Section III C shows external validation, and Sec. III D experiments of model comparison by dividing the dataset. The results further validate the reliability of the model. Section III E uses statistical analysis methods for feature evaluation as well as validation of the proposed model against other models.

A. Class imbalance and feature selection

After class imbalance, our dataset has 8630 samples. It contains 2215 with T2DM and 6415 with non-diabetic. After features selected, the results showed Age, Female, Male, HAN, ZHUANG, Smoke, Drink alcohol, Tea, Carbonate Beverages, Coffee, Hypertension, Retinopathy, Hyperlipidemia, Snore, Hypotensive Drugs, SBP, DBP, BMI, WC, HC, CRP, HDL, LDL, TCHO, TG, AST, Y-GT, FBG, P2hPG, HbA1C, and FINS. Figure 2 illustrates the relationship between the number of features and the accuracy of the model. From the curve, it is found that when the features reach 22, the model improvement begins to slow down, and when the features reach 30, the model can reach the highest accuracy.

B. Comparison of the model performance

Table III shows a comparison of the performance of the five models. As can be seen from the data in the table, the stacking model can combine the advantages of the basic learner to produce better results. The stacking model performed best on this dataset, with a test set accuracy of 0.91, and the precision, recall, F1-score, AUC, and AP were 0.91, 0.90, 0.90, 0.90, and 0.85, respectively. The ROC curves and PR curves are shown in Fig. 3.

CatBoost, XGBoost, and LightGBM are recognized as the three leading implementations of gradient boosting decision trees (GBDT), each representing significant advancements within the GBDT framework. These models have become indispensable tools in machine learning, especially for structured data tasks. XGBoost, CatBoost, and LightGBM are highly efficient and capable of delivering state-of-the-art performance across a wide range of machine learning challenges. Each model offers distinct advantages depending on the characteristics of the dataset and the specific requirements of the task, making them critical components in modern machine learning workflows. In this study, these three models were compared with the stacking model, the results are shown in Fig. 4, and the ROC curve of the stacking model is not inferior to the three models, or even better than XGBoost, CatBoost, and LightGBM.

C. External validation

Table IV outlines the stacking model's performance in the Pima Indian medical association (PIMA) dataset and compares the performance of other models. As shown in the table, the stacking model still reached the best performance in the test set with accuracy, precision, recall, and F1-score of 0.74, 0.73, 0.74, and 0.73, respectively.

TABLE I. Summary of literature review.

Authors	Year	Models	Data sources	Accuracy (%) & AUC
Xie *et al.*[23]	2019	SVM, DT, LR, RF, NN, and Gaussian NB	2014 BRFSS dataset	82.4%
Xiong *et al.*[37]	2019	MLP, AdaBoost, RF, SVM, GTB, and voting	Nanjing Drum Tower Hospital dataset	91%
Xue *et al.*[34]	2020	DT, RF, AdaBoost, and XGBoost	National physical examination	90.6%
Yang *et al.*[29]	2020	DT	PLCO	95.32%
Joshi *et al.*[18]	2021	LR and DT	PIDD	78.26%
Adua *et al.*[20]	2021	NB, KNN, SVM, and DT	A hospital and community for African populations in Ghana	93%
Khanam and Foo[21]	2021	NN, DT, KNN, RF, NB, AB, and LR	PIDD	88.6%
Yang *et al.*[33]	2021	XGBoost	2011–2017 dataset of patients with diabetes in Luzhou City, China	87.68%
Rufo *et al.*[36]	2021	LightGBM	Zewditu Memorial Hospital in Addis Ababa, Ethiopia	98.1% and 98.1%
Deberneh and Kim[39]	2021	LR, RF, SVM, XGBoost, CIM, stacking classifier, and soft Voting	Private medical institutions	73%
Sumathi and Meganathan[38]	2021	MLP, SVM, LR, and stacking	PIDD	78.2%
Chang *et al.*[19]	2022	NB, RF, and J48DT	PIDD	79.57%
Alam *et al.*[27]	2022	AlexNet, InceptionV3, and RegNetY-320	MNIST: HAM10 000 dataset	91%
Onyema Edeh *et al.*[22]	2022	RF, SVM, NB, DT, and K-means	Frankfurt Hospital dataset and PIDD	97.6%
Ali *et al.*[30]	2022	Gradient boosting, light gradient boosting, extreme gradient boosting, and CatBoost	Hubei Institute of Schistosomiasis Prevention and Control, China	87.1%
Liu *et al.*[31]	2022	(Simple averaging, multi-weighted averaging, and majority voting (MVOT)	CSIRO dataset, NIOSH teaching chest x-ray dataset and ILO Standard Radiographs	91.50%
Liu *et al.*[32]	2022	LR, DT, RF, and XGBoost	Health screening data of adults older than 65 years in Wuhan, China from 2018–2020	75.03% and 78.05%
Dong *et al.*[35]	2022	LightGBM, XGBoost, AdaBoost, NN, DT, SVM, and LR	PLA General Hospital	81.5%

D. Comparison of divide the dataset

Table V shows the model performance of the four experiments. It can be seen that stacking models perform best from the table. Figure 5 presents the contributions of the features on the models output ranked in four experiments. Permutation feature importance is an effective method for explaining black box models. This is helpful for us looking for features that are significant risk factors for incident T2DM. Figure 6 provides an overview of the distribution of the effects of age and BMI on diabetes. From the figure, it is found that the yellow dots are mostly concentrated between 60 and 80, while the trend line has an upward trend, but the amplitude is not obvious.

E. Statistical analysis

According to the univariate logistic regression analysis in Table VI, Age, Carbonate Beverages, Han, AST, FBG, Ph2BG, and HbA1c are significant predictors of the occurrence of T2DM in the overall population ($P < 0.05$). Tea drinking was not statistically significant in our dataset. Therefore, we not included tea in the multivariate regression. As can be seen from Table VII, the salient features identified in the univariate analysis described above were included in the multivariate logistic regression analysis. The odds ratios (ORs) calculated indicated the relative risk of T2DM. The results showed that Age, Carbonate Beverages, Han, FBG, Ph2BG, and HbA1c were independent predictors of T2DM.

IV. DISCUSSION

With the development of artificial intelligence, machine learning has been widely integrated into the field of medical diagnosis.[15–21,31–36] The stacking model has also been widely applied in the field of diabetes diagnostics.[37,38,40,41] In this retrospective study, we applied four machine learning models to build a stacked model of the risk of type 2 diabetes in the Guangxi area. It was found that the stacking model showed the best performance in predicting type 2 diabetes through

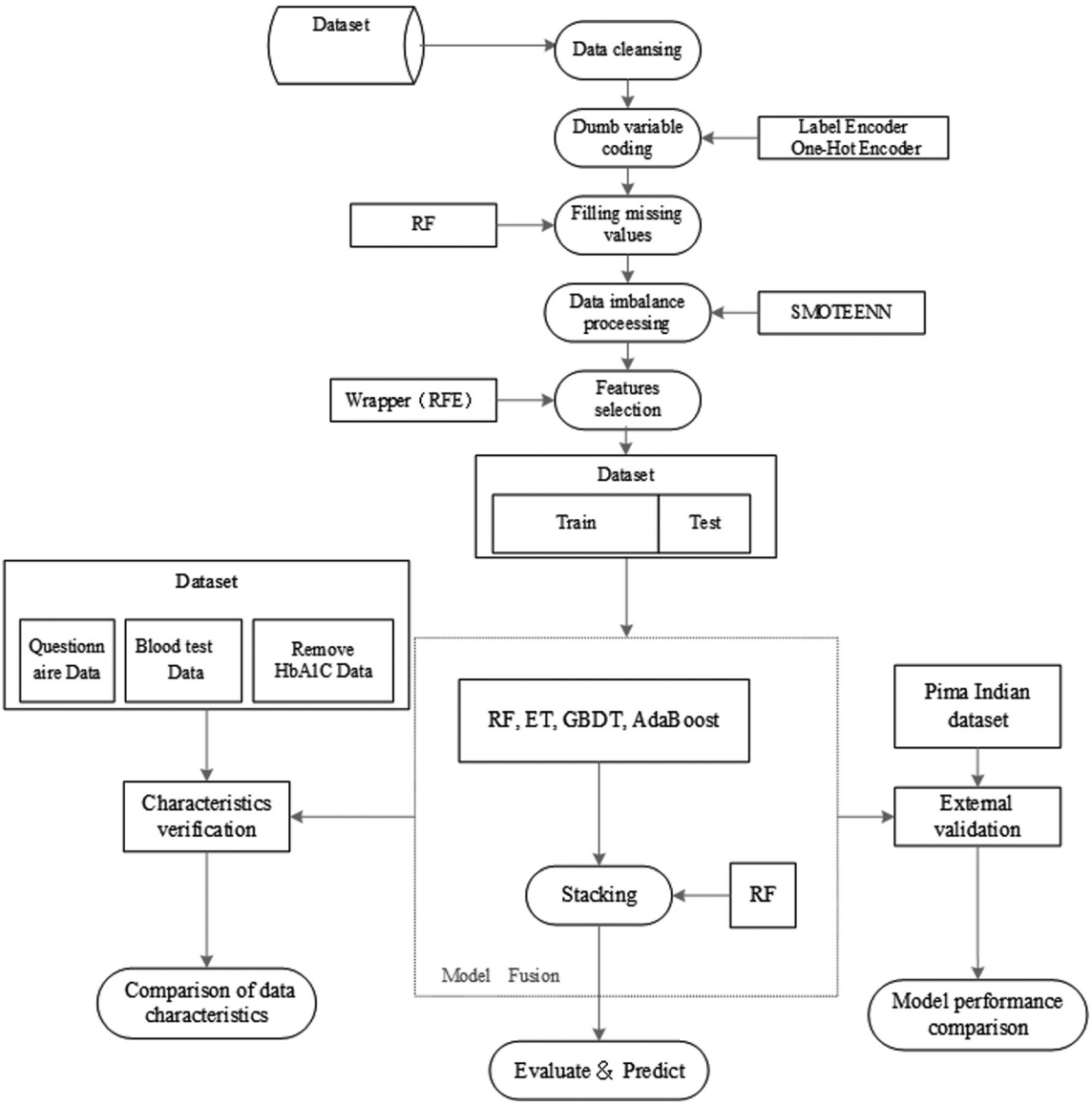

FIG. 1. Study flow chart. Clinical data and blood test data were cleansed, encoded with dumb variables, disposed missing values and class imbalances. Feature selection results were used to construct the dataset. RF, Extra-Trees, GBDT, and AdaBoost were used to build stacking models. The performance of the stacking model was validated in three datasets (questionnaire data, blood test data, and comprehensive data on HbA1C removal). The PIMA dataset serves as an external validation to further validated the model. RF = Random forest; Extra-Trees = extremely randomized trees; and GBDT = gradient boosting decision tree.

model performance comparison and external validation, with a test set accuracy of 0.90, and the precision, recall, F1-score, AUC, and AP were 0.91, 0.90, 0.90, 0.90, and 0.85, respectively. This suggested that the stacking model can use questionnaire data and blood test data to predict early type 2 diabetes, which could benefit the prevention and control of diabetes. Compared with previous research,[40,41] we used random forest to fill the dataset, SMOTEENN to handle class imbalance data, and wrapper for feature selection. Model performance is proven through external verification. In addition, we have added the HbA1c indicator. By ranking the importance of HbA1c indicators in the model, it is proved that the model can predict the occurrence of diabetes based on the currently recognized indicators.

This study also designed four experiments to explore the effect of data preprocessing on the model. Experiment 1 had no imbalance processing and feature selection; experiment 2 did not use imbalance processing but feature selection; experiment 3 used imbalance processing without feature selection; and experiment 4 used imbalance processing and used feature selection. The performance of the stacked fusion model was observed to determine whether unbalanced processing and feature selection were used. We verified the above-mentioned

TABLE II. Features' description.

Type of data	Feature	Description
Questionnaire data	Age	Age at the time of sampling.
	Gender	Male (1) or Female (2)
	Race	Han (1) or Zhuang (2) or other (3)
	Smoke	Whether or not you smoke? No (1) or Yes, but no often (2) or Yes, everyday (3)
	Drink alcohol	Whether or not you drink alcohol? No (1) or Yes, but no often (2) or Yes, every week (3).
	Tea	Whether or not you drink tea in the past year? Never or almost never (1) or occasional drinking (2) or drinking tea often in the past (3) or drinking tea often now (4)
	Carbonated beverages	Whether or not you drink carbonated beverages? Yes (1) or No (2)
	Coffee	Whether or not you drink coffee? Yes (1) or No (2)
	Hypertension	Whether or not you have hypertension? Yes (1) or No (2)
	Retinopathy	Whether or not you have retinopathy? Yes (1) or No (2)
	Hyperlipidemia	Whether or not you have hyperlipidemia? Yes (1) or No (2)
	FLD	Whether or not you have fatty liver? Yes (1) or No (2)
	Snore	Whether or not you snore? Often (1) or Occasionally (2) or Never (3) or Unclear (4)
	Hypotensive Drugs	Whether or not to take hypotensive drugs today? Yes (1) or No (2)
	SBP	Systolic pressure
	DBP	Diastolic pressure
	BMI	Body mass index is a commonly used standard to measure the degree of fat and thinness of the human body and whether it is healthy.
	WC	Waist circumference
	HC	Hip circumference
Blood test data	CRP	C-reactive protein
	HDL	High-density lipoprotein
	LDL	LDL
	TCHO	Total cholesterol.
	TG	Triglyceride.
	AST	Aspartate aminotransferase
	γ-GT	γ-glutamyl transpeptidase
	FBG	Fasting plasma glucose
	P2hPG	Blood glucose 2 h after meals
	HbA1C	Glycated hemoglobin
	FINS	Fasting insulin

hypotheses using the Pima Indians dataset. As shown in Table IV, the integrated model's detection of the Pima Indian dataset improved by about 18% after the imbalance treatment. We conclude that data imbalance affects the performance of the stacked fusion model.

Despite increasing knowledge regarding risk factors for type 2 diabetes and evidence for successful prevention programs, the incidence and prevalence of the disease continue to rise globally.[42] How to design screening programs for early detection and safe and effective treatment will be a key issue in reducing diabetes morbidity and mortality. Notably, the feature importance analysis is an important way to study the factors that influence diabetes in the early stages. To rule out the possible randomness of diabetes factors to model predictions, we divided the dataset and visualized feature importance. In our study, substantial contributions of HbA1c, FBG, Ph2BG, Age, Tea, Han, Carbonate Beverages, AST, etc., were made to the prediction model.

We used statistical methods to verify the importance characteristics of the model, and the results showed that Age, Carbonate Beverages, Han, FBG, Ph2BG, and HbA1c were all risk factors for diabetes mellitus ($OR > 1$) and were statistically significant ($P < 0.05$). Tea drinking is not statistically significant in our data, but many references to polyphenolic compounds in tea can effectively inhibit diabetes,[43,44] which is also consistent with the results shown that the tea drinking has an impact on diabetes by our model. HbA1c concentration is a stable diagnostic measure for type 2 diabetes.[45,46] However, it is not available in all regions. In developing countries, fasting plasma glucose and HbA1c concentrations are inconsistent across ethnicities and with age.[47] This makes sense to look for early diagnostic factors that trigger type 2 diabetes in different regions and ethnicities. As shown in the present study, HbA1c concentration occupies the most important position in the model as a reliable diagnostic indicator for diagnosing type 2

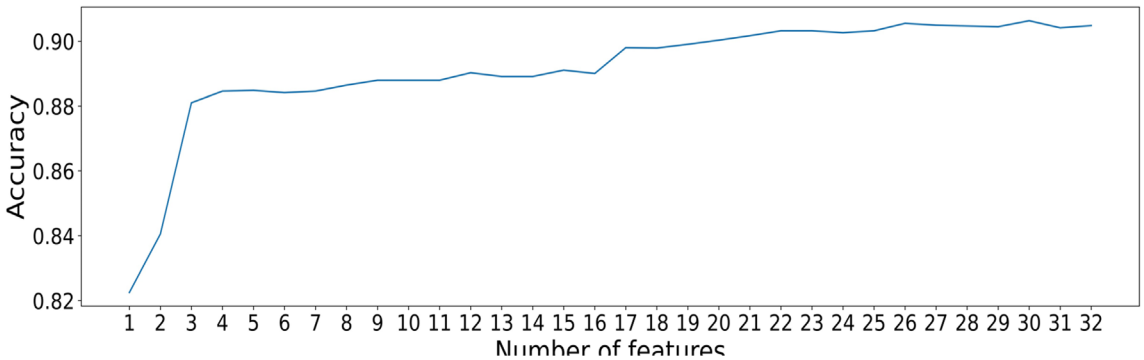

FIG. 2. Wrapper (RFE) curve: the X-axis represents the model accuracy and the Y-axis represents the number of features. The accuracy of the model increases as the number of features increases.

TABLE III. Evaluation attributes results for different models.

Learner	Test set accuracy	Precision (weighted)	Recall (weighted)	F1-score (weighted)	AUC	AP
RF	0.82	0.84	0.82	0.82	0.89	0.83
Extra-trees	0.81	0.83	0.81	0.82	0.88	0.81
GBDT	0.90	0.90	0.89	0.89	0.90	0.83
AdaBoost	0.88	0.88	0.88	0.87	0.88	0.81
Stacking	0.90	0.91	0.90	0.90	0.90	0.85

diabetes, followed by FBG and blood sugar 2 h after a meal. Studies have shown that individuals with a higher blood glucose would have a greater likelihood of developing diabetes.[31,38] These features can also be used as diagnostic indicators for type 2 diabetes detection.[46] However, in our study, FBG was shown to be more reliable in diagnosing type 2 diabetes than blood sugar 2 h after a meal. AST also occupies a certain position, but there are no relevant studies that show an association between AST and diabetes.

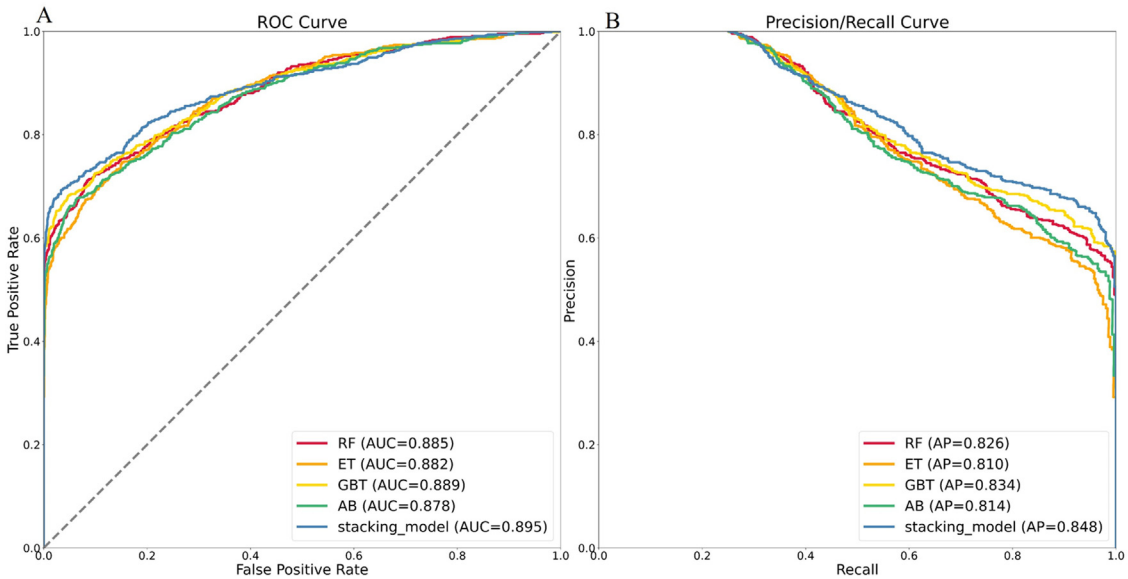

FIG. 3. ROC curves and PR curves. (a) The overall reliability of RF, GBDT, Extra-Trees, AdaBoost, and stacking model. (b) Prediction performance of RF, GBDT, Extra-Trees, AdaBoost, and stacking model. AUC = area under curve; RF = random forest; Extra-Trees = extremely randomized trees; and GBDT = gradient boosting decision tree.

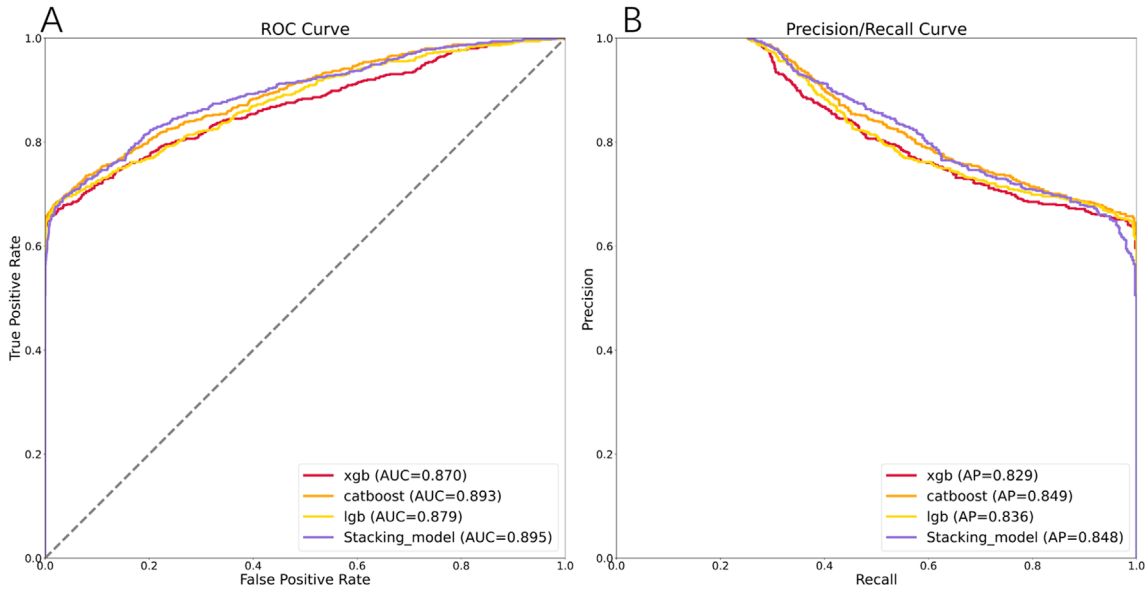

FIG. 4. ROC curves and PR curves. (a) The overall reliability of XGBoost, CatBoost, LightGBM, and stacking model. (b) Prediction performance of XGBoost, CatBoost, LightGBM, and stacking model. AUC = area under curve; xgb = XGBoost; catboost = CatBoost; and lgb = LightGBM.

Obesity and age are important factors in diabetes.[42] We plotted the distribution of the effects of age and BMI on diabetes. As can be seen from Fig. 6, diabetics are concentrated between the ages of 60 and 80. In addition, BMI also increases slightly with age. More diabetic patients have a BMI of more than 30. In our study, another interesting finding was that snoring people are more likely to develop diabetes and the Han population will be more likely to get diabetes than the Zhuang population, which agreed with results from earlier studies.[17,20,48] Our findings further support the views of the above study.

There are limitations in this paper. We only explain the model in terms of feature importance, which is one-sided. Due to data issues, we cannot prove that there is an association between AST and diabetes. In addition, due to the follow-up condition, this study used data from the whole year of 2011, so it may be a bit outdated in time. However, as a chronic disease, diabetes is independent of the timing of the sampled data. Our external validation data also use the already publicly available PIMA dataset. On the other hand, the machine learning framework used in this paper does not incorporate deep learning, which could potentially degrade the performance of the model. Time complexity qualitatively describes the running time of an algorithm and can measure the efficiency of its execution.[49] At present, we do not discuss the time complexity, which will be further supplemented in the future work.

TABLE IV. Evaluation attributes results for different models (PIMA).

Learner	Test set accuracy	Precision (weighted)	Recall (weighted)	F1-score (weighted)
RF	0.72	0.73	0.72	0.72
Extra-trees	0.70	0.71	0.70	0.70
GBDT	0.72	0.72	0.72	0.72
AdaBoost	0.72	0.72	0.72	0.72
Stacking	0.74	0.73	0.74	0.73

In future research, we will explore the robustness of the model against adversarial attacks. In addition, the integration of machine learning algorithms and deep learning models still deserves further exploration by tuning the parameters of the model pairs, including learning rate, batch size, and network structure, in order to find the optimal model configuration. With the continuous updating of data, the models will be continuously trained and updated to adapt to new data distributions and task requirements.

V. CONCLUSIONS

In this retrospective study, we propose a model fusion framework to analyze a cross-sectional dataset of 8981 cases from the First Affiliated Hospital of Guangxi Medical University. Data problems are solved by feature engineering and data preprocessing methods, hyperparameters are optimized using learning curves and grid search, and model performance is evaluated using cross-validation and medical statistical methods. This paper compares other machine learning models with the fusion model, and the results demonstrate that the fusion model outperforms any constructed individual machine learning method and integrated model. The effect of data imbalance handling and feature selection on the model was tested. This study examined the effect of data preprocessing on the model using the PIMA database to demonstrate the robustness of the model. Statistical tests were performed to verify that the proposed model has better generalization. Using feature importance visualization, the reliability of the fusion model is demonstrated and the potential value of other features for diabetes diagnosis is explored. The model fusion framework proposed in this paper can provide assistance in diabetes detection and prevention. In addition, the fusion model can be used in applications or websites to help early warning of diabetic patients.

TABLE V. Model performance comparison in four experiments.

	Learner	Test set accuracy	Precision (weighted)	Recall (weighted)	F1-score (weighted)
Experiment 1 (questionnaire Data)	RF	0.69	0.75	0.69	0.70
	Extra-trees	0.68	0.74	0.68	0.69
	GBDT	0.81	0.81	0.81	0.77
	AdaBoost	0.80	0.80	0.80	0.77
	Stacking	0.81	0.81	0.81	0.77
Experiment 2 (blood test data)	RF	0.72	0.80	0.72	0.74
	Extra-trees	0.72	0.81	0.72	0.74
	GBDT	0.88	0.89	0.88	0.87
	AdaBoost	0.86	0.86	0.86	0.85
	Stacking	0.89	0.89	0.89	0.88
Experiment 3 (remove HbA1C)	RF	0.82	0.83	0.82	0.82
	Extra-trees	0.80	0.82	0.80	0.80
	GBDT	0.89	0.89	0.89	0.88
	AdaBoost	0.87	0.87	0.87	0.87
	Stacking	0.89	0.89	0.89	0.88
Experiment 4 (all data)	RF	0.82	0.84	0.82	0.82
	Extra-trees	0.81	0.83	0.81	0.82
	GBDT	0.89	0.90	0.89	0.89
	AdaBoost	0.88	0.88	0.88	0.87
	Stacking	0.90	0.91	0.90	0.90

VI. METHODS

In this work, we built a two-layer stacking model and demonstrated that the stacking model has more advantages in obtaining a higher predictive accuracy for type 2 diabetes prediction. Figure 1 shows the learning process of the whole study. It includes dataset selection, data preprocessing (data cleaning, class imbalance learning, and feature selection), and model selection. The details of the work are described in the following.

A. Data source

In this retrospective cohort study, the raw data were derived from the Endocrine Department of the First Affiliated Hospital of Guangxi Medical University. Ethical approval was granted by the Ethics Committee of the First Affiliated Hospital of Guangxi Medical University with grant number 2011–14. Data samples were reviewed, and samples containing unreasonable values were removed based on medical criteria. However, the samples with an overly high value in the blood test were not treated since these outliers belong to valid patients. In addition, samples with too many missing features ($\geq$12 features) in a single sample are also deleted. The features in the dataset have been carefully selected based on the available variables in our dataset, clinical expertise, and prior literature evidence of their associations with T2DM. A dataset containing 8981 samples was finally obtained. It contains 30 unique features, where 1596 were diagnosed with T2DM, and 7385 with non-diabetic. Table II outlines the description of the database attributes used in this study. A detailed statistical description of the nominal characteristics and a statistical analysis of the numerical attributes presented in the diabetes dataset, including missing values, centralized trend measures, standard deviations, minimum values, and maximum values in Appendix Tables VIII and IX.

The Pima Indian dataset was used in the external validation. The Pima Indian dataset was downloaded from Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database) and is available via a CC0 public domain license. The dataset is properly anonymized and does not contain any identifiable features of the subjects. This dataset comprised 768 samples, including 500 patients without diabetes and 268 patients with diabetes, as well as their eight characteristics and corresponding classifications.

B. Data preprocessing

Substantial instances of missing data are a serious problem that undermines the scientific credibility of causal conclusions from clinical trials.[50] This study built a random forest regression model to fill in missing values.

Class imbalance is naturally inherent in many real-world applications. Treatment methods of unbalanced data have an important impact on the model performance.[51,52] Since the categories of the incident T2DM in the dataset were imbalanced, the SMOTEENN[23,53] was applied to the training set to resolve the effect of class imbalance. Synthetic minority oversampling technique (SMOTE)[54] is used to analyze minority samples and synthesize new samples based on minority samples to add to the dataset. Edited nearest neighbors (ENN)[23] test each instance with k-NN against the remaining samples in this method. Those incorrectly classified will be discarded, and the remaining samples will form the edited dataset. The hyperparameter sampling strategy is set to 0.3, and the ratio is obtained after many repeated experiments.

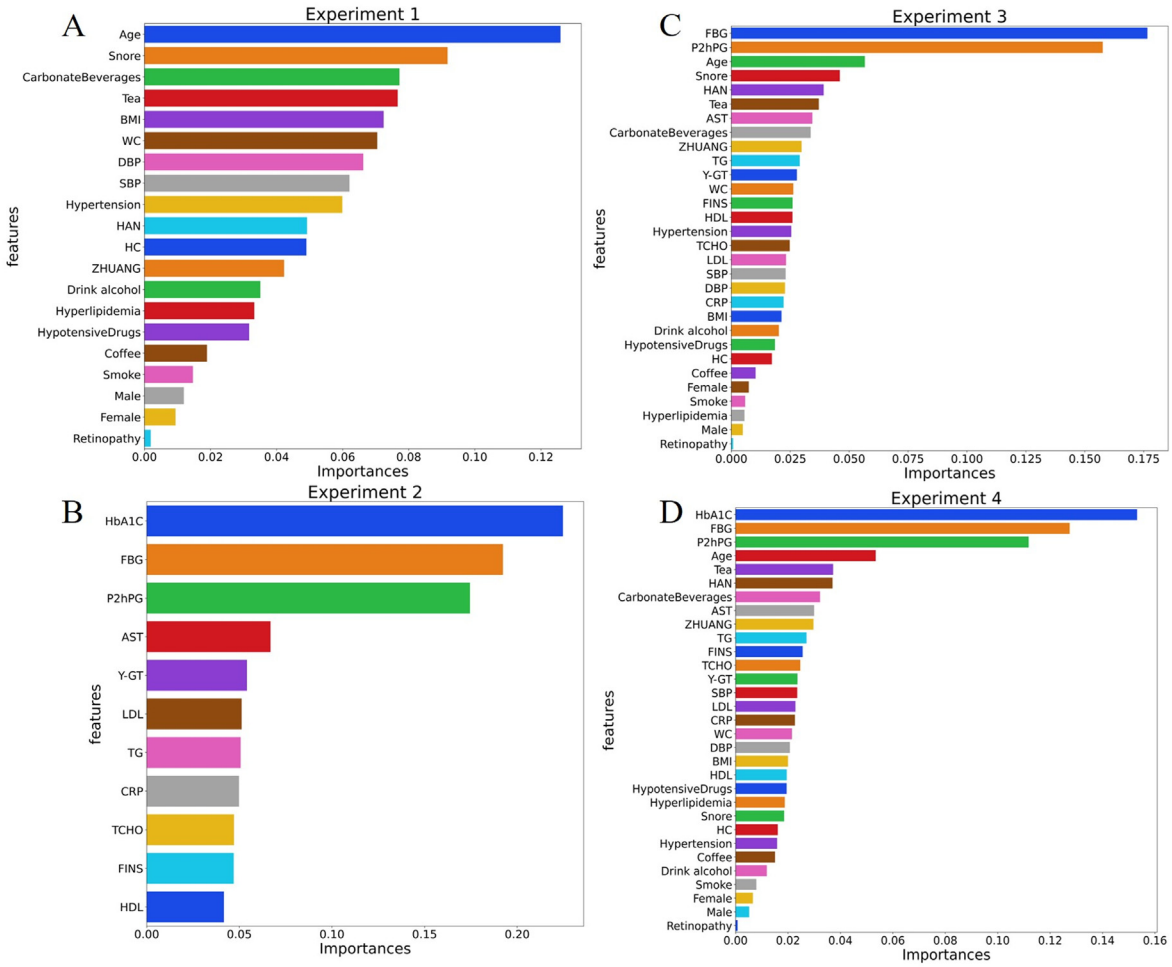

FIG. 5. The feature importance ranking. (a) Permutation feature importance of questionnaire data. (b) Permutation feature importance of blood test data. (c) Permutation feature importance of the comprehensive dataset (remove the HbA1C). (d) Permutation feature importance for the complete dataset.

Using positively or negatively correlated features will cause data redundancy, reduce the accuracy of the model, and increase the computational cost.[55] Feature selection is an important task in data mining and ML applications. It removes irrelevant and redundant features to improve model learning performance.[56]

Wrapper is a method of feature selection that predicts the effect score based on the objective function. It generally finds better feature subset classification performance and relatively high accuracy compared to other feature selection methods. This study set the objective function as recursive feature elimination (RFE).[57] The learning curve of the wrapper is plotted as shown in Fig. 2.

C. Model development

Model fusion refers to building and combining multiple well-performing learners to accomplish a learning task. Different models have their own strengths and differences, and model fusion can make it possible to utilize the strengths of each model, so that these relatively weak learners can be combined by some strategy to achieve a relatively strong learner. Model fusion is derived from, but superior to, model integration. The main difference is that model fusion uses better performing learners, while model integration uses learners from multiple bases, so the training bases are different. From a statistical point of view, model fusion reduces the risk caused by choosing the wrong assumptions, improves the likelihood of capturing real data patterns, and improves the likelihood of having better generalization capabilities. There are two main integration learning methods, boosting and bagging. Stacking combines these two integration methods by utilizing multiple primary learners on the raw data and then sending the features learned by the primary learners to the meta-learner for fitting.

In this study, we built a two-story stacking model. Random forest (RF), Extra-Trees(ET), GradientBoosting (GBDT), and AdaBoost are basic learning algorithms as the first layer, whereas RF is the meta-learner as the second layer. Figure 7 outlines the specific model structure. To avoid overfitting, we use fivefold cross-validation and set random seeds. This study uses the setup learning curve and the GridSearchCV hyper-parameter tuning method to find the best hyper-parameters. The GridSearchCV hyper-parameter tuning method will

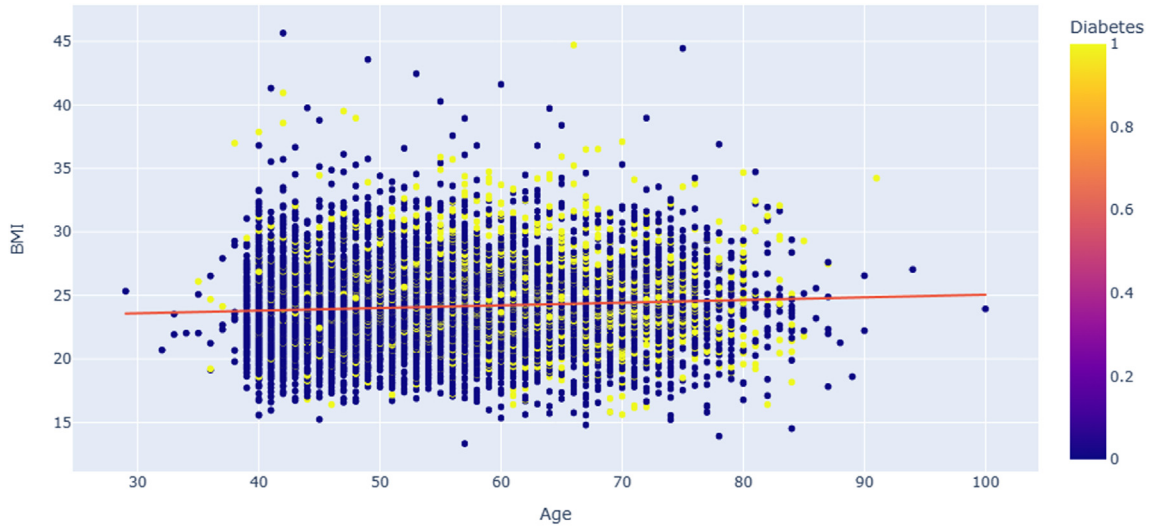

FIG. 6. Age, BMI, and diabetes trends scatterplot. The X-axis is the age indicator, and the Y-axis is the BMI indicator. The label indicates whether the sample has diabetes. Yellow represents diabetic patients, and purple represents non-diabetic participants.

TABLE VI. Univariate logistic regression in the differential diagnosis of diabetes in the feature cohort.

Variables		OR	95%CI	P-value
Age		1.053	(1.047,1.058)	<0.001
Han	(No)	1 (reference)		
	(Yes)	1.799	(1.586,2.041)	<0.001
Tea	(Never)	1 (reference)		
	(Occasional)	0.907	(0.800,1.029)	0.129
	(Used to drink tea)	0.972	(0.602,1.568)	0.906
	(Now often)	0.901	(0.780,1.039)	0.152
Carbonated Beverages	(No)	1 (reference)		
	(Yes)	1.913	(1.660,2.204)	<0.001
AST		1.015	(1.011,1.020)	<0.001
FBG		2.708	(2.547,2.879)	<0.001
P2hPG		1.450	(1.470,1.480)	<0.001
HbA1c		5.194	(4.703,5.736)	<0.001

loop through all candidate parameter selections, trying every possibility to find the best performing hyperparameters. To ensure the stability of the hyperparameters, we chose tenfold cross-validation.

D. Model evaluation

Model performance was evaluated on a test set using accuracy, precision, recall, F1-score, P-R curve, and AUC as model evaluation criteria. Each evaluation method was based on one of four categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

True Positive (TP): predicts positive classes as positive class numbers.
True Negative (TN): predicts a negative class as a negative number.
False Positive (FP): predicts a negative class as a positive class (type 1 error).
False Negative (FN): predicts the positive class as the number of negative classes (type 2 error).

Accuracy: responds to the percentage of samples correctly categorized by the model and is one of the most intuitive metrics to evaluate. However, in unbalanced datasets, accuracy can be misleading because the model may tend to predict the majority of categories while ignoring a few.

$$\text{Accuracy: accuracy} = ((\text{TP} + \text{TN}))/((\text{TP} + \text{TN} + \text{FP} + \text{FN})). \quad (1)$$

TABLE VII. Multivariate logistic regression in the differential diagnosis of diabetes in the feature cohort.

Variables		OR	95%CI	P-value
Age		1.018	(1.011,1.024)	<0.001
Han	(No)	1 (reference)		
	(Yes)	1.484	(1.263,1.743)	<0.001
Carbonated beverages	(No)	1 (reference)		
	(Yes)	1.347	(1.129,1.607)	0.001
AST		1.004	(0.997,1.011)	0.236
FBG		1.586	(1.485,1.693)	<0.001
P2hPG		1.190	(1.159,1.222)	<0.001
HbA1c		2.203	(1.916,2.469)	<0.001

Precision: measures the proportion of true instances in the sample that the model categorizes as positive instances. The level of precision reflects how accurately the model predicts positive cases and is especially important when the cost of false positive cases (false alarms) is high.

$$\text{Precision: precision} = \text{TP}/((\text{TP} + \text{FP})). \tag{2}$$

Recall: measures the model's ability to identify samples of positive examples, i.e., how many true positive examples the model is able to correctly capture. Recall is a key metric in the context of concerns about missing true positive examples, especially when the cost of false negative examples (underreporting) is high.

$$\text{Recall: recall} = \text{TP}/((\text{TP} + \text{FN})). \tag{3}$$

F1-score: combining precision and recall, it is a metric that takes into account both the predictive accuracy and recognition ability of the model. It is particularly useful for the evaluation of unbalanced datasets because it balances the trade-off between precision and recall.

$$\text{F1-score: F1} - \text{score} = \text{TP}/((\text{TP} + \text{FP})). \tag{4}$$

PR Curve (Precision–Recall Curve): the PR curve demonstrates the trade-off between precision and recall of the model at different thresholds. By analyzing the PR curve, the optimal model threshold can be determined and the performance of the model under different thresholds can be understood.

AUC (Area Under the Curve): AUC provides a single value for comparing the performance of different models, with higher AUC indicating better model performance in classification tasks.

E. Model explanation

In this paper, the model is built using stacking and compared to the RF, ET, GBDT, and AdaBoost models. Table III presents the performance metric scores of different methods.

To verify the robustness of the model, external validation is set up. In this paper, the model is tested using the Pima Indian dataset. Table IV shows the performance of stacking models in the datasets.

To validate the reliability of the model and what are the key factors in diagnosing diabetes with different characteristics, in this study, the processed data are divided into four parts for experiments: experiment one: using the model to train and predict the questionnaire data in the dataset; experiment two: using the model to train and predict the blood test data in the dataset; in the third experiment, the model was trained and predicted using the dataset after excluding the gold standard for detecting diabetes (HbA1C); and experiment four uses all the data for model training. To show the impact of model performance and features on diabetes, the model performance was compared for different datasets. Table V outlines the details. In order to evaluate the practical significance of the model, a visual interpretation of the model was performed. The ranking of the importance of the features shows the risk factors that are most relevant to the impact of diabetes.

Characteristics were assessed using one-way logistic regression and multifactor logistic regression. Logistic regression analysis was applied to calculate the odds ratio (OR) with 95% confidence interval

TABLE VIII. Numerical attributes statistical description.

Feature	State	Count	Mean	Std	Min	Max
Age	NonT2DM	7385	54.71	112.174	29	100
	T2DM	1596	60.77	106.321	35	91
SBP	NonT2DM	7277	131.37	427.896	66	216
	T2DM	1575	138.58	437.075	70	228
DBP	NonT2DM	7277	78.53	144.965	33	139
	T2DM	1575	79.69	142.326	49	125
BMI	NonT2DM	7231	23.9897	10.550	13.333	45.6538
	T2DM	1568	24.8180	12.276	15.623	44.7087
WC	NonT2DM	7218	82.442	80.630	53.0	125.0
	T2DM	1560	85.666	81.454	58.0	130.2
HC	NonT2DM	7181	94.002	43.893	59.2	140.0
	T2DM	1550	95.398	50.378	70.0	133.0
CRP	NonT2DM	7310	69.205	331.450	19.6	358.7
	T2DM	1583	72.544	305.577	19.8	259.1
HDL	NonT2DM	7311	1.3039	0.170	0.13	3.00
	T2DM	1583	1.3246	0.147	0.31	2.77
LDL	NonT2DM	7310	2.9151	0.995	0.18	10.46
	T2DM	1583	3.1779	0.934	0.62	11.93
TCHO	NonT2DM	7311	4.9387	1.957	0.40	13.03
	T2DM	1582	5.3598	1.614	1.39	12.17
TG	NonT2DM	7304	1.4754	1.402	0.10	15.59
	T2DM	1579	1.8708	2.352	0.25	14.76
AST	NonT2DM	7288	19.60	106.358	3	254
	T2DM	1583	21.43	98.971	3	136
YGT	NonT2DM	7265	27.21	833.690	4	595
	T2DM	1581	33.97	1668.779	4	768
FBG	NonT2DM	7102	5.51803	0.727	0.110	17.060
	T2DM	1579	7.5948	8.323	1.88	26.95
P2hPG	NonT2DM	7015	6.99696	4.630	3.010	23.200
	T2DM	1554	11.6580	32.563	3.10	33.86
HbA1C	NonT2DM	7218	5.533	0.356	2.7	19.5
	T2DM	1579	6.847	2.952	4.0	15.5
FINS	NonT2DM	7312	8.426	32.041	0.1	163.9
	T2DM	1582	11.283	185.873	0.4	238.7

TABLE IX. Nominal attributes statistical analysis.

Feature	State	Count	Values (count)
Gender	NonT2DM	7385	Male (2850), Female (4535)
	T2DM	1596	Male (575), Female (1021)
Race	NonT2DM	7183	Han (4609), Zhuang (2487), other (87)
	T2DM	1552	Han (1186), Zhuang (341), other (25)
Smoke	NonT2DM	7062	Non-smoke (5935), occasionally (264), smoke (863)
	T2DM	1509	Non-smoke (1327), Occasionally (32), smoke (150)
Drink alcohol	NonT2DM	7113	Non-drink alcohol (4654) occasionally (1780), drink alcohol (679)
	T2DM	1533	Non-drink alcohol (1122) occasionally (294), drink alcohol (117)
Tea	NonT2DM	7307	Non-tea (3404) occasional (2262) often (89) drinking tea (1552)
	T2DM	1580	Non-tea (774) occasional (468) often (20) drinking tea (318)
Carbonate beverages	NonT2DM	7059	Carbonated beverages (1704), non-carbonated beverages (5355)
	T2DM	1536	Carbonated beverages (204), non-carbonated beverages (1332)
Coffee	NonT2DM	7046	Coffee (643), non- coffee (6403)
	T2DM	1530	Coffee (102), non- coffee (1428)
Hypertension	NonT2DM	7324	Hypertension (1291), non- hypertension (6033)
	T2DM	1590	Hypertension (494), non- hypertension (1096)
Retinopathy	NonT2DM	7324	Retinopathy (24), non-retinopathy (7289)
	T2DM	1587	Retinopathy (19), non-retinopathy (1566)
Hyperlipidemia	NonT2DM	7050	Hyperlipidemia (735), non-hyperlipidemia (6580)
	T2DM	1510	Hyperlipidemia (274), non-hyperlipidemia (1312)
FLD	NonT2DM	7176	Fatty liver (768), non-fatty liver (6556)
	T2DM	1546	Fatty liver (249), non-fatty liver (1338)
Snore	NonT2DM	7324	Often (1328) or occasionally (2569) or never (2402) or unclear (751).
	T2DM	1587	Often (372) or occasionally (521) or never (468) or unclear (149).
Hypotensive drugs	NonT2DM	7050	Hypotensive drugs (317), non-hypotensive drugs (6859)
	T2DM	1510	Hypotensive drugs (164), non-hypotensive drugs (1382)

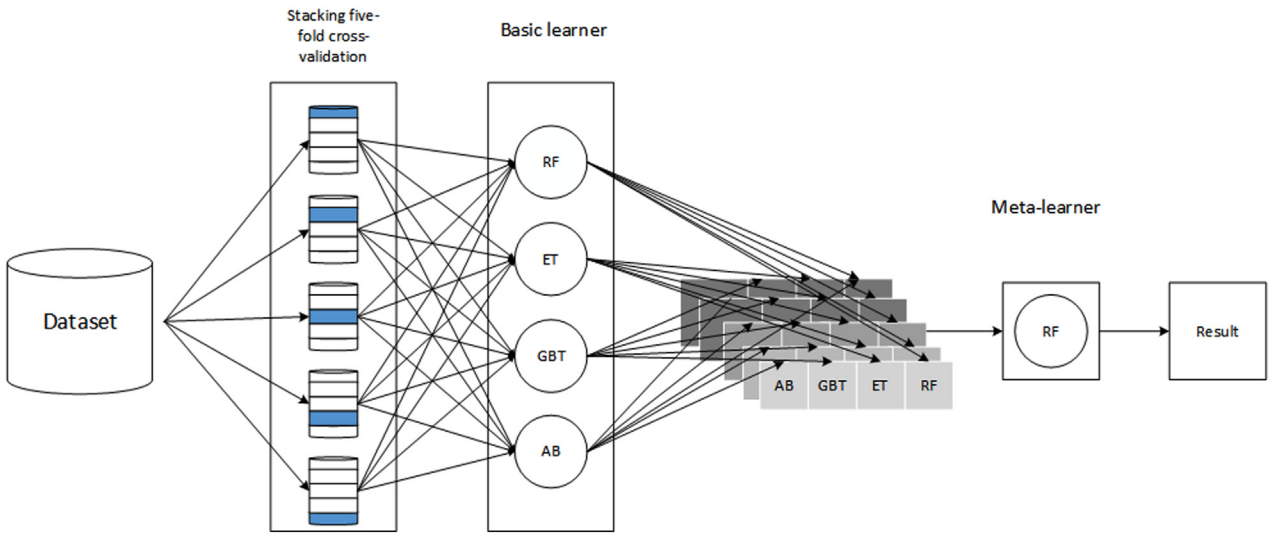

FIG. 7. Model fusion flow chart. Using stacking model fusion, RF, GBDT, ET, and AdaBoost were used as the primary learners for model fusion. The data were subjected to internal fivefold cross-validation using stacking, and the characteristics of each learner output in the training set were extracted and fed into the model fusion meta-learner. Model classification prediction was finally achieved through the data training and learning of the meta-learner.

(CI). $P < 0.05$ was considered to indicate statistical significance. The results are shown in Tables VI and VII. To explore the impact of certain characteristics on diabetes, we present trends in the form of scatter plots.

SUPPLEMENTARY MATERIAL

See the supplementary material for the following: "an epidemiological investigation project on the risk of complications in patients with diabetes mellitus in the community baseline survey informed consent" and "epidemiological study of the risk of complications in patients with type 2 diabetes mellitus in China baseline questionnaire."

ACKNOWLEDGMENTS

We thank all participants who agreed to participate in this study.

This work was support in part by the National Natural Science Foundation of China (No. 62341601,81860604); the National Key Research & Development Plan for Precision Medicine Key Program (Nos. 2016YFC0901200 and 2016YFC0901205); the National Health Commission of China Public Welfare Research Project (No. 201502007); the Innovation Project of Clinical Research Climbing Plan of the First Affiliated Hospital of Guangxi Medical University (No. YYZS2020012); and the Guangxi Medical and Health Appropriate Technology Development and Promotion Application Project (No. S2017026).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

Ethics approval for experiments reported in the submitted manuscript on animal or human subjects was granted. All patients have signed an informed consent form. Authors reporting experiments on humans and all experiments were performed in accordance with relevant guidelines and regulations. The study was approved by Medical Ethics Committee of Ruijin Hospital, Shanghai Jiaotong University (approval number: 2011–14).

Author Contributions

Yong Fu, Xinghuan Liang and Xi Yang contributed equally to this work.

Yong Fu: Conceptualization (equal); Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Xinghuan Liang:** Data curation (equal); Investigation (equal); Writing – original draft (equal). **Xi Yang:** Data curation (equal); Investigation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Li Li:** Data curation (equal); Investigation (equal); Resources (equal). **LiHeng Meng:** Data curation (equal); Methodology (equal). **Yuekun Wei:** Data curation (equal). **Daizheng Huang:** Conceptualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Yingfen Qin:** Conceptualization (equal); Funding acquisition (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

NOMENCLATURE

List of Abbreviations

AUC	area under curve
CI	confidence interval
ENN	edited nearest neighbors
ET	extra-trees
GBDT	gradientboosting
KNN	K-nearest neighbor
LR	logistic regression
ML	machine learning
NB	Naïve–Bayes
OR	odds ratio
RF	random forest
RFE	recursive feature elimination
ROC	receiver operating characteristic curve
SMOTE	synthetic minority oversampling technique
SVM	support vector machine
T2DM	type 2 diabetes mellitus

APPENDIX: STATISTICAL ANALYSIS OF NUMERICAL AND NOMINAL ATTRIBUTES IN A DIABETIC POPULATION

Numerical attributes statistical description. Nominal attributes statistical analysis.

REFERENCES

[1]WHO, see https://www.who.int/news-room/fact-sheets/detail/diabetes for "Diabetes WHO diabetes" (2022).
[2]IDF, see https://diabetesatlas.org/ for "Diabetes around the world in 2021 IDF Diabetes Atlas" (2022).
[3]S. Demir, P. P. Nawroth, S. Herzig *et al.*, "Emerging targets in type 2 diabetes and diabetic complications," Adv. Sci. **8**(18), 2100275 (2021).
[4]N. Peer, Y. Balakrishna, and S. Durao, "Screening for type 2 diabetes mellitus," Cochrane Database Syst. Rev. **5**(5), CD005266 (2020).
[5]K. Shaukat, S. Luo, V. Varadharajan *et al.*, "Performance comparison and current challenges of using machine learning techniques in cybersecurity," Energies **13**, 2509 (2020).
[6]K. Shaukat, S. Luo, V. Varadharajan *et al.*, "A survey on machine learning techniques for cyber security in the last decade," IEEE Access **8**, 222310–222354 (2020).
[7]T. M. Alam, K. Shaukat, I. A. Hameed *et al.*, "A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining," Biomed. Signal Process. Control **68**, 102726 (2021).
[8]T. M. Alam, K. Shaukat, A. Khelifi *et al.*, "A fuzzy inference-based decision support system for disease diagnosis," Comput. J. **66**(9), 2169–2180 (2023).
[9]*Computational Methods for Medical and Cyber Security*, edited by S. Luo and K. Shaukat (MDPI Books, 2022).
[10]M. R. Kumar, S. Vekkot, S. Lalitha *et al.*, "Dementia detection from speech using machine learning and deep learning architectures," Sensors **22**(23), 9311 (2022).
[11]A. Siddique, K. Shaukat, and T. Jan, "An intelligent mechanism to detect multifactor skin cancer," Diagnostics **14**(13), 1359 (2024).

[12]Srinivas, CK. S. N. P. and Zakariah, M *et al.*, "Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images," J. Healthcare Eng. **2022**(1), 3264367.
[13]M. G. Alsubaie, S. Luo, and K. Shaukat, "ConvADD: Exploring a novel CNN architecture for Alzheimer's disease detection," Int. J. Adv. Comput. Sci. Appl. **15**(4), 300–313 (2024).
[14]K. Shaukat, S. Luo, and V. Varadharajan, "A novel deep learning-based approach for malware detection," Eng. Appl. Artif. Intell. **122**, 106030 (2023).
[15]K. Shaukat, S. Luo, and V. Varadharajan, "A novel machine learning approach for detecting first-time-appeared malware," Eng. Appl. Artif. Intell. **131**, 107801 (2024).
[16]K. Shaukat, S. Luo, and V. Varadharajan, "A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks," Eng. Appl. Artif. Intell. **116**, 105461 (2022).
[17]F. Anwar, M. Y. Ejaz, and A. Mosavi, "A comparative analysis on diagnosis of diabetes mellitus using different approaches–A survey," Inf. Med. Unlocked **21**, 100482 (2020).
[18]R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," Int. J. Environ. Res. Public Health **18**(14), 7346 (2021).
[19]V. Chang, J. Bailey, Q. A. Xu *et al.*, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Comput. Appl. **35**(22), 16157–16173 (2023).
[20]E. Adua, E. A. Kolog, E. Afrifa-Yamoah *et al.*, "Predictive model and feature importance for early detection of type II diabetes mellitus," Transl. Med. Commun. **6**, 1–15 (2021).
[21]J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," Ict Express **7**(4), 432–439 (2021).
[22]M. O. Edeh, O. I. Khalaf, C. A. Tavera *et al.*, "A classification algorithm-based hybrid diabetes prediction model," Front. Public Health **10**, 829519 (2022).
[23]Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," Prev. Chronic Dis. **16**, E130 (2019).
[24]I. Goodfellow, *Deep Learning* (MIT Press, 2016).
[25]C. Zhang, S. Bengio, M. Hardt *et al.*, "Understanding deep learning (still) requires rethinking generalization," Commun. ACM **64**(3), 107–115 (2021).
[26]M. Khushi, K. Shaukat, T. M. Alam *et al.*, "A comparative performance analysis of data resampling methods on imbalance medical data," IEEE Access **9**, 109960–109975 (2021).
[27]T. M. Alam, K. Shaukat, W. A. Khan *et al.*, "An efficient deep learning-based skin cancer classifier for an imbalanced dataset," Diagnostics **12**(12(9), 2115 (2022).
[28]T. M. Alam, K. Shaukat, H. Mahboob *et al.*, "A machine learning approach for identification of malignant mesothelioma etiological factors in an imbalanced dataset," Comput. J. **65**(7), 1740–1751 (2022).
[29]X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (IEEE, 2020), pp. 1–6.
[30]Z. Ali, M. F. Hayat, K. Shaukat *et al.*, "A proposed framework for early prediction of schistosomiasis," Diagnostics **12**(12), 3138 (2022).
[31]L. Devnath, S. Luo, P. Summons *et al.*, "Deep ensemble learning for the automatic detection of pneumoconiosis in coal worker's chest X-ray radiography," J. Clin. Med. **11**(18), 5342 (2022).
[32]Q. Liu, M. Zhang, Y. He *et al.*, "Predicting the risk of incident type 2 diabetes mellitus in Chinese elderly using machine learning techniques," J. Pers. Med. **12**(12(6), 905 (2022).
[33]H. Yang, Y. Luo, X. Ren *et al.*, "Risk prediction of diabetes: Big data mining with fusion of multifarious physical examination indicators," Inf. Fusion **75**, 140–149 (2021).
[34]M. Xue, Y. Su, C. Li *et al.*, "Identification of potential type II diabetes in a large-scale chinese population using a systematic machine learning framework," J. Diabetes Res. **2020**(1), 6873891.
[35]Z. Dong, Q. Wang, Y. Ke, W. Zhang, Q. Hong, C. Liu *et al.*, "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records," J. Transl. Med. **20**(1), 143 (2022).
[36]D. D. Rufo, T. G. Debelee, A. Ibenthal *et al.*, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," Diagnostics **11**(9), 1714 (2021).
[37]X. Xiong, R. Zhang, Y. Bi *et al.*, "Machine learning models in type 2 diabetes risk prediction: Results from a cross-sectional retrospective study in Chinese adults," Curr. Med. Sci. **39**(4), 582–588 (2019).
[38]A. Sumathi and S. Meganathan, "Ensemble classifier technique to predict gestational diabetes mellitus (GDM)," Comput. Syst. Sci. Eng. **40**(1), 313–325 (2022).
[39]H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," Int. J. Environ. Res. Public Health **18**(6), 3317 (2021).
[40]M. Gollapalli, A. Alansari, H. Alkhorasani *et al.*, "A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM," Comput. Biol. Med. **147**, 105757 (2022).
[41]S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," Comput. Biol. Med. **135**, 104554 (2021).
[42]S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," Lancet **389**(10085), 2239–2251 (2017).
[43]H. Cao, J. Ou, L. Chen, Y. Zhang, T. Szkudelski, D. Delmas *et al.*, "Dietary polyphenols and type 2 diabetes: Human study and clinical trial," Crit. Rev. Food Sci. Nutr. **59**(20), 3371–3379 (2019).
[44]W. Koch, "Dietary Polyphenols—Important non-nutrients in the prevention of chronic noncommunicable diseases. A systematic review," Nutrients **11**(5), 1039 (2019).
[45]S. Colagiuri, C. M. Lee, T. Y. Wong, B. Balkau, J. E. Shaw, K. Borch-Johnsen *et al.*, "Glycemic thresholds for diabetes-specific retinopathy: Implications for diagnostic criteria for diabetes," Diabetes Care **34**(1), 145–150 (2011).
[46]Chinese Elderly Type 2 Diabetes Prevention and Treatment of Clinical Guidelines Writing Group; Geriatric Endocrinology and Metabolism Branch of Chinese Geriatric Society; Geriatric Endocrinology and Metabolism Branch of Chinese Geriatric Health Care Society; Geriatric Professional Committee of Beijing Medical Award Foundation; National Clinical Medical Research Center for Geriatric Diseases (PLA General Hospital), Zhonghua Nei Ke Za Zhi. **61**(1), 12–50 (2022).
[47]K. J. Welsh, M. S. Kirkman, and D. B. Sacks, "Role of glycated proteins in the diagnosis and management of diabetes: Research gaps and future directions," Diabetes Care **39**(8), 1299 (2016).
[48]L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck, "Type 2 diabetes with artificial intelligence machine learning: Methods and evaluation," Arch. Comput. Methods Eng. **29**(1), 313–333 (2022).
[49]K. Shaukat, S. Luo, S. Chen *et al.*, "Cyber threat detection using machine learning techniques: A performance evaluation perspective," in *2020 International Conference on Cyber Warfare and Security (ICCWS)* (IEEE, 2020), pp. 1–6.
[50]R. J. Little, R. D'agostino, M. L. Cohen *et al.*, "The prevention and treatment of missing data in clinical trials," N. Engl. J. Med. **367**(14), 1355–1360 (2012).
[51]S. C. K. Tékouabou, I. Chabbar, H. Toulni *et al.*, "Optimizing the early glaucoma detection from visual fields by combining preprocessing techniques and ensemble classifier with selection strategies," Expert Syst. Appl. **189**, 115975 (2022).
[52]X. Chen, C. Faviez, M. Vincent *et al.*, "Patient-patient similarity-based screening of a clinical data warehouse to support ciliopathy diagnosis," Front. Pharmacol. **13**, 786710 (2022).
[53]G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," SIGKDD Explor. Newsl. **6**(1), 20–29 (2004).
[54]S. Maldonado, C. Vairetti, A. Fernandez *et al.*, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," Pattern Recognit. **124**, 108511 (2022).
[55]J. Han, J. Pei, and H. Tong, "Data mining: Concepts and techniques" (published online) (2022).
[56]R. Sheikhpour, M. A. Sarram, S. Gharaghani *et al.*, "A survey on semi-supervised feature selection methods," Pattern Recognit. **64**, 141–158 (2017).
[57]W. Liu and J. Wang, "Recursive elimination–election algorithms for wrapper feature selection," Appl. Soft Comput. **113**, 107956 (2021).