Research paper

Machine learning predicts diabetes risk in high-risk populations: based on the National Health and Nutrition Examination Survey database

Keywords
diabetes, machine learning, National Health and Nutrition Examination Survey, prediction model

Abstract

Introduction

This project intended to develop and validate a diabetes prediction model for high-risk populations based on machine learning algorithms.

Material and methods

A total of 2,355 samples from the National Health and Nutrition Examination Survey (NHANES) database covering three cycles from 2013 to 2018 were included. The data were divided into training and testing sets in a 7:3 ratio. Nineteen risk prediction factors were selected as feature variables, including demographic baseline data, measurement data, medical history, and psychological health. Five machine learning models, including decision tree, random forest (RF), multilayer perceptron (MLP), Adaboost, and XGBoost,

Results

The present work ultimately included 2,355 individuals at high risk of diabetes for analysis, with 260 cases of diabetes and 2,095 cases without diabetes. Among the five machine learning models established in this project, the RF and XGBoost models exhibited better overall performance compared to other models. In the test set, the RF model had an AUC of 0.896, accuracy of 0.784, sensitivity of 0.739, specificity of 0.849, and MCC of 0.418. The XGBoost model had corresponding values of AUC as 0.903, accuracy of 0.815, sensitivity of 0.962, and MCC of 0.443. According to the importance analysis of features in these two optimal models, waist circumference, age, BMI, gender

Conclusions

The RF and XGBoost models in machine learning demonstrate good performance in predicting the occurrence of diabetes in high-risk populations, which can aid in developing more precise intervention measures and personalized treatment plans to effectively reduce the incidence of diabetes and related risks in this population.

Machine learning predicts diabetes risk in high-risk populations: based on the National Health and Nutrition Examination Survey database

Short title: Machine learning predicts diabetes risk in high-risk populations

Authors:

Xiaohua Yang[1], Meiqi Yao[1], Jia Huang[2], Zhuojing Cheng[2], Ting Sun[2, *]

Affiliation:

1Nursing department, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, 310009, China

2Physical examination center, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, 310009, China

***Corresponding Author Information:**

Ting Sun, Physical examination center, The Second Affiliated Hospital of Zhejiang University School of Medicine, 88 Jiefang Road, Shangcheng District, Hangzhou, 310009, China.

Tel: 13906503158, Email: 1195037@zju.edu.cn

Abstract

Objectives: This project intended to develop and validate a diabetes prediction model for high-risk populations based on machine learning algorithms.

Method: A total of 2,355 samples from the National Health and Nutrition Examination Survey (NHANES) database covering three cycles from 2013 to 2018 were included. The data were divided into training and testing sets in a 7:3 ratio. Nineteen risk prediction factors were selected as feature variables, including demographic baseline data, measurement data, medical history, and psychological health. Five machine learning models, including decision tree, random forest (RF), multilayer perceptron (MLP), Adaboost, and Extreme Gradient Boosting (XGBoost), were developed based on the data and variables mentioned above. Model performance was evaluated using accuracy, sensitivity, specificity, the Area Under Curve (AUC) values of receiver operating characteristic (ROC) curves, and Matthews Correlation Coefficient (MCC) scores. Finally, the Shapley feature importance measurement tool was employed to select features in the optimal model.

Results: The present work ultimately included 2,355 individuals at high risk of diabetes for analysis, with 260 cases of diabetes and 2,095 cases without diabetes. Among the five machine learning models established in this project, the RF and XGBoost models exhibited better overall performance compared to other models. In the test set, the RF model had an AUC of 0.896, accuracy of 0.784, sensitivity of 0.739, specificity of 0.849, and MCC of 0.418. The XGBoost model had corresponding values of AUC as 0.903, accuracy of 0.815, sensitivity of 0.962, and MCC of 0.443. According to the importance analysis of features in these two optimal models, waist circumference, age, BMI, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), education level, poverty income ratio (PIR), Patient Health Questionnaire (PHQ)-9 score, and race were the top ten key risk factors for diabetes in the high-risk population.

Conclusion: The RF and XGBoost models in machine learning demonstrate good performance in predicting the occurrence of diabetes in high-risk populations, which can aid in developing more precise intervention measures and personalized treatment plans to effectively reduce the incidence of diabetes and related risks in this population.

Keywords: diabetes; machine learning; National Health and Nutrition Examination Survey; prediction model

1 Introduction

Diabetes is a chronic metabolic disorder characterized by abnormal blood sugar levels, caused by improper use of insulin or insufficient insulin secretion, leading to severe long-term damage to multiple organs and body systems (including kidneys, heart, nerves, blood vessels, and eyes)[1], ultimately becoming a major contributor to death[2]. On a global scale, diabetes has become a daunting public health challenge, with its prevalence continuing to rise[3]. From 2021 to 2050, the global burden of diabetes will rise from 529 million people to 1.31 billion people[4]. Despite the projected dramatic increase in the future diabetic population, high-risk individuals often underestimate their own risk of developing the disease[5, 6]. Once high-risk individuals progress to confirmed cases, although there are treatment methods available to slow down the progression of the disease, there is still a lack of curative treatment options[7]. Considering the high prevalence of diabetes and the relatively large size of the high-risk population[8], it is crucial at a population level to further identify risk factors and take preventive measures before high-risk individuals develop the disease state[9].

In recent years, the application of machine learning techniques in the medical field has become increasingly widespread, especially in the prediction of disease risks and diagnosis, exhibiting potential value[10-12]. Utilizing rich clinical data and advanced algorithms, machine learning studies based on large-scale databases have become a hot topic in diabetes research, aiding in identifying individuals at risk of diabetes and providing personalized prevention and management strategies[13, 14]. For example, a project used five machine learning models, including Logistic Regression, Support Vector Machine, Random Forest (RF), Extreme Gradient Boosted Tree, and Weighted Voting Classifier to predict diabetes in adolescents and identify factors leading to diabetes in adolescents, such as waist circumference, gender, BMI, and leg length[15]. However, research on the risk assessment of diabetes in high-risk populations has not been fully developed yet.

Therefore, this project utilized the National Health and Nutrition Examination Survey (NHANES) database with a large sample size to construct a diabetes risk prediction model for high-risk populations using machine learning techniques. The model was designed to early identify individuals at high risk of diabetes in high-risk populations and to take timely prevention and treatment measures. This work is of great significance for reducing the incidence of diabetes and related complications.

2 Methods

2.1 Data source and study population

This investigation conducted data analysis using the NHANES public database, which was established and continuously updated and improved by the Centers for Disease Control and Prevention (CDC) in the United States. The NHANES employs a layered, multi-stage probability sampling method to select a nationally representative sample of the population, and collects data through direct physical examinations, clinical and laboratory tests, personal interviews, and relevant measurement procedures. Relevant questionnaires and study protocols can be obtained from the NHANES official webpage on the CDC website[16]. The NHANES has obtained ethical approval from the National Health Statistics Research Ethics Review Committee in the United States, and all participants have signed informed consent forms to ensure that they understand and agree to participate in the survey.

2.2 Participants

In this project, we selected a sample of individuals at high risk of diabetes from 29,400 participants in the NHANES from 2013 to 2018. The definition criteria for high-risk groups of diabetic patients were those who meet any of the following conditions: (1) age ≥ 40 years old; (2) impaired glucose tolerance (fasting glucose < 126 mg/dL and 140 mg/dL ≤ OGTT (oral glucose tolerance test) < 200 mg/dL)[17] or abnormal fasting glucose (100 mg/dL ≤ fasting glucose < 126 mg/dL)[18]; (3) overweight (BMI ≥ 25 kg/m^2); (4) lack of physical activity (moderate or equivalent intensity activity time < 150 min per week) (5) family history of diabetes; (6) A history of gestational diabetes; (7) High blood pressure or taking antihypertensive drugs; (8) A history of coronary heart disease; (9) patients with polycystic ovarian cancer syndrome; (10) taking antidepressants for more than 3 months and depression diagnosed by ICD-10 coding (F32.9 and F33.9). By applying the above screening criteria, 18,649 individuals at high risk of diabetes were identified. Subsequently, 16,294 samples with missing feature data were excluded, resulting in 2,355 eligible samples that met the criteria. These samples were divided into training and testing sets in a 7:3 ratio for the construction and evaluation of machine learning prediction models. The specific inclusion process is shown in Figure 1.

2.3 Outcome variables

Diabetes as the outcome variable was defined as meeting one of the following criteria: (a) diagnosed with diabetes by a doctor; (b) taking antidiabetic medication; (c) glycosylated hemoglobin HbA1c > 6.5%; (d) fasting blood glucose > 126 mg/dL[19, 20]. In this project, the outcome variable was defined as a categorical variable, encoded in numerical form for binary classification of high-risk individuals: 0 indicating non-diabetes and 1 indicating diabetes. Categorical features were encoded with numerical values for analysis.

2.4 Feature variables

The feature variables in this project included demographic baseline data, measurement data, medical history, and the Patient Health Questionnaire (PHQ)-9. Demographic baseline data included gender, age, race, education level[21], poverty-income ratio (PIR), and family history of diabetes. Measurement data included waist circumference, BMI, diastolic blood pressure (DBP), and systolic blood pressure (SBP). History of disease included heart disease, hypertension, arthritis, cancer, stroke, hearing impairment, vision impairment, asthma, and kidney failure[22, 23]. The kidney failure was defined as eGFR ≤ 60 mL/min/1. 73 m^2 or albumin to creatinine ratio ≥ 30 mg/g[24]. As a depression screening scale. The PHQ-9 score included 9 items such as “not at all”, “a few days”, “more than half”, and “almost every day” with a score of 0 ~ 3, yielding a full score of 27 points[25, 26].

2.5 Machine learning

We applied five machine learning algorithms to train the classification model. The first one was the decision tree, a classification model based on a tree-like structure, which was utilized to make decisions by progressively splitting the data into multiple nodes. The decision tree model is easy to understand and interpret, and suitable for handling non-linear relationships and complex rules in data[27]. Random forest (RF) is an ensemble learning model that reduces errors and improves prediction reliability by constructing multiple independent decision trees (similar to a flowchart judgment model) and then synthesizing the results of all trees[28]. Multilayer perceptron (MLP) is a network model that simulates the connections of human brain neurons. By continuously adjusting the connection strength of each neuron in the network (i.e. backpropagation algorithm), it reduces prediction errors and improves the ability to judge disease risk[29]. Adaboost, an ensemble learning method, can first train a simple model (weak classifier), then adjust weights based on its incorrectly predicted samples (making hard-to-predict samples more noticeable), and then train the next model, ultimately integrating the results of all models to improve prediction accuracy[30]. XGBoost, an efficient gradient boosting algorithm that continuously builds new decision trees to correct errors in previous models, can gradually improve predictive performance. It is suitable for processing complex data and is widely used in classification and regression tasks, with great performance on large-scale datasets[31].

2.6 Statistical analysis

Continuous variables were represented in the form of mean and standard deviation, while categorical variables were represented as percentages. We utilized the *t*-test for inter-group comparison of continuous variables and a chi-square test for inter-group comparison of categorical variables. The data was split into a 7:3 ratio for training and testing sets. Machine learning models were developed using Python 3.9.7 and the *sklearn* package[32], and the receiver operating characteristic (ROC) curves were plotted using the *matplotlib* package[33]. Five machine learning algorithms, including Decision Tree, RF, MLP, Adaboost, and XGBoost, were applied to train the prediction model. The grid search method was employed to generate the optimal model parameters by adjusting the model parameters for different models and evaluating their performance. The trained models were evaluated on a test set using 10-fold cross-validation to determine the stability of the model. The following indices were used in evaluation: accuracy (the proportion of correct overall model predictions), sensitivity (the ability to correctly identify actual patients, that is, the proportion of “no missed diagnosis”), specificity (the ability to correctly identify actual unaffected individuals, that is, the proportion of “no misdiagnosis”), the area under the curve (AUC) of the ROC curves (the comprehensive ability of the model to distinguish between diseased and non-diseased individuals, with the value closer to 1 indicating greater ability to distinguish patients), and the Matthews correlation coefficient (MCC, the accuracy of model classification, ranging from -1 to 1, with the value closer to 1 indicating the more reliable ability of classification). Based on the best-performing machine learning model, the SHAP (SHapley Additive exPlanations) model, a tool grounded in mathematical theory, was utilized to analyze the impact of each factor on the model's prediction results, identifying more important factors for diabetes risk prediction. The partial SHAP values were plotted as a summary plot, which included the relative ranking of features and the relationship between each feature and the outcome. The SHAP values for each feature were calculated for each sample to reflect the impact of the feature on the prediction result. Next, we aggregated the average absolute SHAP values and summarized the global contribution of each feature in a bar chart form[34]. To address the issue of data imbalance in the study, the combination technique of SMOTEENN from the *imblearn* package was applied to handle imbalanced data. First, we oversampled SMOTE, then cleaned the samples with ENN to reduce noisy samples and refine model generalization performance[35] (P<0.05: statistically significant).

3 Results

3.1 Baseline characteristics

The baseline features included in this project are displayed in Table 1. With 260 diabetic patients and 2,095 non-diabetic patients in high-risk groups, the results suggested that the average age of those with diabetes was higher (46.37 vs. 39.15, P < 0.001), and the proportion of people with an education level of high school education or equivalent education was even lower (46.9% vs. 48.4%, P = 0.020). Waist circumference (113.35 vs. 99.32, P < 0.001) and SBP (126.22 vs. 121.06, P < 0.001) in diabetic patients were significantly higher than those in the unaffected group. Heart disease (9.6% vs. 4.2%), hypertension (68.5% vs. 46.0%, P < 0.001), arthritis (35.4% vs. 20.9%, P < 0.001), stroke (35.4% vs. 20.9%, P < 0.001), asthma (25.0% vs. 18.5%, *P* = 0.016), chronic kidney disease (23.5% vs. 9.3%, *P* < 0.001), hearing loss (11.5% vs. 9.3%, *P* < 0.001). 5.5%, *P* < 0.001), and visual loss (12.7% vs. 5.6%, *P* < 0.001) were more likely to occur in diabetic patients than in unaffected people. In addition, patients with diabetes also showed elevated scores of PHQ-9 (5.50 vs. 5.06, *P* = 0.01).

3.2 Model performance comparison

Table 2 shows the performance of five models on the test set. The decision tree model had an accuracy of 0.744, sensitivity of 0.714, specificity of 0.789, and AUC of 0.751. In contrast, the RF model had an accuracy of 0.784, sensitivity of 0.739, specificity of 0.849, and AUC of 0.896, exhibiting better performance in all aspects. The MLP model had an AUC of 0.900 on the test set, with high accuracy (0.822) and sensitivity (0.905), but slightly lower specificity (0.704). The AUC of the AdaBoost model on the test set was 0.895, with a specificity of 0.805 and good accuracy (0.837) and sensitivity (0.859). The XGBoost model had an accuracy of 0.815, sensitivity of 0.962, specificity of 0.602, and AUC of 0.903 on the test set. Compared to MLP and AdaBoost, the XGBoost model had lower specificity but possessed the best sensitivity and classification ability.

Additionally, we calculated the MCC for each model to provide a more comprehensive evaluation of model performance. As shown in Table 2, the MCC for the decision tree model was 0.361, for the RF model was 0.418, for the MLP model was 0.447, for the Adaboost model was 0.463, and for the XGBoost model was 0.443. The results indicated that the Adaboost and MLP models performed well in balancing sensitivity and specificity. However, the MCC results further supported the conclusion that the RF and XGBoost models excelled in classification accuracy and recognition capability, respectively. The MCC values of these two models still demonstrated their good classification capabilities. Therefore, from an overall assessment of model performance, RF and XGBoost were the top-performing models.

3.3 Feature importance

Based on the comparison of model performance above, we employed RF and XGBoost to calculate the importance of each feature. Figure 2 shows the impact of the baseline values of the top 10 features on the model output, i.e., the development of relative risk for diabetes. Combining the SHAP summary plot (Figure 2A) with the bar plot (Figure 2B), the top three features in the RF model were age (SHAP = 0.11), waist circumference (SHAP = 0.11), and BMI (SHAP = 0.06), while sex, DBP, education level, SBP, PHQ-9 score, PIR, and race were the next most important features. Similarly, in the XGBoost model, the top three features were waist circumference (SHAP = 2.1), age (SHAP = 1.7), and BMI (SHAP = 0.67), while sex, DBP, SBP, education level, PHQ score, PIR, and race were the next most important features (Figure 2C-D). In the RF and XGBoost models, waist circumference, age, BMI, and PHQ-9 score were positively correlated with diabetes risk, while education level and PIR were negatively correlated with diabetes risk, and women had a higher diabetes risk relative to men.

4 Discussion

In this project, 2,355 individuals at high risk of diabetes from the NHANES in the years 2013-2018 were considered for building the risk model. We applied five machine learning methods (decision tree, RF, MLP, Adaboost, and XGBoost) and evaluated their performance, finding the AUC values of the RF and XGBoost models in the test set were 0.896 and 0.903, respectively. The accuracy of the RF model on the test set was 0.784, with sensitivity of 0.739 and specificity of 0.849, while the XGBoost model had an accuracy of 0.815, with sensitivity of 0.962, indicating that these two machine learning algorithms possessed high predictive ability in diabetes risk assessment. Moreover, the MCC values of the RF and XGBoost models were 0.418 and 0.443, respectively, further validating their robust classification performance when handling imbalanced datasets. This also reinforces the suitability of RF and XGBoost for developing personalized diabetes risk assessment tools. Furthermore, we also conducted feature importance analysis on these two models and found that waist circumference, age, and BMI were closely linked to the development of diabetes, while gender, SBP, DBP, education level, PIR, PHQ score, and race were identified as important predictive factors. The present work provided important clues for innovating personalized disease risk assessment tools in the future, with great potential to refine the early prevention and management of diabetes.

In this study, XGBoost performed the best in overall performance with its gradient boosting architecture, which is consistent with the conclusions of multiple studies. For example, XGBoost achieved an AUROC of 0.92 in health literacy prediction and nutrition score modeling, outperforming RF (0.90) and logistic regression (0.88), and leading in sensitivity (91%), specificity (84%), and other indicators[36]. Another NHANES study showed that the AUC of XGBoost (0.8168) was significantly higher than that of RF and logistic regression (about 0.79), and the three were similar in accuracy (about 85%)[37]. In addition, in the Patient Generated Subject Global Assessment (PG-SGA) score prediction, the AUC values of XGBoost and RF were 0.75 and 0.77, respectively, showing the best performance[38]. The high sensitivity of XGBoost (>96% in this study) makes it an ideal tool for preliminary screening of large-scale populations, minimizing missed diagnosis rates to the greatest extent possible. In contrast, although RF has slightly lower AUC (0.896) and sensitivity (0.739) than XGBoost, its specificity (0.849) is significantly higher. This "low false positive" advantage stems from its ability to capture non-linear relationships and feature interactions[39]. Therefore, the RF model is more suitable for clinical diagnosis, such as conducting secondary validation on individuals with XGBoost initial positive screening to reduce misdiagnosis rates and avoid wasting limited medical resources on false-positive individuals. This two-stage strategy (XGBoost preliminary screening + RF verification) can be effectively integrated into the existing management process of high-risk groups of diabetes. More specifically, the XGBoost model is used to quickly identify a large number of potential high-risk individuals in community physical examination, health file system organization, or outpatient preliminary screening. Subsequently, for those who tested positive in the initial screening, the RF model was applied in the clinical environment for more accurate review and risk assessment, and intervention priorities were determined based on the doctor's judgment. It is worth noting that traditional models still have value in specific scenarios. Logistic regression often performs robustly in external validation sets. Among Bayes logistic regression, decision tree comparisons, logistic regression repeatedly ranks among the top three[40], but its performance may be limited when dealing with nonlinear relationships[41]. Meanwhile, support vector machines (SVM) are comparable to the optimal model in certain tasks (such as AUROC 0.83)[42], but most studies show that their performance is slightly inferior to RF or XGBoost[43, 44].

This study found that BMI, age, waist circumference, and depression were positively correlated with diabetes in high-risk groups, and these key predictors were highly consistent with the existing literature on diabetes risk. Firstly, waist circumference, as a core index to measure abdominal obesity, has been identified as the most important predictor (the highest SHAP value) in the RF and XGBoost models of this study, which is consistent with a large body of evidence that abdominal visceral fat is the core pathophysiological mechanism of diabetes[45]. The visceral adipose tissue has strong metabolic activity and secretes a large amount of pro-inflammatory factors and free fatty acids, directly leading to insulin resistance and β-cell dysfunction[45, 46]. Secondly, as an immutable risk factor, age growth has been widely confirmed to be associated with disease[47, 48]. This study again supported the key role of β-cell function decline and insulin sensitivity decline in the development of diabetes during aging[49, 50]. Thirdly, as an indicator of overall obesity, BMI has a strong association with diabetes[51-53]. This study confirmed that BMI is still an important risk marker in high-risk groups. Obesity (whether overall or abdominal) promotes the development of diabetes by increasing pancreatic fat deposition, increasing the burden of β cells, and systemic insulin resistance[54, 55]. Finally, this study found that depression is an important psychosocial risk factor, which is consistent with previous research[56, 57]. It is worth noting that this study not only confirmed the risk of depression, but also included the screening criteria of “continuous use of antidepressant drugs for more than 3 months”. The results also suggest that this population has a higher risk, which is consistent with the literature exploring the possible impact of antidepressant drug use on blood glucose control[58].

The positive correlation of BMI, age, waist circumference, and depression with the risk of diabetes in high-risk groups not only provides strong support for the study of diabetes risk mechanisms but also has a clear application value in clinical practice and public health. From the perspective of clinical practice, these key predictors can directly guide the hierarchical management and precise intervention of high-risk groups. First, the waist measurement can be included in the core indicators of routine screening of high-risk groups of diabetes in clinical practice, and in-depth tests such as fasting blood glucose and glycosylated hemoglobin will be given priority to individuals with significantly increased waist circumference. At the same time, targeted abdominal fat reduction programs will be developed, such as combining diet control and core muscle group training, to reduce the risk of insulin resistance caused by visceral fat deposition[59]. Secondly, in response to the irreversible risk factor of aging, in clinical practice, it is necessary to strengthen regular follow-up for high-risk populations over 40 years old, especially focusing on their blood glucose fluctuations and changes in β-cell function. Early initiation of lifestyle interventions can delay the decline of β-cell function. Thirdly, for individuals with high BMI, weight management should be the core intervention goal, achieved through personalized nutrition guidance and exercise prescriptions to reduce pancreatic fat burden and improve insulin sensitivity[60]. Fourthly, for high-risk populations with depression and long-term use of antidepressants, psychiatric and endocrinology departments should collaborate to evaluate and prioritize the selection of antidepressants with minimal impact on blood sugar. At the same time, regular monitoring of glycated hemoglobin and fasting blood sugar should be conducted to avoid adverse effects of medication on blood sugar control. In addition, from the perspective of public health applications, a simple risk scoring tool can be developed based on waist circumference, age, BMI, and depression status to quickly identify high-risk individuals by primary healthcare institutions. For high-risk subgroups such as the elderly who are easily overlooked, theme health education should be implemented by combining community resources to lower intervention thresholds. For the population who use antidepressants for a long time, medical institutions should be promoted to establish a linkage mechanism between medication and blood glucose monitoring, and blood glucose indicators should be incorporated into the routine evaluation system for depression treatment[61]. Through the combination of clinical practice and public health measures, the transformation from risk prediction to active prevention can be realized to ultimately reduce the incidence rate of diabetes and the burden of related complications.

In addition, SBP and DBP in this study are also related to the risk of diabetes; that is, blood pressure level affects the risk of diabetes. Multiple population studies have confirmed the universality of this association. Among African Americans and Caucasians aged 35-54, higher blood pressure is associated with a higher risk of diabetes compared with normal blood pressure[62]. The Korean adult cohort study showed that even for people in prehypertension (120-139/80-89 mmHg), their risk of diabetes was significantly higher than that of normotension[63]. These studies indicate that blood pressure management should not be limited to patients with diagnosed hypertension. Blood pressure monitoring of high-risk groups (including early status) should be the core component of diabetes prevention strategies. In addition, race has been identified as a key sociobiological predictor. American data shows that the prevalence of diabetes among non-Hispanic blacks, Asians, and Hispanics (12% -14%) is significantly higher than that of other ethnic groups[64], and this difference persists among high-risk elderly people[65]. It suggests that when developing public health interventions, it is important to focus on high-risk racial/ethnic groups and integrate culturally sensitive support programs in community screening and health management.

In our study, women showed a higher risk of diabetes in high-risk groups of diabetes. This is closely related to the physiological changes unique to women, especially the lack of estrogen during menopause. Premature menopause (<40 years old) or surgical menopause significantly increases the risk of type 2 diabetes[66, 67]. Estrogen deficiency affects the development of diabetes through multiple mechanisms, including changes in insulin secretion of pancreatic β cells, decreased sensitivity of targeted organs and tissues to insulin, and increased sensitivity of major organs of diabetes related pathology to glucose[67]. In addition, in an epidemiological study, the risk of insomnia in women of all ages is found to be generally 40% higher than that in men, and there is a close relationship between insomnia and diabetes[68]. Sleep disorders are closely related to obesity and insulin resistance[69]. Lack of sleep can disrupt key hormones that regulate appetite and energy balance (such as leptin, ghrelin, adiponectin), increase intake of high-calorie foods, and worsen blood sugar control[70-72]. This suggests that in the risk screening of diabetes, women's reproductive history (such as menopausal age, surgical menopause history) and sleep quality evaluation should be routinely included. At the same time, for perimenopausal and postmenopausal women, weight management and education and support on lifestyle intervention (healthy diet, regular exercise) should be strengthened.

Education level and income are equally crucial social determinants. High levels of education typically promote healthier lifestyles, reduce shift work (lowering stress and obesity risk), and enhance health awareness and proactive prevention behaviors[73-76]. On the contrary, low income and poverty significantly increased the risk of diabetes (the probability increased 2-3 times)[77]. Improving the socio-economic environment (such as moving out of high poverty areas) can reduce the prevalence of diabetes[78]. Poverty is often accompanied by resource limitations such as malnutrition and lack of safe exercise space, exacerbating risk factors such as obesity[79]. Therefore, for the intervention of high-risk groups of diabetes, we must pay attention to improving the health literacy of low-education/low-income groups, and provide culturally appropriate and easy-to-understand educational materials and support services. At the same time, a sound social security system should be established to ensure their access to nutrition and basic medical services, and create a supportive environment to encourage physical exercise[80].

This study showed that RF and XGBoost models had the best prediction performance in the risk prediction among groups with high-risk diabetes, effectively identifying key risk factors such as waist circumference, age, BMI, depression, SBP, and DBP, as well as socioeconomic factors such as gender, education level, and income. These findings support the development of RF and XGBoost models as personalized risk assessment tools that can be embedded in electronic health records systems or mobile health applications to assist clinicians in achieving risk stratification management, improving the efficiency of early screening and intervention for diabetes, and ultimately reducing the incidence rate and burden of complications.

However, this study also has some limitations. Firstly, the study adopts a cross-sectional design and cannot directly infer causal relationships. Compared to longitudinal studies that can reveal the temporal correlation of disease occurrence through long-term follow-up data (such as tracking changes in blood glucose and dynamic evolution of risk factors), this study fails to capture such dynamic effects based on the cross-sectional data. Future research needs to adopt a prospective cohort design, combined with time series analysis, to more accurately clarify the causal path between various risk factors and the onset of diabetes. Secondly, although the feature variables included in this study cover multidimensional information, key predictive factors may still be overlooked. In the future, genetic data, physical activity monitoring data, dietary habits, and other information can be further integrated to improve the predictive accuracy of the model. In addition, considering the high specificity of the RF model and the high sensitivity of the XGBoost model, the exploration of the integration algorithm of the two may further optimize the classification performance, balancing the missed diagnosis rate and misdiagnosis rate. Finally, this study only evaluates the performance of the model through internal validation, and external generalizability still needs to be verified. In the future, external validation should be conducted among different populations, especially focusing on the applicability of the model in resource-limited areas and underserved populations. Based on the above direction, future research can further develop a personalized risk assessment tool that integrates multi-source data, assisting medical personnel and patients to make joint decisions, ultimately achieving the transformation from risk prediction to accurate prevention, and especially providing feasible solutions for diabetes prevention and control in resource scarce regions to promote the fairness of global diabetes prevention.

Declarations

Consent for publication

All of the authors have given their consent to submit the manuscript for publication.

Availability of date and materials

The date and materials utilized in the current study are accessible upon reasonable request from the corresponding author.

Competing interest

The authors declare that they have no conflicts of interest.

Ethics approval and consent to participate

Not applicable.

Funding

Not applicable.

Authors' contributions

Xiaohua Yang and Ting Sun conceived and designed the study.

Xiaohua Yang and Meiqi Yao conducted the assays.

Xiaohua Yang and Ting Sun wrote the manuscript.

Jia Huang and Zhuojing Cheng reviewed and edited the manuscript.

All authors read and approved the final manuscript.

Acknowledgments

Not applicable.

Reference

1. Organization WH. Diabetes. 2023; Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes.
2. Shin J, Kim J, Lee C, Yoon JY, Kim S, Song S, et al. Development of Various Diabetes Prediction Models Using Machine Learning Techniques. Diabetes Metab J. 2022; 46:650-7.
3. Liu T, Zhao J, Lin C. Sprouty-related proteins with EVH1 domain (SPRED2) prevents high-glucose induced endothelial-mesenchymal transition and endothelial injury by suppressing MAPK activation. Bioengineered. 2022; 13:13882-92.
4. Collaborators GBDD. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2023; 402:203-34.
5. Vornanen M, Konttinen H, Peltonen M, Haukkala A. Diabetes and Cardiovascular Disease Risk Perception and Risk Indicators: a 5-Year Follow-up. Int J Behav Med. 2021; 28:337-48.

6. Adriaanse MC, Twisk JW, Dekker JM, Spijkerman AM, Nijpels G, Heine RJ, et al. Perceptions of risk in adults with a low or high risk profile of developing type 2 diabetes; a cross-sectional population-based study. Patient Educ Couns. 2008; 73:307-12.

7. Naina Marikar S, Al-Hasani K, Khurana I, Kaipananickal H, Okabe J, Maxwell S, et al. Pharmacological inhibition of human EZH2 can influence a regenerative beta-like cell capacity with in vitro insulin release in pancreatic ductal cells. Clin Epigenetics. 2023; 15:101.

8. Walther F, Heinrich L, Schmitt J, Eberlein-Gonska M, Roessler M. Prediction of inpatient pressure ulcers based on routine healthcare data using machine learning methodology. Sci Rep. 2022; 12:5044.

9. Gong Q, Zhang P, Wang J, Ma J, An Y, Chen Y, et al. Morbidity and mortality after lifestyle intervention for people with impaired glucose tolerance: 30-year results of the Da Qing Diabetes Prevention Outcome Study. Lancet Diabetes Endocrinol. 2019; 7:452-61.

10. Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med. 2018; 24:1304-5.

11. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 2018; 71:2668-79.

12. Myszczynska MA, Ojamies PN, Lacoste AMB, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol. 2020; 16:440-56.

13. Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int J Environ Res Public Health. 2021; 18.

14. Olusanya MO, Ogunsakin RE, Ghai M, Adeleke MA. Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach. Int J Environ Res Public Health. 2022; 19.

15. Hu H, Lai T, Farid F. Feasibility Study of Constructing a Screening Tool for Adolescent Diabetes Detection Applying Machine Learning Methods. Sensors (Basel). 2022; 22.

16. Prevention CfDCa. National Health and Nutrition Examination Survey. Available from: http://www.cdc.gov/nchs/nhanes.htm.

17. Expert Committee on the D, Classification of Diabetes M. Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care. 2003; 26 Suppl 1:S5-20.

18. Genuth S, Alberti KG, Bennett P, Buse J, Defronzo R, Kahn R, et al. Follow-up report on the diagnosis of diabetes mellitus. Diabetes Care. 2003; 26:3160-7.

19. Hajian-Tilaki K, Heidari B, Hajian-Tilaki A. Are Gender Differences in Health-related Quality of Life Attributable to Sociodemographic Characteristics and Chronic Disease Conditions in Elderly People? Int J Prev Med. 2017; 8:95.

20. Zhu C, Zhang H, Shen Z, Chen J, Gu Y, Lv S, et al. Cystatin C-based estimated GFR performs best in identifying individuals with poorer survival in an unselected Chinese population: results from the China Health and Retirement Longitudinal Study (CHARLS). Clin Kidney J. 2022; 15:1322-32.

21. Patel JS, Oh Y, Rand KL, Wu W, Cyders MA, Kroenke K, et al. Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005-2016. Depress Anxiety. 2019; 36:813-23.

22. Ferguson JM, Jacobs J, Yefimova M, Greene L, Heyworth L, Zulman DM. Virtual care expansion in the Veterans Health Administration during the COVID-19 pandemic: clinical services and patient characteristics associated with utilization. J Am Med Inform Assoc. 2021; 28:453-62.

23. Walker EA, Mertz CK, Kalten MR, Flynn J. Risk perception for developing diabetes: comparative

risk judgments of physicians. Diabetes Care. 2003; 26:2543-8.

24. Sakhuja S, Jaeger BC, Akinyelure OP, Bress AP, Shimbo D, Schwartz JE, et al. Potential impact of systematic and random errors in blood pressure measurement on the prevalence of high office blood pressure in the United States. J Clin Hypertens (Greenwich). 2022; 24:263-70.

25. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16:606-13.

26. Li W, Zeng L, Yuan S, Shang Y, Zhuang W, Chen Z, et al. Machine learning for the prediction of cognitive impairment in older adults. Front Neurosci. 2023; 17:1158141.

27. Haque UM, Kabir E, Khanam R. Early detection of paediatric and adolescent obsessive-compulsive, separation anxiety and attention deficit hyperactivity disorder using machine learning algorithms. Health Inf Sci Syst. 2023; 11:31.

28. KDnuggets. Random Forests®, Explained. 2017; Available from: https://www.kdnuggets.com/2017/10/random-forests-explained.html.

29. Mohammed M, Munir M, Aljabr A. Prediction of Date Fruit Quality Attributes during Cold Storage Based on Their Electrical Properties Using Artificial Neural Networks Models. Foods. 2022; 11.

30. Ullah Z, Saleem F, Jamjoom M, Fakieh B. Reliable Prediction Models Based on Enriched Data for Identifying the Mode of Childbirth by Using Machine Learning Methods: Development Study. J Med Internet Res. 2021; 23:e28856.

31. Bavaro DA, Fanizzi A, Iacovelli S, Bove S, Comes MC, Cristofaro C, et al. A Machine Learning Approach for Predicting Capsular Contracture after Postmastectomy Radiotherapy in Breast Cancer Patients. Healthcare (Basel). 2023; 11.

32. Ferre F, Laurent R, Furelau P, Doumard E, Ferrier A, Bosch L, et al. Perioperative Risk Assessment of Patients Using the MyRISK Digital Score Completed Before the Preanesthetic Consultation: Prospective Observational Study. JMIR Perioper Med. 2023; 6:e39044.

33. Chen W, Zhang L, Cai G, Zhang B, Lian Z, Li J, et al. Machine learning-based multimodal MRI texture analysis for assessing renal function and fibrosis in diabetic nephropathy: a retrospective study. Front Endocrinol (Lausanne). 2023; 14:1050078.

34. Liu X, Morelli D, Littlejohns TJ, Clifton DA, Clifton L. Combining machine learning with Cox models to identify predictors for incident post-menopausal breast cancer in the UK Biobank. Sci Rep. 2023; 13:9221.

35. Yu S, Zhang M, Ye Z, Wang Y, Wang X, Chen YG. Development of a 32-gene signature using machine learning for accurate prediction of inflammatory bowel disease. Cell Regen. 2023; 12:8.

36. Inceoglu F, Deniz S, Yagin FH. Prediction of effective sociodemographic variables in modeling health literacy: A machine learning approach. Int J Med Inform. 2023; 178:105167.

37. Riveros Perez E, Avella-Molano B. Learning from the machine: is diabetes in adults predicted by lifestyle variables? A retrospective predictive modelling study of NHANES 2007-2018. BMJ Open. 2025; 15:e096595.

38. Qian G, Jiaxin H, Minghua C, Beijia L, Yinfeng L, Guiyu H, et al. Rapid identification of tumor patients with PG-SGA >/= 4 based on machine learning: a prospective study. BMC Cancer. 2025; 25:902.

39. Zhang Y, Zhang X, Razbek J, Li D, Xia W, Bao L, et al. Opening the black box: interpretable machine learning for predictor finding of metabolic syndrome. BMC Endocr Disord. 2022; 22:214.

40. Qi J, Lei J, Li N, Huang D, Liu H, Zhou K, et al. Machine learning models to predict in-hospital mortality in septic patients with diabetes. Front Endocrinol (Lausanne). 2022; 13:1034251.

41. Chu WM, Tsan YT, Chen PY, Chen CY, Hao ML, Chan WC, et al. A model for predicting physical

function upon discharge of hospitalized older adults in Taiwan-a machine learning approach based on both electronic health records and comprehensive geriatric assessment. Front Med (Lausanne). 2023; 10:1160013.

42. Rus Prelog P, Matic T, Pregelj P, Sadikov A. A pilot predictive model based on COVID-19 data to assess suicidal ideation indirectly. J Psychiatr Res. 2023; 163:318-24.

43. Obagbuwa IC, Danster S, Chibaya OC. Supervised machine learning models for depression sentiment analysis. Front Artif Intell. 2023; 6:1230649.

44. Asnake AA, Gebrehana AK, Asebe HA, Seifu BL, Fente BM, Bezie MM, et al. Application of machine learning algorithm for prediction of abortion among reproductive age women in Ethiopia. Sci Rep. 2025; 15:17924.

45. Sam S. Differential effect of subcutaneous abdominal and visceral adipose tissue on cardiometabolic risk. Horm Mol Biol Clin Investig. 2018; 33.

46. Joshi RD, Dhakal CK. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int J Environ Res Public Health. 2021; 18.

47. Meshram, II, Vishnu Vardhana Rao M, Sudershan Rao V, Laxmaiah A, Polasa K. Regional variation in the prevalence of overweight/obesity, hypertension and diabetes and their correlates among the adult rural population in India. Br J Nutr. 2016; 115:1265-72.

48. De Tata V. Age-related impairment of pancreatic Beta-cell function: pathophysiological and cellular mechanisms. Front Endocrinol (Lausanne). 2014; 5:138.

49. Hernandez-Bautista RJ, Alarcon-Aguilar FJ, Del CE-VM, Almanza-Perez JC, Merino-Aguilar H, Fainstein MK, et al. Biochemical alterations during the obese-aging process in female and male monosodium glutamate (MSG)-treated mice. Int J Mol Sci. 2014; 15:11473-94.

50. Lee JH, Lee J. Endoplasmic Reticulum (ER) Stress and Its Role in Pancreatic beta-Cell Dysfunction and Senescence in Type 2 Diabetes. Int J Mol Sci. 2022; 23.

51. Par F, Sarvi F, Khodadost M, Pezeshki B, Doosti H, Tabrizi R. A Nonlinear Association of Body Mass Index and Fasting Blood Glucose: A Dose-Response Analysis From Fasa Adults Cohort Study (FACS). Health Sci Rep. 2025; 8:e70560.

52. Poulsen K, Cleal B, Clausen T, Andersen LL. Work, diabetes and obesity: a seven year follow-up study among Danish health care workers. PLoS One. 2014; 9:e103425.

53. Ng ACT, Delgado V, Borlaug BA, Bax JJ. Diabesity: the combined burden of obesity and diabetes on heart disease and the role of imaging. Nat Rev Cardiol. 2021; 18:291-304.

54. Skudder-Hill L, Sequeira IR, Cho J, Ko J, Poppitt SD, Petrov MS. Fat Distribution Within the Pancreas According to Diabetes Status and Insulin Traits. Diabetes. 2022; 71:1182-92.

55. Al-Mrabeh A, Hollingsworth KG, Shaw JAM, McConnachie A, Sattar N, Lean MEJ, et al. 2-year remission of type 2 diabetes and pancreas morphology: a post-hoc analysis of the DiRECT open-label, cluster-randomised trial. Lancet Diabetes Endocrinol. 2020; 8:939-48.

56. Mezuk B, Eaton WW, Albrecht S, Golden SH. Depression and type 2 diabetes over the lifespan: a meta-analysis. Diabetes Care. 2008; 31:2383-90.

57. Rubin RR, Ma Y, Marrero DG, Peyrot M, Barrett-Connor EL, Kahn SE, et al. Elevated depression symptoms, antidepressant medicine use, and risk of developing diabetes during the diabetes prevention program. Diabetes Care. 2008; 31:420-6.

58. Kammer JR, Hosler AS, Leckman-Westin E, DiRienzo G, Osborn CY. The association between antidepressant use and glycemic control in the Southern Community Cohort Study (SCCS). J Diabetes Complications. 2016; 30:242-7.

59. Russell LE, Tse J, Bowie J, Richardson CR, Trubek A, Maruthur N, et al. Cooking behaviours after Diabetes Prevention Program (DPP) participation among DPP participants in Baltimore, MD. Public Health Nutr. 2023; 26:2492-7.

60. Crandall JP, Dabelea D, Knowler WC, Nathan DM, Temprosa M, Group DPPR. The Diabetes Prevention Program and Its Outcomes Study: NIDDK's Journey Into the Prevention of Type 2 Diabetes and Its Public Health Impact. Diabetes Care. 2025; 48:1101-11.

61. Murteira R, Cary M, Galante H, Romano S, Guerreiro JP, Rodrigues AT. Effectiveness of a collaborative diabetes screening campaign between community pharmacies and general practitioners. Prim Care Diabetes. 2023; 17:314-20.

62. Wei GS, Coady SA, Goff DC, Jr., Brancati FL, Levy D, Selvin E, et al. Blood pressure and the risk of developing diabetes in african americans and whites: ARIC, CARDIA, and the framingham heart study. Diabetes Care. 2011; 34:873-9.

63. Cho NH, Kim KM, Choi SH, Park KS, Jang HC, Kim SS, et al. High Blood Pressure and Its Association With Incident Diabetes Over 10 Years in the Korean Genome and Epidemiology Study (KoGES). Diabetes Care. 2015; 38:1333-8.

64. Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and Trends in Diabetes Among Adults in the United States, 1988-2012. JAMA. 2015; 314:1021-9.

65. Odlum M, Moise N, Kronish IM, Broadwell P, Alcantara C, Davis NJ, et al. Trends in Poor Health Indicators Among Black and Hispanic Middle-aged and Older Adults in the United States, 1999-2018. JAMA Netw Open. 2020; 3:e2025134.

66. Shen L, Song L, Li H, Liu B, Zheng X, Zhang L, et al. Association between earlier age at natural menopause and risk of diabetes in middle-aged and older Chinese women: The Dongfeng-Tongji cohort study. Diabetes Metab. 2017; 43:345-50.

67. Mauvais-Jarvis F, Manson JE, Stevenson JC, Fonseca VA. Menopausal Hormone Therapy and Type 2 Diabetes Prevention: Evidence, Mechanisms, and Clinical Implications. Endocr Rev. 2017; 38:173-88.

68. Schmid SM, Hallschmid M, Schultes B. The metabolic burden of sleep loss. Lancet Diabetes Endocrinol. 2015; 3:52-62.

69. Cappuccio FP, D'Elia L, Strazzullo P, Miller MA. Quantity and quality of sleep and incidence of type 2 diabetes: a systematic review and meta-analysis. Diabetes Care. 2010; 33:414-20.

70. Spiegel K, Leproult R, L'Hermite-Baleriaux M, Copinschi G, Penev PD, Van Cauter E. Leptin levels are dependent on sleep duration: relationships with sympathovagal balance, carbohydrate regulation, cortisol, and thyrotropin. J Clin Endocrinol Metab. 2004; 89:5762-71.

71. Taheri S, Lin L, Austin D, Young T, Mignot E. Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. PLoS Med. 2004; 1:e62.

72. Hibi M, Kubota C, Mizuno T, Aritake S, Mitsui Y, Katashima M, et al. Effect of shortened sleep on energy expenditure, core body temperature, and appetite: a human randomised crossover trial. Sci Rep. 2017; 7:39640.

73. Borrell LN, Dallo FJ, White K. Education and diabetes in a racially and ethnically diverse population. Am J Public Health. 2006; 96:1637-42.

74. Hanprathet N, Lertmaharit S, Lohsoonthorn V, Rattananupong T, Ammaranond P, Jiamjarasrangsi W. Increased Risk Of Type 2 Diabetes And Abnormal FPG Due To Shift Work Differs According To Gender: A Retrospective Cohort Study Among Thai Workers In Bangkok, Thailand. Diabetes Metab Syndr Obes. 2019; 12:2341-54.

75. Suwazono Y, Dochi M, Sakata K, Okubo Y, Oishi M, Tanaka K, et al. A longitudinal study on the

effect of shift work on weight gain in male Japanese workers. Obesity (Silver Spring). 2008; 16:1887-93.

76. Allen K, McFarland M. How Are Income and Education Related to the Prevention and Management of Diabetes? J Aging Health. 2020; 32:1063-74.

77. Dinca-Panaitescu S, Dinca-Panaitescu M, Bryant T, Daiski I, Pilkington B, Raphael D. Diabetes prevalence and income: Results of the Canadian Community Health Survey. Health Policy. 2011; 99:116-23.

78. Ludwig J, Sanbonmatsu L, Gennetian L, Adam E, Duncan GJ, Katz LF, et al. Neighborhoods, obesity, and diabetes--a randomized social experiment. N Engl J Med. 2011; 365:1509-19.

79. Gaskin DJ, Thorpe RJ, Jr., McGinty EE, Bower K, Rohde C, Young JH, et al. Disparities in diabetes: the nexus of race, poverty, and place. Am J Public Health. 2014; 104:2147-55.

80. Okwechime IO, Roberson S, Odoi A. Prevalence and Predictors of Pre-Diabetes and Diabetes among Adults 18 Years or Older in Florida: A Multinomial Logistic Modeling Approach. PLoS One. 2015; 10:e0145781.

Figure legends

Figure 1 Flow chart visualizing the data processing and model development process

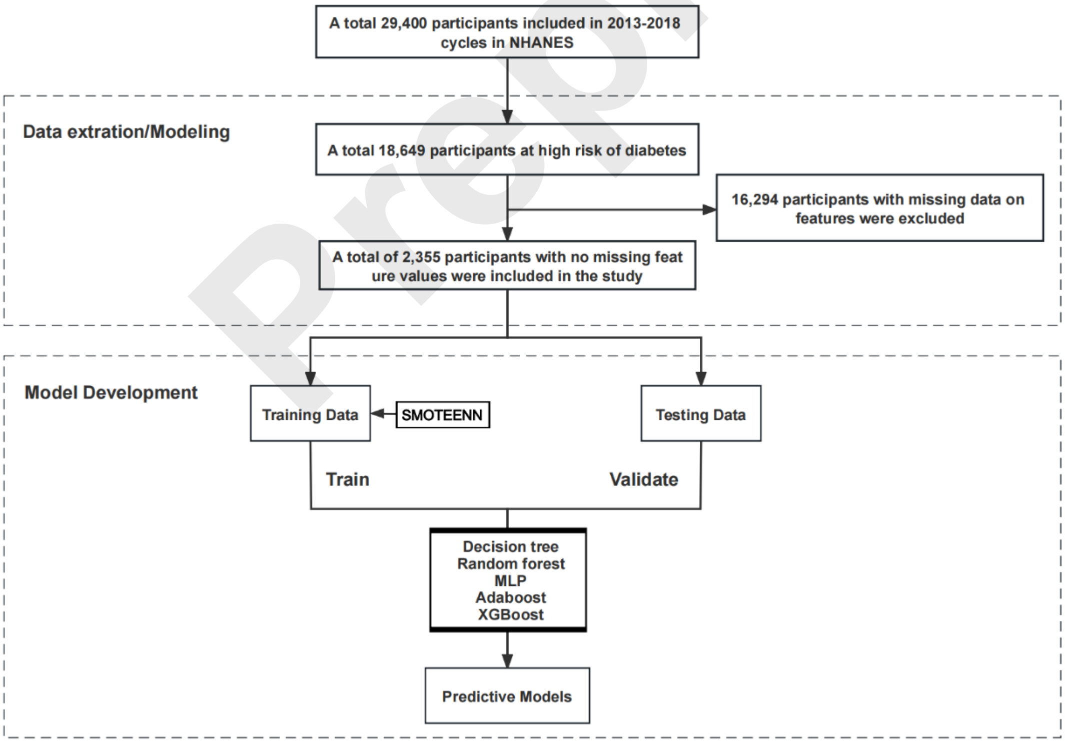
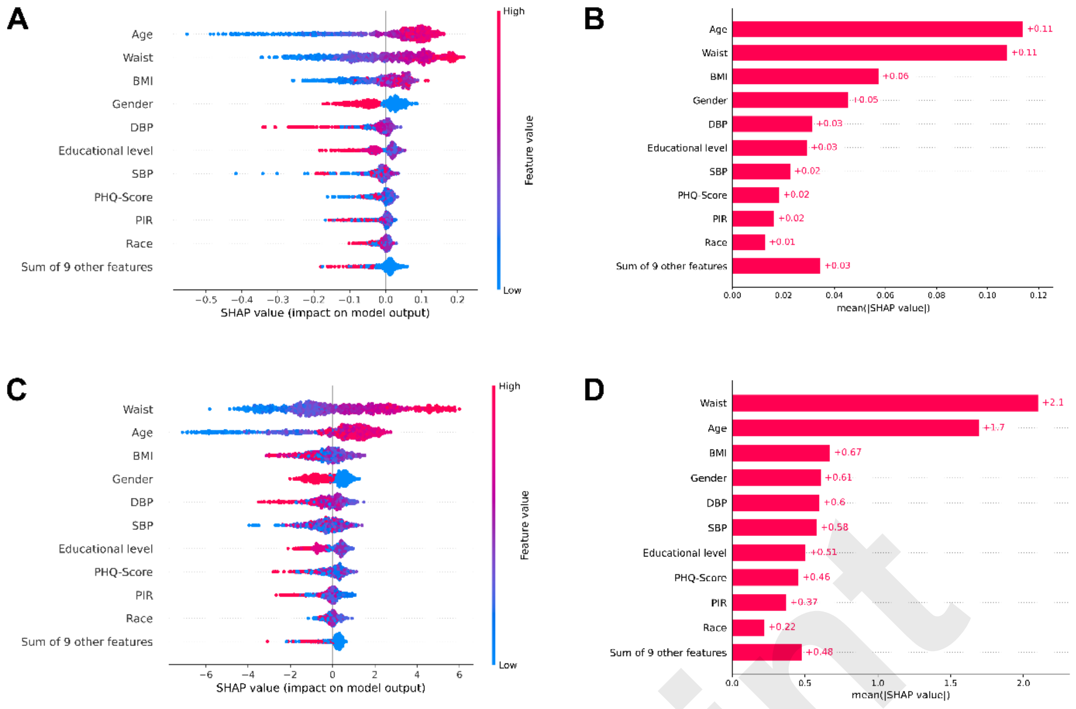

A total 29,400 participants included in 2013-2018 cycles in NHANES
Data extration/Modeling
A total 18,649 participants at high risk of diabetes
16,294 participants with missing data on features were excluded
A total of 2,355 participants with no missing feat ure values were included in the study
Model Development
Training Data
SMOTEENN
Testing Data
Train
Validate
Decision tree
Random forest
MLP
Adaboost
XGBoost
Predictive Models

Figure 2 Summary plot and feature importance for SHAP values in the testing set. Summary SHAP plots (A) and bar plots (B) of the global SHAP values of the RF model. Summary SHAP plots (C) and bar plots (D) of the global SHAP values of the XGBoost

model.



SHAP summary plot provides three aspects of information: (1) ranking indicates the relative importance of features; (2) Color gradients indicate the relative size of each feature, with red indicating high values of the feature (e.g., older age) and blue indicating the opposite (e.g., younger age), where females are shown in blue and males in red. A negative SHAP value indicates a decreased relative risk, whereas a negative SHAP value indicates an increased relative risk. (3) The discretization of points indicates whether the relationship between each feature and the outcome is linear. The bars show the global SHAP values.

Table 1 Characteristics of NHANES participants

Characters	Total	Non-diabetes	Diabetes	P-value
Overall	2355	2095 (89.0)	260 (11.0)	
Gender				0.952
Female	1159 (49.2)	1032 (49.3)	127 (48.8)	
Male	1196 (50.8)	1063 (50.7)	133 (51.2)	
Age	39.95 (11.66)	39.15 (11.65)	46.37 (9.62)	<0.001
Race				0.259
Mexican American	265 (11.3)	227 (10.8)	38 (14.6)	
Other Hispanic	200 (8.5)	175 (8.4)	25 (9.6)	
Non-Hispanic White	1020 (43.3)	920 (43.9)	100 (38.5)	
Non-Hispanic Black	587 (24.9)	524 (25.0)	63 (24.2)	
Other race	283 (12.0)	249 (11.9)	34 (13.1)	

Education level				0.020
Less than 9th grade	104 (4.4)	85 (4.1)	19 (7.3)	
9th to 12th grade (no diploma)	401 (17.0)	346 (16.5)	55 (21.2)	
High school graduate/GED equivalent	714 (30.3)	650 (31.0)	64 (24.6)	
Some college or associate degree	848 (36.0)	757 (36.1)	91 (35.0)	
College graduate or above	288 (12.2)	257 (12.3)	31 (11.9)	
PIR	2.02 (1.48)	2.03 (1.49)	1.92 (1.41)	0.260
Waist (cm)	100.87 (17.89)	99.32 (17.21)	113.35 (18.33)	<0.001
BMI (kg/m^2)	28.96 (6.68)	29.04 (6.67)	28.27 (6.71)	0.131
DBP (mmHg)	72.12 (12.53)	71.95 (12.40)	73.46 (13.51)	0.067
SBP (mmHg)	121.63 (16.05)	121.06 (15.80)	126.22 (17.26)	<0.001
Heart disease				<0.001
No	2243 (95.2)	2008 (95.8)	235 (90.4)	
Yes	112 (4.8)	87 (4.2)	25 (9.6)	
Hypertension				<0.001
No	1214 (51.5)	1132 (54.0)	82 (31.5)	
Yes	1141 (48.5)	963 (46.0)	178 (68.5)	
Arthritis				<0.001
No	1826 (77.5)	1658 (79.1)	168 (64.6)	
Yes	529 (22.5)	437 (20.9)	92 (35.4)	
Cancer				0.408
No	2231 (94.7)	1988 (94.9)	243 (93.5)	
Yes	124 (5.3)	107 (5.1)	17 (6.5)	
Stroke				0.001
No	2291 (97.3)	2047 (97.7)	244 (93.8)	
Yes	64 (2.7)	48 (2.3)	16 (6.2)	
Asthma				0.016
No	1902 (80.8)	1707 (81.5)	195 (75.0)	
Yes	453 (19.2)	388 (18.5)	65 (25.0)	
Chronic kidney disease				<0.001
No	2100 (89.2)	1901 (90.7)	199 (76.5)	
Yes	255 (10.8)	194 (9.3)	61 (23.5)	
Hearing loss				<0.001
No	2210 (93.8)	1980 (94.5)	230 (88.5)	
Yes	145 (6.2)	115 (5.5)	30 (11.5)	
Seeing loss				<0.001
No	2205 (93.6)	1978 (94.4)	227 (87.3)	
Yes	150 (6.4)	117 (5.6)	33 (12.7)	
PHQ-9 score	4.39 (5.12)	4.27 (5.06)	5.40 (5.50)	0.010

Note: PIR, Poverty income ratio. DBP, Diastolic blood pressure. SBP, Systolic blood pressure. PHQ-9, Patient Health Questionnaire-9.

Table 2 Results from 10-fold cross-validation for diabetes classification.

Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
Decision tree	0.744	0.714	0.789	0.751	0.361
Random forest	0.784	0.739	0.849	0.896	0.418
MLP	0.822	0.905	0.704	0.900	0.447
Adaboost	0.837	0.859	0.805	0.895	0.463
XGBoost	0.815	0.962	0.602	0.903	0.443

Note: AUC, Area Under the Curve; MCC, Matthews Correlation Coefficient.

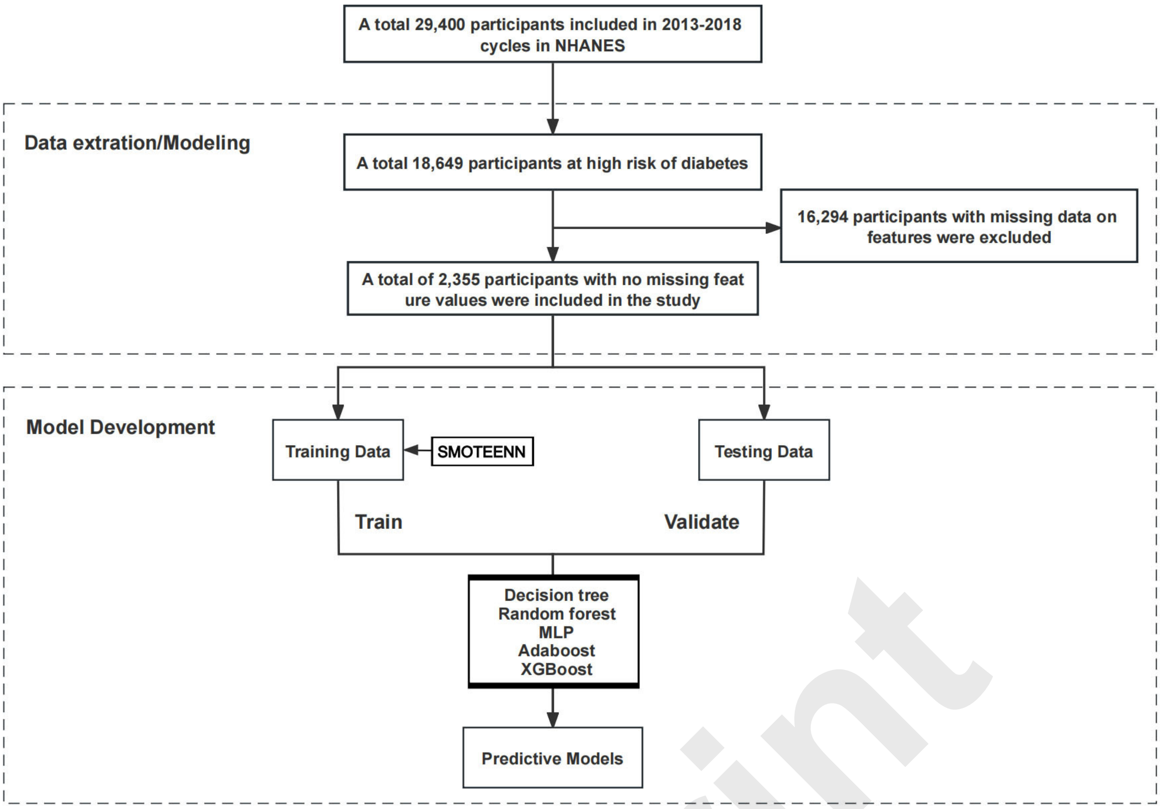

A total 29,400 participants included in 2013-2018 cycles in NHANES
Data extration/Modeling
A total 18,649 participants at high risk of diabetes
16,294 participants with missing data on features were excluded
A total of 2,355 participants with no missing feat ure values were included in the study
Model Development
Training Data
SMOTEENN
Testing Data
Train
Validate
Decision tree
Random forest
MLP
Adaboost
XGBoost
Predictive Models

Table 1 Characteristics of NHANES participants

Characters	Total	Non-diabetes	Diabetes	P-value
Overall	2355	2095 (89.0)	260 (11.0)	
Gender				0.952
Female	1159 (49.2)	1032 (49.3)	127 (48.8)	
Male	1196 (50.8)	1063 (50.7)	133 (51.2)	
Age	39.95 (11.66)	39.15 (11.65)	46.37 (9.62)	<0.001
Race				0.259
Mexican American	265 (11.3)	227 (10.8)	38 (14.6)	
Other Hispanic	200 (8.5)	175 (8.4)	25 (9.6)	
Non-Hispanic White	1020 (43.3)	920 (43.9)	100 (38.5)	
Non-Hispanic Black	587 (24.9)	524 (25.0)	63 (24.2)	
Other race	283 (12.0)	249 (11.9)	34 (13.1)	
Education level				0.020
Less than 9th grade	104 (4.4)	85 (4.1)	19 (7.3)	
9th to 12th grade (no diploma)	401 (17.0)	346 (16.5)	55 (21.2)	
High school graduate/GED equivalent	714 (30.3)	650 (31.0)	64 (24.6)	
Some college or associate degree	848 (36.0)	757 (36.1)	91 (35.0)	
College graduate or above	288 (12.2)	257 (12.3)	31 (11.9)	
PIR	2.02 (1.48)	2.03 (1.49)	1.92 (1.41)	0.260
Waist (cm)	100.87 (17.89)	99.32 (17.21)	113.35 (18.33)	<0.001
BMI (kg/m^2)	28.96 (6.68)	29.04 (6.67)	28.27 (6.71)	0.131
DBP (mmHg)	72.12 (12.53)	71.95 (12.40)	73.46 (13.51)	0.067
SBP (mmHg)	121.63 (16.05)	121.06 (15.80)	126.22 (17.26)	<0.001
Heart disease				<0.001
No	2243 (95.2)	2008 (95.8)	235 (90.4)	
Yes	112 (4.8)	87 (4.2)	25 (9.6)	
Hypertension				<0.001
No	1214 (51.5)	1132 (54.0)	82 (31.5)	
Yes	1141 (48.5)	963 (46.0)	178 (68.5)	
Arthritis				<0.001
No	1826 (77.5)	1658 (79.1)	168 (64.6)	
Yes	529 (22.5)	437 (20.9)	92 (35.4)	
Cancer				0.408
No	2231 (94.7)	1988 (94.9)	243 (93.5)	
Yes	124 (5.3)	107 (5.1)	17 (6.5)	
Stroke				0.001
No	2291 (97.3)	2047 (97.7)	244 (93.8)	
Yes	64 (2.7)	48 (2.3)	16 (6.2)	
Asthma				0.016
No	1902 (80.8)	1707 (81.5)	195 (75.0)	
Yes	453 (19.2)	388 (18.5)	65 (25.0)	
Chronic kidney disease				<0.001
No	2100 (89.2)	1901 (90.7)	199 (76.5)	

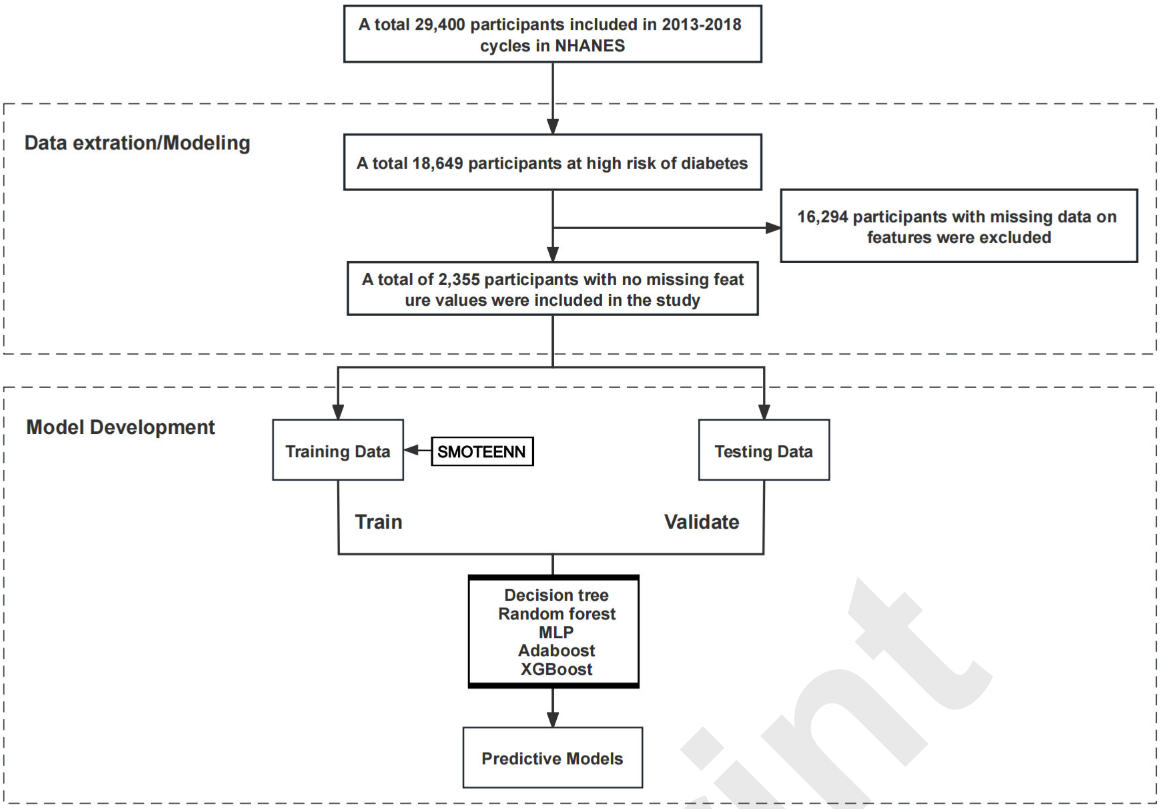
Yes	255 (10.8)	194 (9.3)	61 (23.5)	
Hearing loss				<0.001
No	2210 (93.8)	1980 (94.5)	230 (88.5)	
Yes	145 (6.2)	115 (5.5)	30 (11.5)	
Seeing loss				<0.001
No	2205 (93.6)	1978 (94.4)	227 (87.3)	
Yes	150 (6.4)	117 (5.6)	33 (12.7)	
PHQ-9 score	4.39 (5.12)	4.27 (5.06)	5.40 (5.50)	0.010

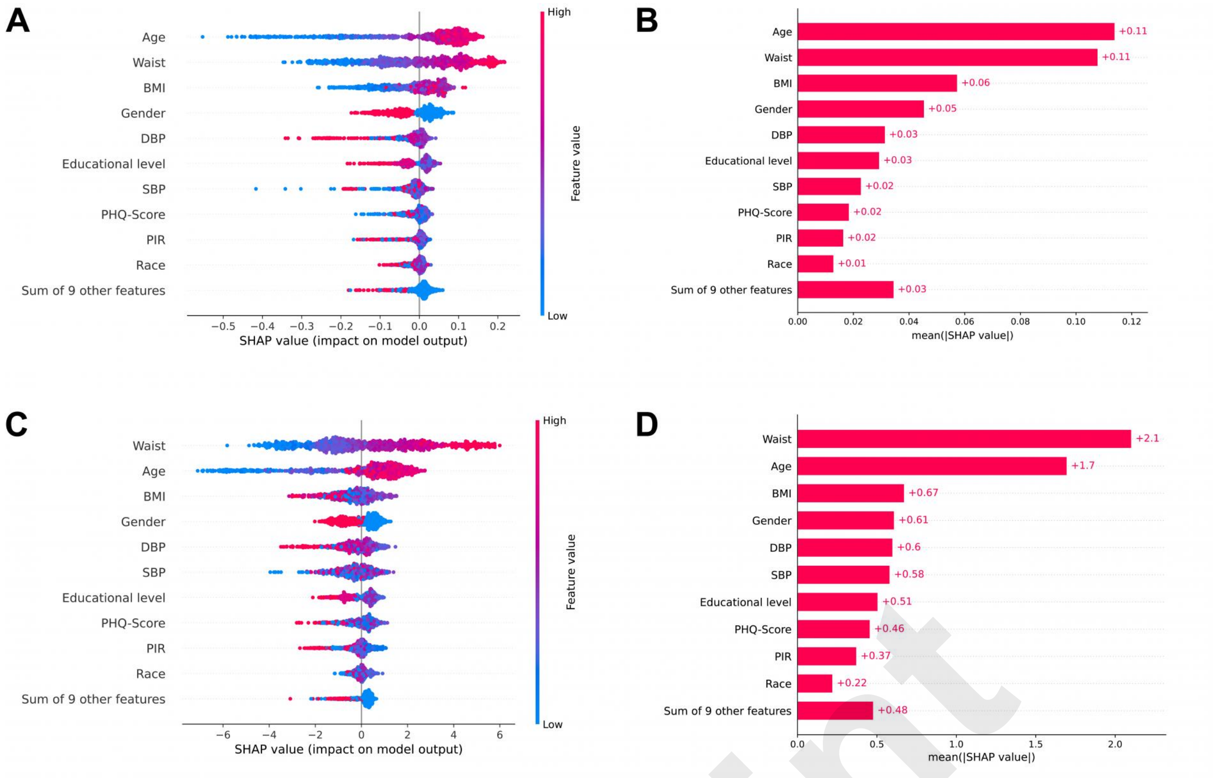
Note: PIR, Poverty income ratio. DBP, Diastolic blood pressure. SBP, Systolic blood pressure. PHQ-9, Patient Health Questionnaire-9.

Table 2 Results from 10-fold cross-validation for diabetes classification.

Model	Accuracy	Sensitivity	Specificity	AUC	MCC
Decision tree	0.744	0.714	0.789	0.751	0.361
Random forest	0.784	0.739	0.849	0.896	0.418
MLP	0.822	0.905	0.704	0.900	0.447
Adaboost	0.837	0.859	0.805	0.895	0.463
XGBoost	0.815	0.962	0.602	0.903	0.443

Note: AUC, Area Under the Curve; MCC, Matthews Correlation Coefficient.


A total 29,400 participants included in 2013-2018 cycles in NHANES
Data extration/Modeling
A total 18,649 participants at high risk of diabetes
16,294 participants with missing data on features were excluded
A total of 2,355 participants with no missing feat ure values were included in the study
Model Development
Training Data
SMOTEENN
Testing Data
Train
Validate
Decision tree
Random forest
MLP
Adaboost
XGBoost
Predictive Models


A
Age
Waist
BMI
Gender
DBP
Educational level
SBP
PHQ-Score
PIR
Race
Sum of 9 other features
−0.5 −0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2
SHAP value (impact on model output)
High
Feature value
Low
B
Age +0.11
Waist +0.11
BMI +0.06
Gender +0.05
DBP +0.03
Educational level +0.03
SBP +0.02
PHQ-Score +0.02
PIR +0.02
Race +0.01
Sum of 9 other features +0.03
0.00 0.02 0.04 0.06 0.08 0.10 0.12
mean(|SHAP value|)
C
Waist
Age
BMI
Gender
DBP
SBP
Educational level
PHQ-Score
PIR
Race
Sum of 9 other features
−6 −4 −2 0 2 4 6
SHAP value (impact on model output)
High
Feature value
Low
D
Waist +2.1
Age +1.7
BMI +0.67
Gender +0.61
DBP +0.6
SBP +0.58
Educational level +0.51
PHQ-Score +0.46
PIR +0.37
Race +0.22
Sum of 9 other features +0.48
0.0 0.5 1.0 1.5 2.0
mean(|SHAP value|)