SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(3), 1461.
https://doi.org/10.62617/mcb1461

Article

Machine learning-based diagnosis of Type 2 Diabetes Mellitus using Social Determinants of Health

Guopeng Hu[1,*,†], Lihan Lin[1,†], Xiangju Hu[2,3], Yikun Zheng[1], Xiaoyang Liu[1], Zhenduo Xu[4], Yuqi He[1], Yinghui Zhang[1]

[1] College of Physical Education, Huaqiao University, Quanzhou 362021, China
[2] School of Public Health, Fujian Medical University, Fuzhou 350005, China
[3] Department for Chronic and Noncommunicable Disease Control and Prevention, Fujian Provincial Center for Disease Control and Prevention, Fuzhou 350001, China
[4] Institute of Advanced Manufacturing, Shantou Polytechnic, Shaotou 515078, China
*** Corresponding author:** GuoPeng Hu, hugp@hqu.edu.cn
[†] These authors contributed equally to this work

CITATION

Hu G, Lin L, Hu X, et al. Machine learning-based diagnosis of Type 2 Diabetes Mellitus using Social Determinants of Health. Molecular & Cellular Biomechanics. 2025; 22(3): 1461.
https://doi.org/10.62617/mcb1461

ARTICLE INFO

Received: 25 January 2025
Accepted: 21 February 2025
Available online: 27 February 2025

Abstract: Background: In China, half of Type 2 Diabetes Mellitus (T2DM) cases remain undiagnosed, worsening patient health and increasing complication risks and socioeconomic burdens. This study aims to develop a T2DM prediction model by integrating machine learning (ML) methods with Social Determinants of Health (SDoH) data from Fujian Province, China. **Methods:** This study utilized a cross-sectional design and multi-stage cluster random sampling to assess SDoH and T2DM prevalence in 26,298 participants from April 2019 to April 2020 in Fujian, China. To predict T2DM, the study leveraged 5 machine learning algorithms—Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM), with the Synthesized Minority Oversampling Technique (SMOTE) algorithm balancing samples. hyperparameters were tuned through RandomizedSearchCV and GridSearchCV to obtain optimal parameters. Model evaluation metrics included accuracy, recall, precision, Area under Curve (AUC) and *F*1 Score. SHapley Additive exPlanations (SHAP) analysis elucidated the impact of specific SDoH variables on T2DM risk prediction. **Results:** Among the 26,298 participants in the study population, the mean (SD) age was 53.77 years (14.41) and 13.99% were T2DM (*N* = 3680). All ML models had AUC values above 0.70, with LightGBM performing best (AUC 0.723, Accuracy 0.659, Recall 0.709, Precision 0.641). SHAP analysis showed that older age and higher Body Mass Index (BMI) significantly increases diabetes risk, along with hypertension, poor self-rated health, and dyslipidemia. **Conclusion:** The predictive model, combined with SDoH data, provides a non-invasive, efficient, and low-cost tool for T2DM prediction, targeting China's large undiagnosed diabetic population. Key factors influencing the model include older age, higher BMI, hypertension, dyslipidemia, and urban residency, which are critical T2DM risk factors. This model supports early detection and targeted interventions, helping to reduce healthcare burdens in resource-limited settings.

Keywords: Type 2 Diabetes prediction; risk factors; predictive medicine; Noncommunicable Disease; public health

1. Introduction

Diabetes Mellitus (DM) is a metabolic disorder characterized by defects in insulin secretion or impaired insulin action. Type 2 Diabetes Mellitus (T2DM) is the most prevalent, accounting for approximately 90% of cases [1]. With rapid population growth in recent years [2], along with issues like urbanization and aging, the number

of people with DM in China has significantly increased, rising from 98.4 million to 140.9 million between 2013 and 2021 [3,4]. Furthermore, due to the complex pathogenesis of DM, insufficient health education, and inadequate healthcare resources in rural areas, the undiagnosed rate in China is high, with 60% in urban areas and 80% in rural areas [5,6]. This worsens public health risks and economic burdens, as delayed or undiagnosed T2DM leads to complications like cardiovascular disease, kidney failure, and neuropathy, which require costly treatments and threaten life expectancy, while also severely impacting quality of life through chronic pain, disability, and reduced mobility [7]. Therefore, providing a rapid, low-cost, and effective screening method for the substantial potential T2DM patient population in China is of paramount importance.

ML-based T2DM prediction models have demonstrated cost-effectiveness, but many existing ML models rely on biomedical data, such as blood glucose, HbA1c, and triglyceride levels, all of which need to be measured in a medical laboratory [8]. This limitation significantly narrows the applicability of machine learning prediction models, especially in certain underdeveloped medical regions in China.

In contrast, utilizing Social Determinants of Health (SDoH) as data inputs in predictive models represents a groundbreaking approach to diagnosing and forecasting T2DM. SDoH encompasses a wide array of conditions affecting individuals' lives, including socioeconomic status, education levels, living and working conditions, access to healthcare, and community environments [9–12]. Research shows that in China, populations with lower education levels and socioeconomic development [13–15], along with unhealthy lifestyle behaviors such as poor diet, lack of physical activity (PA), and insufficient sleep, experience higher rates of diabetes incidence [16,17]. Studies in northern rural China have developed T2DM risk prediction models using SDoH factors such as age, weight, obesity, family history, dietary habits, and hypertension [18]. Another study incorporated factors like age, gender, education, income, marital status, and diet to predict T2DM risk [19], highlighting the influence of these factors on diabetes incidence. These findings support integrating SDoH, which are relatively low-cost and accessible predictive data, into machine learning models to enhance T2DM prediction accuracy.

This study aims to develop a machine learning-based model for screening T2DM, leveraging SDoH over invasive medical tests for quick, accurate, and cost-effective detection. By assessing feature importance, the model evaluates how these determinants influence diabetes risk, informing targeted interventions. As shown in **Figure 1**, this approach provides a more efficient alternative to conventional screenings by prioritizing prompt, non-invasive detection. In particular, in rural China, where medical resources are scarce, this model can effectively identify undiagnosed T2DM patients and individuals at high risk of developing T2DM. This enables early blood sugar management, targeted prevention strategies, and timely treatment, which are crucial for reducing disease burden and improving public health outcomes.

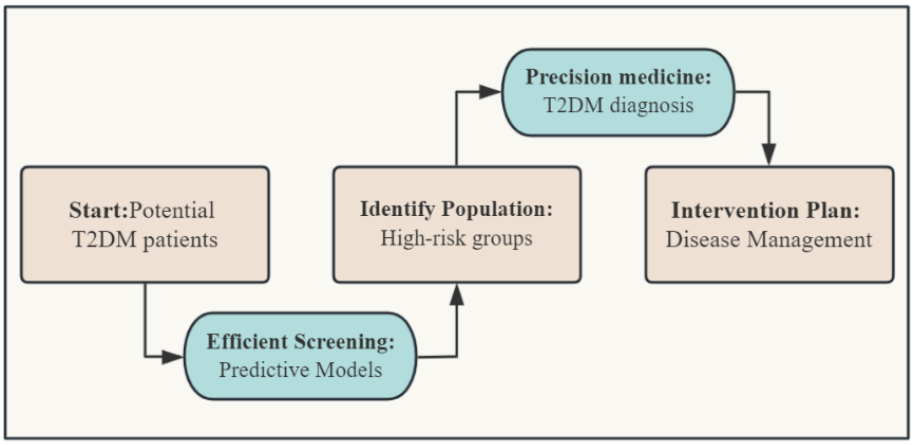

Figure 1. Flowchart of T2DM risk predictive model strategy.

2. Materials and method

2.1. Data source

This study was based on the Chinese Adults Noncommunicable Disease and Nutrition Surveillance (Fujian segment), a cross-sectional study investigating SDoH and T2DM among adults in Fujian Province, China. The baseline dataset used in this study was collected from April 2019 to April 2020 (**Table S4**. Raw data). All participants were fully informed of the study's purpose, procedures, potential risks, and benefits prior to their participation. Each participant provided their consent by signing a written informed consent form, thereby confirming their understanding and agreement to participate under the outlined conditions.

In the first phase, a total of 16 administrative regions, comprising 5 districts and 11 counties, were selected from the 86 administrative divisions of Fujian Province through a probability proportionate to size (PPS) sampling method. This approach utilized population data from the Fujian Province Population Annual Report (2020) to ensure proportional representation of areas with larger populations. This method ensured that districts and counties with larger populations had a higher chance of being selected, thus reflecting the demographic diversity of the province. Districts, streets, and communities represent the three levels of the urban population structure, while counties, townships, and villages form the three levels of the rural population structure.

In the second stage, within each selected district or county, 6 townships (or streets) were randomly selected using the same method. In the third stage, within each selected township (or street), 3 village committees (or communities) were chosen using simple random sampling, with each having at least 100 households. In the fourth and final stage, within each selected household, 1 individual was surveyed, with the sample size calculated using the Kish Leslie formula [20]. The survey targeted permanent residents of Fujian Province who had lived in the survey areas for 6 months or more, were aged 18 years or older, and excluded pregnant women and individuals with cognitive or language impairments.

A national standard questionnaire, the Social Factors Special Survey Form (SFSSF-2019, shown in **File S6**), developed by the Chinese Center for Disease Control and Prevention, was utilized. This questionnaire combined face-to-face interviews

with medical examinations. Uniformly calibrated instruments were employed for physical and laboratory exams to measure blood pressure, blood lipids, height, and weight, as referenced in studies such as [21–24]. Such as, blood pressure was measured using the Omron HBP-1300 electronic blood pressure monitor on the right upper arm [25]. Measurements were taken three times in a resting state, with intervals of more than five minutes between each measurement. Blood glucose levels were determined using fasting plasma glucose (FPG) [26] and a 2-hour post-75 g oral glucose tolerance test (OGTT) [27] venous blood samples (participants with a history of diabetes did not undergo the glucose challenge).

2.2. Study population

Participants were excluded if they met any of the following criteria: 1) Missing information on age, Body Mass Index (BMI), daily sleep duration, sedentary time and the diagnosis of T2DM; 2) more than 10% of SDoH variables were missing [28,29]. Ultimately, 26,298 participants were analyzed, among whom 3680 were diagnosed with T2DM during the study, including both pre-existing cases and newly identified ones, yielding a proportion of 13.99% (3680/26,298).

2.3. Dependent variable

The diagnosis of T2DM was defined by the criterion issued by the American Diabetes Association (ADA) [30], which was defined as FPG $\geq$ 7.0 mmol/L and/or 2Hrs OGTT $\geq$ 11.1 mmol/L. An equivalent diagnosis of T2DM, previously established by a top-tier (Tier 3) hospital in China, is similarly acknowledged as a valid determination of the condition.

2.4. SDoH variables

To develop a T2DM prediction model driven by SDoH data, this study combined findings from a literature review [31-34] with data collected in Fujian Province using the SFSSF-2019, **File S3** Methods supplement has provided more details on the decision-making process for variable inclusion. After excluding variables with over 10% missing data, 19 SDoH variables associated with T2DM prevalence were selected, prioritizing their ease of collection, particularly in regions with limited medical resources. These variables were categorized into three groups: Demographic variables, lifestyle variables, and physiological health variables, as shown in **Figure 2**.

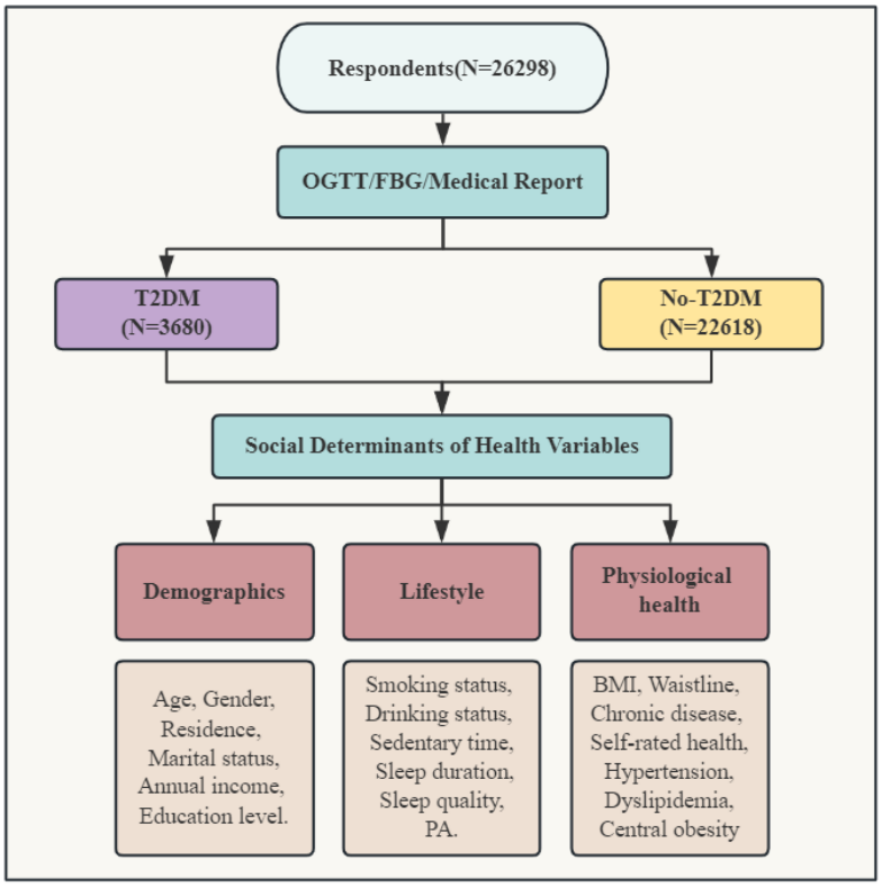

Figure 2. The SDoH variables selected.

The categorization of annual income levels was based on the distribution of per capita disposable income for 2023 as reported by the National Bureau of Statistics of China, with classifications into three tiers: Lower income (under 20,442 yuan, approximately under $2788), Middle income (20,442–50,220 yuan, approximately $2788–$6850), and Higher income (over 50,220 yuan, approximately over $6850). Central obesity was defined according to the "Guidelines for the Prevention and Control of Overweight and Obesity in Chinese Adults", with a waist circumference ≥ 90 cm for men or ≥ 85 cm for women indicating central obesity [35]. The data on sleep quality (categorized as Bad, Average, or Good) and sedentary time (measured in hours per day) were obtained through face-to-face interviews using SFSSF-2019 with participants. The definition of chronic disease incidence included the presence of coronary heart disease, malignant tumors, chronic digestive system diseases, neck and lumbar diseases, chronic obstructive pulmonary disease (COPD), osteoarthritis, cerebrovascular disease, and chronic urinary system diseases, among eight types of chronic conditions. Physical activity (PA) levels were classified into low, moderate, and high based on the scoring rules of the International Physical Activity Questionnaire (IPAQ) Short Form [36]. The remaining SDoH variables are shown in **Table 1** below.

Table 1. SDoH variables assignment.

Variable name	Variable assignment
Dependent variable	
T2DM	No T2DM = 0, T2DM = 1
Demographic variable	
Age	Continuous variables
BMI	Continuous variables
Residence	Rural area = 0, Urban area = 1
Gender	Female = 0, Male = 1
Education	Below primary school = 0, Primary school = 1, Junior high school = 2, Junior college and above = 3
Annual income	Lower income = 0 (under 20,442 yuan); Middle income = 1 (20,442 ≤ income ≤ 50,220); Higher income = 2 (over 50,220 yuan)
Marital status	Solitary = 0, Cohabitation = 1
Medical insurance	None = 0, Yes = 1
Lifestyle variable	
Daily sleep duration	Continuous variables
Sleep quality	Bad = 0, Average = 1, Good = 2
Sedentary time	Continuous variables
PA	Low PA = 0, Moderate PA = 1, High PA = 2
Drinking status	< 1/month, 1–4/month, Every week
Smoking status	Never = 0, Former = 1, Current = 2
Physiological health variable	
Chronic disease	None = 0, Yes = 1
Hypertension	None = 0, Yes = 1
Self-rated health	Bad = 0, Average = 1, Good = 2
Dyslipidemia	None = 0, Yes = 1
Central obesity	None = 0, Yes = 1

2.5. Machine learning algorithms

This study rigorously followed the TRIPOD process to construct prediction models [37]. We selected 5 machine learning algorithms—LR [38], SVM [39], RF [40], XGBoost [41] and LightGBM [42]—to build the prediction of T2DM through the analysis of SDoH variables. Detailed descriptions of these algorithms are provided in **Table S1**.

First, we preprocessed the data by detecting and removing outliers using the Isolation Forest algorithm [43]. Subsequently, missing data were imputed using the "IterativeImputer" function from scikit-learn, following the approach of previous studies [44–46], Detailed information on the number of outliers excluded and the amount of missing data imputed can be found in the **File S3** Methods supplement. The original dataset was split into training and test sets in a 7:3 ratio. To avoid data leakage and biased results, we performed data imputation on the training set only. To address the class imbalance, where T2DM cases were underrepresented (3680/26,298), we employed Synthesized Minority Oversampling Technique (SMOTE) to oversample the minority class in the training set. This balanced the dataset without affecting the

test set, preserving model generalizability. SMOTE improved model robustness and accuracy, particularly for predicting T2DM in imbalanced datasets, enhancing the reliability of our risk prediction [47]. In the training set, machine learning models were trained, and hyperparameters were tuned through RandomizedSearchCV and GridSearchCV to obtain optimal parameters [48,49].

Model performance is evaluated using a comprehensive set of metrics, including accuracy, recall, precision, *F*1 Score, Receiver operating characteristic Area under the Receiver Operating Characteristic Curve curves (AUC-ROC) measured on the test dataset [50]. Model interpretation and feature importance scores were calculated and represented via SHAP values from the optimal prediction model [51,52]. By assessing the incremental contribution of each feature's value against a baseline, SHAP values offer a rigorous quantification of feature impact on specific predictions. The application of Shapley values in predictive modeling enables a granular quantification of each predictor variable's influence, enhancing the understanding of feature importance. These ML modeling methodologies are depicted in **Figure 3** below.

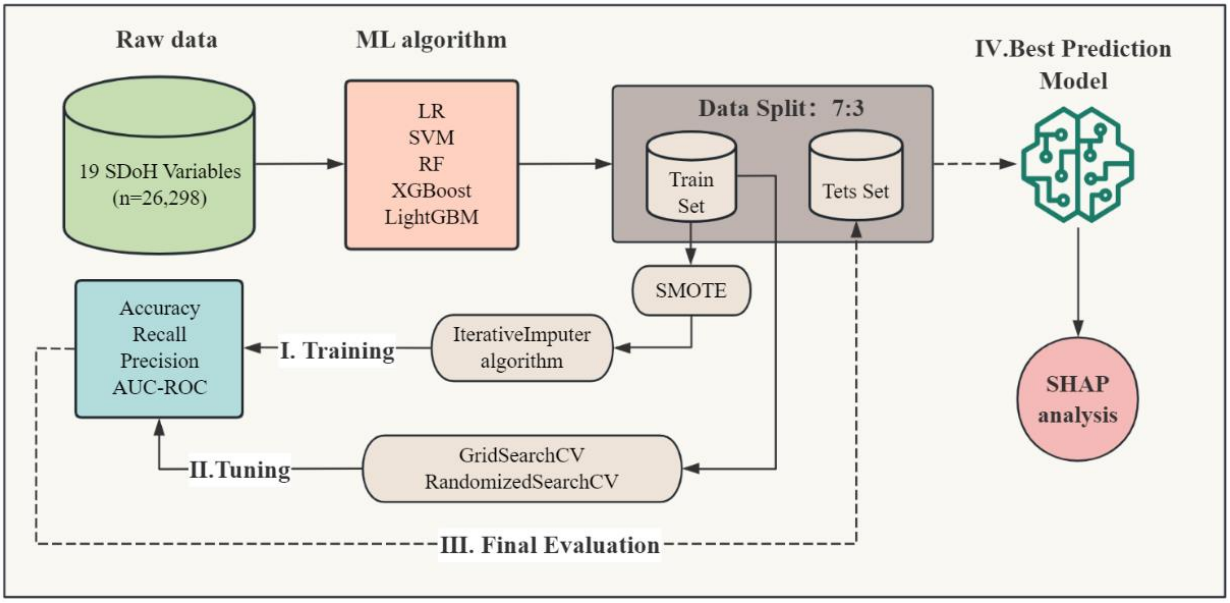

Figure 3. The flow chart of the ML modeling process.

2.6. Statistical methods

Data preprocessing and the construction of ML models were completed in Python 3.9 using Sklearn, NumPy, Matplotlib, and Pandas (in **File S7**). Descriptive statistics in this study utilized SPSS 26.0, presenting continuous variables as mean (SD) or mean (range) and categorical variables as percentages. Continuous variables were subjected to t-tests, while categorical variables underwent chi-square tests, with a $P < 0.05$ deemed to indicate statistical significance.

2.7. Ethics

This research is a branch of Chinese Adults Noncommunicable Disease and Nutrition Surveillance project, conducted within the Fujian province of China. The project is led by the National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and has

received approval from its ethics committee (#201819, **File S5**). Further, we assure that all sociodemographic data utilized for disease prediction were anonymized and de-identified prior to analysis. This precautionary measure was taken to ensure that individual participants could not be identified, thereby safeguarding patient confidentiality. This process underscores our commitment to protecting the privacy and confidentiality of all participants involved in our study, aligning with the ethical standards set forth by the overseeing ethical committee.

3. Results

3.1. Baseline characteristics of dataset

Among the 26,298 participants, a significant proportion of T2DM respondents (n = 3,680) had not pursued education beyond primary school (69.5%). For T2DM participants, the mean BMI was 24.75 (SD 3.56), significantly exceeding the mean BMI of 23.43 (SD 3.28) in non-T2DM individuals, and the average age was 60.79 (SD 11.85), markedly higher than the 52.63 (SD 14.46) of non-T2DM participants. Furthermore, a significant fraction of T2DM individuals, representing 65.2%, belong to the low-income tier. Furthermore, gender and marital status, having *P*-values above 0.05, are excluded as predictors in the ML model due to insufficient statistical significance. These findings are elaborated in **Table 2** below.

Table 2. SDoH variables descriptive.

	Participants, No. (%)				
		No T2DM	**T2DM**	**Total**	
Variable	**Classification**	**(N = 22,618)**	**(N = 3680)**	**(N = 26,298)**	***P* value**
Age, mean (SD)	-	52.63 (14.46)	60.79 (11.85)	53.77 (14.41)	< 0.001
BMI, mean (SD)	-	23.43 (3.28)	24.75 (3.56)	23.61 (3.36)	< 0.001
Residence	Rural	16,084 (71.1)	2354 (64.0)	18,438 (70.1)	< 0.001
	Urban	6534 (28.9)	1326 (36.0)	7860 (29.9)	
Gender	Female	12,527 (55.40)	1991 (54.10)	14,518 (55.20)	0.152
	Maled	10,091 (44.60)	1689 (45.90)	11,780 (44.80)	
Education	Below primary school	9273 (41.00)	1924 (52.30)	11,197 (42.60)	< 0.001
	Primary school	3820 (16.90)	634 (17.20)	4454 (16.90)	
	Junior high school	4919 (21.70)	689 (1d8.70)	5608 (21.30)	
	Junior college and above	4606 (20.40)	433 (11.80)	5039 (19.20)	
Annual income	Lower income	13,864 (61.30)	2399 (65.20)	16,263 (61.80)	< 0.001
	Middle income	3908 (17.30)	582 (15.80)	4490 (17.10)	
	Higher income	4846 (21.40)	699 (19.00)	5545 (21.10)	
Marital status	Solitary	3052 (13.50)	610 (16.60)	3662 (13.90)	< 0.001
	Cohabitation	19,566 (86.50)	3070 (83.40)	22,636 (86.10)	
Medical insurance	None	104 (0.50)	11 (0.30)	115 (0.40)	0.216
	Yes	22,514 (99.50)	3669 (99.70)	26,183 (99.60)	
Daily sleep duration, mean (SD)	-	7.21 (1.49)	7.14 (1.67)	7.20 (1.52)	0.009

Table 2. (*Continued*).

	Participants, No. (%)				
		No T2DM	**T2DM**	**Total**	
Variable	**Classification**	**(N = 22,618)**	**(N = 3680)**	**(N = 26,298)**	***P* value**
Sleep quality	Bad	3098 (13.70)	617 (16.80)	3715 (14.10)	< 0.001
	Average	7590 (33.60)	1305 (35.50)	8895 (33.80)	
	Good	11,930 (52.70)	1758 (47.80)	13,688 (52.00)	
Sedentary time, mean (SD)	-	4.49 (2.91)	4.64 (3.02)	4.51 (2.93)	0.003
PA	Low PA	15,352 (67.90)	2407 (65.40)	17,759 (67.50)	0.007
	Moderate PA	4868 (21.50)	833 (22.60)	5701 (21.70)	
	High PA	2398 (10.60)	440 (12.00)	2838 (10.80)	
Drinking status	< 1/month	15,988 (70.70)	2723 (74.00)	18,711 (71.10)	< 0.001
	1–4/month	4055 (17.90)	520 (14.10)	4575 (17.40)	
	Every week	2575 (11.40)	437 (11.90)	3012 (11.50)	
Smoking status	Never	16,001 (70.70)	2572 (69.90)	18,573 (70.60)	< 0.001
	Former	1435 (6.30)	323 (8.80)	1758 (6.70)	
	Current	5182 (22.90)	785 (21.30)	5967 (22.70)	
Chronic disease	None	12,606 (55.70)	1715 (46.60)	14,321 (54.50)	< 0.001
	Yes	10,012 (44.30)	1965 (53.40)	11,977 (45.50)	
Self-rated health	Bad	1770 (7.80)	565 (15.40)	2335 (8.90)	< 0.001
	Average	10,183 (45.00)	1985 (53.90)	12,168 (46.30)	
	Good	10,665 (47.20)	1130 (30.70)	11,795 (44.90)	
Hypertension	None	16,776 (74.20)	1990 (54.10)	18,766 (71.40)	< 0.001
	Yes	5842 (25.80)	1690 (45.90)	7532 (28.60)	
Dyslipidemia	None	14,078 (62.20)	1727 (46.90)	15,805 (60.10)	< 0.001
	Yes	8540 (37.80)	1953 (53.10)	10,493 (39.90)	
Central obesity	None	16,567 (73.20)	1987 (54.00)	18,554 (70.60)	< 0.001
	Yes	6051 (26.80)	1693 (46.00)	7744 (29.40)	

3.2. Features selection

Random Forest-Recursive Feature Elimination (RF-RFE) is a feature selection technique that iteratively removes the least important features based on their importance scores, identifying a subset that enhances prediction accuracy. In this study, RF-RFE (Random Forest-Recursive Feature Elimination) was used to identify the most predictive variables, with the optimal number of variables determined by maximizing the AUC. A higher AUC indicates better model prediction accuracy. As shown in **Figure 4**, the model achieves the highest AUC when it includes 18 variables. The optimal set of variables selected by the RF-RFE algorithm includes age, hypertension, BMI, central obesity, sedentary time, self-rated health, daily sleep duration, education, dyslipidemia, sleep quality, drinking status, annual income, PA, residence, smoking status, chronic disease, gender, and marital status, while medical insurance was excluded.

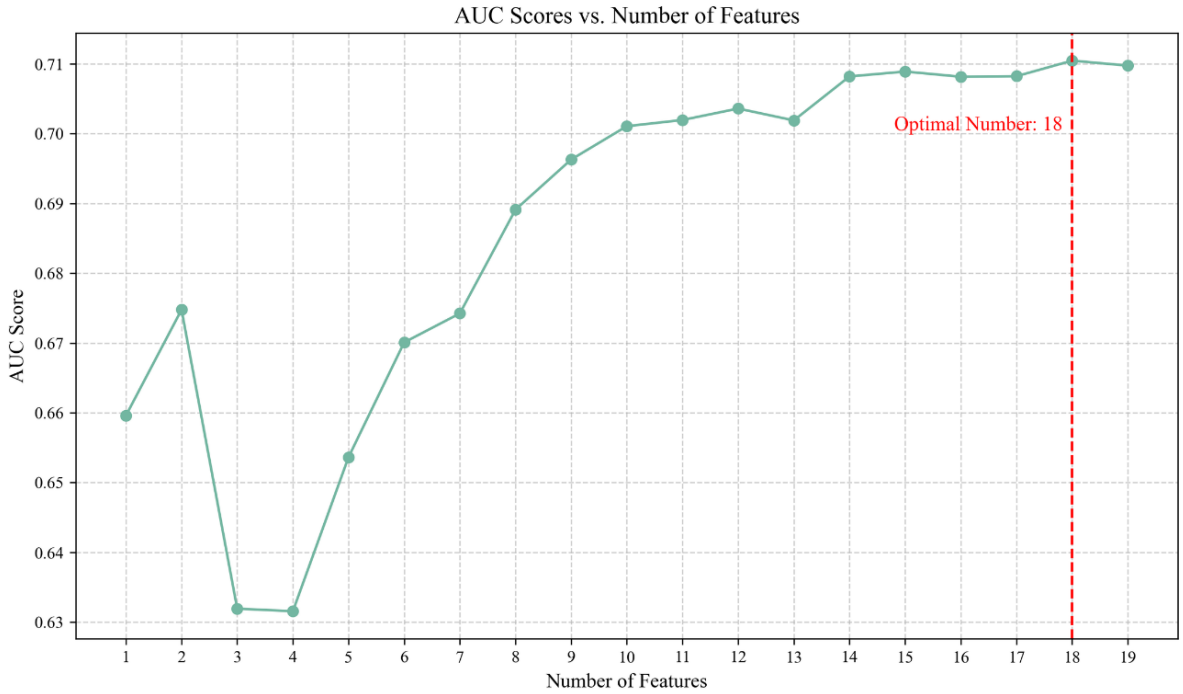

Figure 4. Relationship between the number of features and AUC score for T2DM prediction.

This plot shows the relationship between the number of features and the AUC score for T2DM prediction, with optimal performance achieved at 18 features (red dashed line). AUC refers to the Area Under the Curve, indicating the model's ability to discriminate between classes.

3.3. Model hyperparameter tuning

The hyperparameters of each algorithm are optimized using grid search, where the Best AUC represents the highest evaluated AUC value obtained during the hyperparameter tuning process through 5-fold cross-validation. The Model Hyperparameter Tuning process involves initial tuning by defining a model pipeline with preprocessing steps and the classifier, followed by setting a broad range of hyperparameters. Using RandomizedSearchCV, we perform an initial hyperparameter search, fit the model on the training data, and identify the best parameters. For fine-tuning, we narrow the hyperparameter range based on the initial results and employ GridSearchCV to conduct exhaustive search. The model is then refitted with the refined parameters, and the final best parameters and corresponding scores are reported. This two-step approach efficiently explores and optimizes the hyperparameter space. The hyperparameters at this point are considered the optimal hyperparameters for the model. The optimal hyperparameters for each model, along with their Best Score and default values, are shown in **Table S2**.

3.4. Model evaluation

This study validated the performance of five models on the dataset, with all models achieving AUC-ROC values above 0.70, shown in **Table 3**. The LightGBM model demonstrated the best performance across all metrics. It achieved accuracy of 0.659, recall of 0.709, and precision of 0.641. Additionally, LightGBM had the highest

AUC-ROC value of 0.723, indicating its superior ability to distinguish between positive and negative samples. The AUC-ROC curve for the 5 model is shown in **Figure 5**.

Table 3. The performance of ML models.

Metrics	LR	SVM	RF	XGBoost	LightGBM
Accuracy	0.658	0.658	0.656	0.658	0.659
Recall	0.669	0.706	0.693	0.691	0.709
Precision	0.650	0.640	0.641	0.643	0.641
AUC-ROC	0.718	0.722	0.717	0.722	0.723

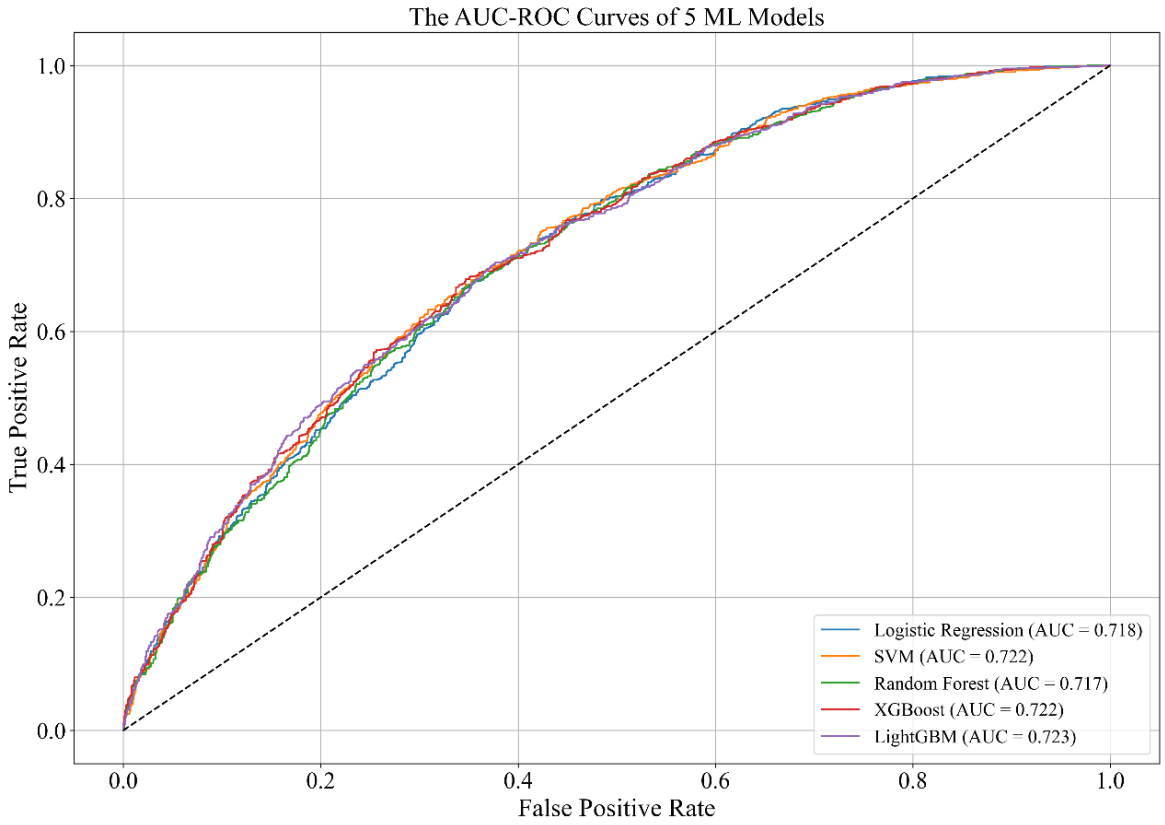

Figure 5. The AUC-ROC curves of ML models.

The AUC-ROC curve compares the performance of various machine learning models in predicting T2DM risk, with LightGBM achieving the highest AUC of 0.723, demonstrating strong discriminatory ability across models.

3.5. SHAP value of ML model

The SHAP values reflect the contribution of each feature to the model's prediction of T2DM risk. According to **Figure 6**, the most significant predictors of T2DM risk include age, BMI, hypertension, dyslipidemia, and residence (urban/rural). Specifically, older age, higher BMI, the presence of hypertension and dyslipidemia, and living in urban areas are all associated with an increased likelihood of developing T2DM. Additionally, poor self-rated health, central obesity, and longer sedentary time are strongly positively correlated with T2DM risk.

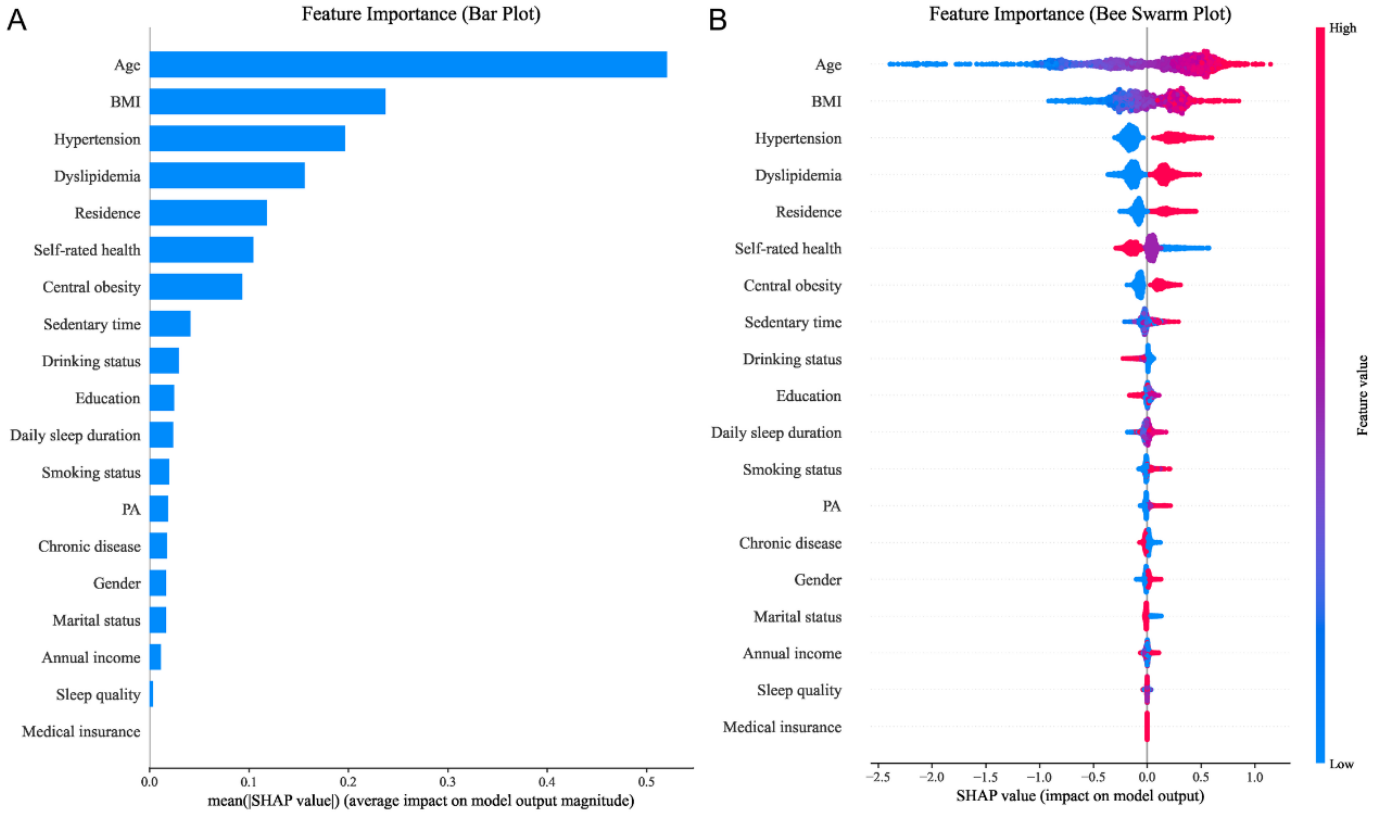

Figure 6. SHAP-based feature importance for predicting T2DM risk using LightGBM model.

Panel A shows feature importance from SHAP values of an XGBoost model predicting T2DM risk, with features ranked by importance along the y-axis. The x-axis represents the mean absolute SHAP values, indicating each feature's average impact on the model. Higher values indicate a greater positive contribution to T2DM risk, while lower values suggest a negative impact. Panel B illustrates how individual feature values impact the model's prediction of T2DM risk. The colors represent feature values, with blue indicating low values and red indicating high values. Positive SHAP values show a positive contribution to T2DM risk, while negative SHAP values indicate a negative impact.

4. Discussion

4.1. Novelty and importance of the T2DM prediction model using SDoH

In this study, T2DM prediction models using SDoH data offer a non-invasive, efficient, and low-cost method for identifying potential T2DM patients. This machine learning-based advancement not only reduces the burden on undiagnosed cases but also promises significant improvements in public health through personalized, proactive healthcare approaches [53–57]. These models incorporate socioeconomic, environmental, and behavioral factors, providing a comprehensive understanding of an individual's T2DM risk. Using non-clinical SDoH data, they are more cost-effective than traditional methods reliant on clinical or biological markers, as SDoH data can be obtained through simple online questionnaires or interviews, minimizing the need for healthcare personnel and medical resources during screening. Traditional

diabetes screening methods, including the FPG, the 2-hour OGTT, or the HbA1c test, are invasive, inconvenient, and expensive. Studies have shown that the cost of diagnosing each case of diabetes is $758 in the United States [58] (no screening strategy) and €831 in Germany [59]. Applying the Chinese Diabetes Risk Score (CDRS) to screen for pre-diabetes in China also costs $299.67 per case [60]. So, for a populous country like China, with a large number of potential T2DM patients, achieving large-scale T2DM screening represents a significant financial burden. This affordability and ease of access to SDoH data make these prediction models particularly valuable in under-resourced areas where healthcare infrastructure may be lacking or where there are significant barriers to healthcare access. By identifying at-risk individuals based on a broader set of determinants, healthcare providers can target interventions more effectively and allocate resources more efficiently, helping to mitigate the impacts of healthcare inequalities and ensuring that preventative measures reach those most in need.

4.2. Comparison with existing Chinese T2DM prediction models

In the realm of T2DM risk prediction, our model represents a paradigm shift from traditional biomarker-centric approaches to a holistic evaluation encompassing SDoH. Contrasting with existing models, our algorithm amalgamates an extensive dataset integrating both clinical and non-clinical parameters, addressing a critical gap for a multifaceted risk assessment suitable for the Chinese population.

Early studies on predictive models for T2DM in the Chinese population, such as those by Zhang et al. [61] and Liu et al. [62], primarily relied on medical examination data including FPG, total cholesterol (TC), triglyceride (TG), high-density lipoprotein (HDL-C), low-density lipoprotein (LDL-C), alanine aminotransferase (ALT), aspartate transaminase (AST), total bilirubin (TBIL), which have limited accessibility. Prevailing non-invasive T2DM prediction models, Zhang et al. [54] made substantial strides with a model for rural populations in Henan Province, underscoring the predictive value of readily obtainable clinical data. Wang et al. [63] contribute significantly to this body of research through the Kailuan prospective study, which employs risk scores to forecast the incidence of T2DM. Xiong et al. [64] conducted a retrospective study in Nanjing and applied machine learning to urban clinical data with a notable degree of success, drawing parallels in performance to our own model. Lastly, Zhang et al. [65] also utilized machine learning in their Henan Rural Cohort Study to identify new risk factors for T2DM, demonstrating the expanding capability of these algorithms in rural settings.

Our research builds upon the groundwork established by these pivotal studies, eliminating predictive indicators that are challenging to obtain in areas with scarce medical resources. By incorporating a broader range of SDoH into our predictive model, we have maintained commendable predictive accuracy, with an AUC value of 0.723, ranking it in the middle tier of previous studies (AUC: 0.65–0.89) [65–67]. This approach underscores the significance of socio-economic and lifestyle variables in assessing T2DM risk, thereby offering a nuanced and comprehensive tool for early detection and intervention strategies within the Chinese context.

4.3. Implications of SDoH T2DM prediction for public health

T2DM prediction models incorporating SDoH reshape public health planning and policymaking by highlighting the critical influence of environmental, economic, and social factors on health outcomes. Our model, focused on the Fujian Province dataset, exemplifies this approach by aligning closely with the region's unique socio-economic conditions and health characteristics. This localized specificity enhances the precision of T2DM screening and prevention in Fujian, underscoring the importance of tailoring models to reflect the comprehensive health landscapes of specific communities.

Utilizing SDoH data for T2DM prediction catalyzes cross-disciplinary collaboration, uniting healthcare, public health, community organizations, and policy sectors. This collaboration is essential for addressing broad health challenges and advocating for policies that simultaneously enhance social, environmental, and health conditions. Our work serves as a case study in the effective use of non-clinical data to predict health outcomes, demonstrating the potential for similar region-specific models to contribute to a cohesive, preventative healthcare system.

These models are pivotal for health monitoring and policy adjustment, allowing for the timely identification of trends and the evaluation of intervention efficacy. By implementing T2DM prediction models using SDoH data, such as our targeted approach in Fujian, we mark a significant stride in public health. This strategy aims to improve health outcomes, reduce disparities, and foster a data-led, preventative healthcare environment that is equitable and universally effective.

4.4. Limitations

The study of T2DM using SDoH encounters limitations that include dataset representativeness, subjective survey responses, and an imbalance of sample sizes. Firstly, the dataset's focus on Fujian province, one of China's fastest-growing economic regions and ranked seventh in Gross Domestic Product (GDP) nationwide, may not fully capture the health conditions and socio-economic variations observed across other regions. Fujian exhibits pronounced internal disparities, with highly developed coastal cities such as Fuzhou and Xiamen coexisting alongside economically underdeveloped mountainous areas. This unique intra-provincial imbalance, which is less prevalent in other provinces, may limit the broader applicability and accuracy of the model. Secondly, the reliance on self-reported data for key variables such as sleep quality and sedentary time introduces a level of subjectivity. This subjectivity could lead to bias, as individual perceptions and reporting accuracy vary, potentially skewing the data and affecting the model's outputs.

5. Conclusion

This study developed a non-invasive, low-cost predictive model for T2DM using SDoH and five machine learning algorithms, with LightGBM showing the best performance (AUC = 0.723). This model offers a reliable tool for T2DM screening, particularly in regions with limited medical resources. The most influential features for model prediction include older age, higher BMI, the presence of hypertension and

dyslipidemia, and living in urban areas, which reflect key risk factors for T2DM. Future research should expand the geographic scope of data collection beyond Fujian Province to improve model generalizability. Additionally, reducing reliance on self-reported data for variables such as sleep quality and self-rated health will enhance input accuracy. Future studies should also explore alternative machine learning models and incorporate environmental and genetic data to provide a more comprehensive view of T2DM risk.

Supplementary materials: S1 Table. Description of Machine Learning Algorithms. (XLSX); S2 Table. Model Hyperparameter Tuning result. (XLSX); S3 File. Methods supplement. (RAR); S4 Table. Raw Dataset. (XLSX); S5 File. Ethical Approval. (PDF); S6 File. SFSSF-2019. (PDF); S7 File. Code for ML models. (RAR).

Author contributions: Conceptualization, GH and LL; methodology, LL; software, LL; validation, GH, XH and LL; formal analysis, LL and XL; investigation, GH; resources, GH and XH; data curation, ZX and XH; writing—original draft preparation, LL and XL; writing—review and editing, GH, LL, XH, YZ, XL, ZX, YH and YZ; visualization, LL; supervision, GH and YH; project administration, GH; funding acquisition, GH and YZ. All authors have read and agreed to the published version of the manuscript.

Funding: This study was sponsored by the Fujian Province Natural Science Foundation, grant number 2020J01087, and Fujian Provincial Social Science Foundation, grant number FJ2021B130.

Acknowledgments: We thank the National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese CDC, for their support of our Fujian branch project under the Chinese Adults Noncommunicable Disease and Nutrition Surveillance. Special gratitude to Li Xinhua for his invaluable guidance. Appreciation extends to our team and participants for their crucial contributions.

Ethical approval: The study was conducted in accordance with the Declaration of Helsinki, and approved by the National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control Ethics Committee (Protocol code: #201819, on April 27, 2018). Informed consent was obtained from all subjects involved in the study.

Conflict of interest: The authors declare no conflict of interest.

References

1. American Diabetes Association Professional Practice Committee. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022. Diabetes Care. 2022; 45: S17–S38. doi: 10.2337/dc2 2-S002
2. GBD 2017 Population and Fertility Collaborators. Population and fertility by age and sex for 195 countries and territories, 1950–2017: A systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018; 392: 1995–2051. doi: 10.1016/S0140-6736(18)32278-5
3. Farrell K, Westlund H. China's rapid urban ascent: An examination into the components of urban growth. Asian Geographer. 2018; 35: 85–106. doi: 10.1080/10225706.2018.1476256
4. Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice. 2022; 183: 109119.

5. Ma RCW. Epidemiology of diabetes and diabetic complications in China. Diabetologia. 2018; 61: 1249–1260. doi: 10.1007/s00125-018-4557-7
6. Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res. Clin. Pract. 2019; 157: 107843. doi: 10.1016/j.diabres.2019.107843
7. Avogaro A, Fadini GP. Microvascular complications in diabetes: A growing concern for cardiologists. International Journal of Cardiology. 2019; 291: 29–35. doi: 10.1016/j.ijcard.2019.02.030
8. Barber SR, Davies MJ, Khunti K, et al. Risk assessment tools for detecting those with pre-diabetes: A systematic review. Diabetes Res. Clin. Pract. 2014; 105: 1–13. doi: 10.1016/j.diabres.2014.03.007
9. Braveman P, Egerter S, Williams DR. The Social Determinants of Health: Coming of Age. Annu. Rev. Public Health. 2011; 32: 381–398. doi: 10.1146/annurev-publhealth-031210-101218
10. Preda A, Voigt K. The Social Determinants of Health: Why Should We Care? The American Journal of Bioethics. 2015; 15: 25–36. doi: 10.1080/15265161.2014.998374
11. Duncan GJ, Daly MC, McDonough P, et al. Optimal Indicators of Socioeconomic Status for Health Research. Am. J. Public Health. 2002; 92: 1151–1157. doi: 10.2105/AJPH.92.7.1151
12. Viner RM, Ozer EM, Denny S, et al. Adolescence and the social determinants of health. The Lancet. 2012; 379: 1641–1652. doi: 10.1016/S0140-6736(12)60149-4
13. Pan XR, Yang WY, Li GW, et al. Prevalence of Diabetes and Its Risk Factors in China, 1994. Diabetes Care. 1997; 20: 1664–1669. doi: 10.2337/diacare.20.11.1664
14. Tao X, Li J, Zhu X, et al. Association between socioeconomic status and metabolic control and diabetes complications: A cross-sectional nationwide study in Chinese adults with type 2 diabetes mellitus. Cardiovasc Diabetol. 2016; 15: 61. doi: 10.1186/s12933-016-0376-7
15. Zhang H, Xu W, Dahl AK, et al. Relation of socio-economic status to impaired fasting glucose and Type 2 diabetes: Findings based on a large population-based cross-sectional study in Tianjin, China. Diabet. Med. 2013; 30. doi: 10.1111/dme.12156
16. Zhang Y, Wang Y, Zhang S, et al. Complex Association Among Diet Styles, Sleep Patterns, and Obesity in Patients with Diabetes. Diabetes Metab. Syndr. Obes. 2023; 16: 749–767. doi: 10.2147/DMSO.S390101
17. Li Y, Wang DD, Ley SH, et al. Time Trends of Dietary and Lifestyle Factors and Their Potential Impact on Diabetes Burden in China. Diabetes Care. 2017; 40: 1685–1694. doi: 10.2337/dc17-0571
18. Chen X, Wu Z, Chen Y, et al. Risk score model of type 2 diabetes prediction for rural Chinese adults: The Rural Deqing Cohort Study. J. Endocrinol. Invest. 2017; 40: 1115–1123. doi: 10.1007/s40618-017-0680-4
19. Shao X, Wang Y, Huang S, et al. Development and validation of a prediction model estimating the 10-year risk for type 2 diabetes in China. PLoS ONE. 2020; 15: e0237936. doi: 10.1371/journal.pone.0237936
20. Kish L. Sampling Organizations and Groups of Unequal Sizes. Am. Sociol. Rev. 1965; 30: 564. doi: 10.2307/2091346
21. Yu W, Li X, Zhong W, et al. Rural-urban disparities in the associations of residential greenness with diabetes and prediabetes among adults in southeastern China. Science of the Total Environment. 2023; 860: 160492. doi: 10.1016/j.scitotenv.2022.160492
22. Huang S, Lin X, Yin P, et al. Assessment of disability weights at the provincial and city levels based on 93,254 respondents in Fujian, China: Findings from the Fujian disability weight measurement study. Chinese Medical Journal. 2024; 137: 1375–1377. doi: 10.1097/CM9.0000000000002812
23. Xie XX, Zhou WM, Lin F, et al. Ischemic heart disease deaths, disability-adjusted life years and risk factors in Fujian, China during 1990–2013: Data from the Global Burden of Disease Study 2013. International Journal of Cardiology. 2016; 214: 265–269. doi: 10.1016/j.ijcard.2016.03.236
24. Hu X, Fang X, Wu M. Prevalence, awareness, treatment and control of type 2 diabetes in southeast China: A population-based study. J. of Diabetes Invest. 2024; 15(8): 1034-1041. doi: 10.1111/jdi.14213
25. Meng L, Zhao D, Pan Y, et al. Validation of Omron HBP-1300 professional blood pressure monitor based on auscultation in children and adults. BMC Cardiovasc. Disord. 2016; 16: 9. doi: 10.1186/s12872-015-0177-z
26. Zhou X, Pang Z, Gao W, et al. Performance of an A1C and Fasting Capillary Blood Glucose Test for Screening Newly Diagnosed Diabetes and Pre-Diabetes Defined by an Oral Glucose Tolerance Test in Qingdao, China. Diabetes Care. 2010; 33: 545–550. doi: 10.2337/dc09-1410

27. Bartoli E, Fra GP, Schianca GPC. The oral glucose tolerance test (OGTT) revisited. Eur. J. Intern. Med. 2011; 22: 8–12. doi: 10.1016/j.ejim.2010.07.008
28. Bennett DA. How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health. 2001; 25: 464–469. doi: 10.1111/j.1467-842X.2001.tb00294.x
29. Chen X, He L, Shi K, et al. Interpretable Machine Learning for Fall Prediction Among Older Adults in China. American Journal of Preventive Medicine. 2023; 65: 579–586. doi: 10.1016/j.amepre.2023.04.006
30. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2014; 37: S81–S90. doi: 10.2337/dc14-S081
31. Zhang N, Du SM, Ma GS. Current lifestyle factors that increase risk of T2DM in China. Eur. J. Clin. Nutr. 2017; 71: 832–838. doi: 10.1038/ejcn.2017.41
32. Dendup T, Feng X, Clingan S, et al. Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. Int. J. Environ. Res. Public Health. 2018; 15: 78. doi: 10.3390/ijerph15010078
33. Lin L, Hu X, Liu X, et al. Key influences on dysglycemia across Fujian's urban-rural divide. PLoS One. 2024 Jul 31;19(7): e0308073. doi: 10.1371/journal.pone.0308073.
34. Wu Y, Ding Y, Tanaka Y, et al. Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. Int. J. Med. Sci. 2014; 11: 1185–1200. doi: 10.7150/ijms.10001
35. Wei J, Liu X, Xue H, et al. Comparisons of Visceral Adiposity Index, Body Shape Index, Body Mass Index and Waist Circumference and Their Associations with Diabetes Mellitus in Adults. Nutrients. 2019; 11: 1580. doi: 10.3390/nu11071580
36. Lee PH, Macfarlane DJ, Lam T, et al. Validity of the international physical activity questionnaire short form (IPAQ-SF): A systematic review. Int. J. Behav. Nutr. Phy. 2011; 8: 115. doi: 10.1186/1479-5868-8-115
37. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. Circulation. 2015; 131: 211–219. doi: 10.1161/CIRCULATIONAHA.114.014508
38. Sperandei S. Understanding logistic regression analysis. Biochem. Med. 2014; 12–18. doi: 10.11613/BM.2014.003
39. Noble WS. What is a support vector machine? Nat. Biotechnol. 2006; 24: 1565–1567. doi: 10.1038/nbt1206-1565
40. Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32. doi: 10.1023/A:1010933404324
41. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17 August 2016; San Francisco, CA, USA. pp. 785–794.
42. Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, et al. (editors). Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.
43. Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining; 15–19 December 2008; Pisa, Italy. pp. 413–422.
44. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning Research. 2011; 12: 2825–2830.
45. McAndrew T, Codi A, Cambeiro J, et al. Chimeric forecasting: Combining probabilistic predictions from computational models and human judgment. BMC Infect. Dis. 2022; 22: 833. doi: 10.1186/s12879-022-07794-5
46. Younus S, Rönnstrand L, Kazi JU. Xputer: Bridging data gaps with NMF, XGBoost, and a streamlined GUI experience. Front. Artif. Intell. 2024; 7: 1345179. doi: 10.3389/frai.2024.1345179
47. Han H, Wang WY, Mao BH. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang DS, Zhang XP, Huang GB (editors). Advances in Intelligent Computing. Springer; 2005. pp. 878–887.
48. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 2012; 13: 281–305.
49. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence; 20–25 August 1995; Montreal, Canada. pp. 1137–1145.
50. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015; 349: 255–260. doi: 10.1126/science.aaa8415
51. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017; 30.
52. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2020; 2: 56–67. doi: 10.1038/s42256-019-0138-9

53. Pham TD. Classification of COVID-19 chest X-rays with deep learning: New models or fine tuning? Health Inf. Sci. Syst. 2021; 9: 2. doi: 10.1007/s13755-020-00135-3
54. Khanagar SB, Alkadi L, Alghilan MA, et al. Application and Performance of Artificial Intelligence (AI) in Oral Cancer Diagnosis and Prediction Using Histopathological Images: A Systematic Review. Biomedicines. 2023; 11: 1612. doi: 10.3390/biomedicines11061612
55. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: A systematic review. Journal of the American Medical Informatics Association. 2021; 28: 2716–2727. doi: 10.1093/jamia/ocab170
56. Segar MW, Hall JL, Jhund PS, et al. Machine Learning–Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients with Heart Failure. JAMA Cardiol. 2022; 7: 844. doi: 10.1001/jamacardio.2022.1900
57. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. Journal of the American Medical Informatics Association. 2020; 27: 1764–1773. doi: 10.1093/jamia/ocaa143
58. Zhang P, Engelgau MM, Valdez R, et al. Costs of Screening for Pre-diabetes Among U.S. Adults. Diabetes Care. 2003; 26: 2536–2542. doi: 10.2337/diacare.26.9.2536
59. Icks A, Haastert B, Gandjour A, et al. Cost-Effectiveness Analysis of Different Screening Procedures for Type 2 Diabetes. Diabetes Care. 2004; 27: 2120–2128. doi: 10.2337/diacare.27.9.2120
60. Hao J, Yao Q, Lin Y, et al. Cost-effectiveness of two screening strategies based on Chinese diabetes risk score for pre-diabetes in China. Front. Public. Health. 2022; 10: 1018084. doi: 10.3389/fpubh.2022.1018084
61. Zhang L, Wang Y, Niu M, et al. Nonlaboratory-Based Risk Assessment Model for Type 2 Diabetes Mellitus Screening in Chinese Rural Population: A Joint Bagging-Boosting Model. IEEE J Biomed Health Inform. 2021; 25: 4005–4016. doi: 10.1109/JBHI.2021.3077114
62. Liu Q, Zhang M, He Y, et al. Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques. J. Pers. Med. 2022;12: 905. doi: 10.3390/jpm12060905
63. Wang A, Chen G, Su Z, et al. Risk scores for predicting incidence of type 2 diabetes in the Chinese population: The Kailuan prospective study. Sci. Rep. 2016; 6: 26548. doi: 10.1038/srep26548
64. Xiong XL, Zhang RX, Bi Y, et al. Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Study in Chinese Adults. Curr. Med. Sci. 2019; 39: 582–588. doi: 10.1007/s11596-019-2077-4
65. Zhang L, Wang Y, Niu M, et al. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. Sci. Rep. 2020; 10: 4406. doi: 10.1038/s41598-020-61123-x
66. Wang Y, Zhang L, Niu M, et al. Genetic Risk Score Increased Discriminant Efficiency of Predictive Models for Type 2 Diabetes Mellitus Using Machine Learning: Cohort Study. Front. Public. Health. 2021; 9: 606711. doi: 10.3389/fpubh.2021.606711
67. Wang H, Liu T, Qiu Q, et al. Development and validation of a simple risk score for prevalent undiagnosed type 2 diabetes in Southern Chinese population. Int. J. Diabetes Dev. Ctries. 2015; 35: 318–326. doi: 10.1007/s13410-014-0285-9
68. Awa WL, Fach E, Krakow D, et al. Type 2 diabetes from pediatric to geriatric age: analysis of gender and obesity among 120183 patients from the German/Austrian DPV database. European Journal of Endocrinology. 2012; 167(2): 245-254. doi: 10.1530/eje-12-0143
69. Seiglie JA, Marcus ME, Ebert C, et al. Diabetes Prevalence and Its Relationship With Education, Wealth, and BMI in 29 Low- and Middle-Income Countries. Diabetes Care. 2020; 43(4): 767-775. doi: 10.2337/dc19-1782
70. Kivimäki M, Virtanen M, Kawachi I, et al. Long working hours, socioeconomic status, and the risk of incident type 2 diabetes: a meta-analysis of published and unpublished data from 222 120 individuals. The Lancet Diabetes & Endocrinology. 2015; 3(1): 27-34. doi: 10.1016/S2213-8587(14)70178-0
71. Wang S, Ma W, Yuan Z, et al. Association between obesity indices and type 2 diabetes mellitus among middle-aged and elderly people in Jinan, China: a cross-sectional study. BMJ Open. 2016; 6(11): e012742. doi: 10.1136/bmjopen-2016-012742
72. Beulens JWJ, Pinho MGM, Abreu TC, et al. Environmental risk factors of type 2 diabetes—an exposome approach. Diabetologia. 2021; 65(2): 263-274. doi: 10.1007/s00125-021-05618-w

73. Gassasse Z, Smith D, Finer S, et al. Association between urbanisation and type 2 diabetes: an ecological study. BMJ Global Health. 2017; 2(4): e000473. doi: 10.1136/bmjgh-2017-000473
74. Karimi MA, Binaei S, Hashemi SH, et al. Marital status and risk of type 2 diabetes among middle-aged and elderly population: a systematic review and meta-analysis. Frontiers in Medicine. 2025; 11. doi: 10.3389/fmed.2024.1485490
75. Cornelis MC, Chiuve SE, Glymour MM, et al. Bachelors, Divorcees, and Widowers: Does Marriage Protect Men from Type 2 Diabetes? Sen U, ed. PLoS ONE. 2014; 9(9): e106720. doi: 10.1371/journal.pone.0106720
76. Knutson KL. Role of Sleep Duration and Quality in the Risk and Severity of Type 2 Diabetes Mellitus. Archives of Internal Medicine. 2006; 166(16): 1768. doi: 10.1001/archinte.166.16.1768
77. Shan Z, Ma H, Xie M, et al. Sleep Duration and Risk of Type 2 Diabetes: A Meta-analysis of Prospective Studies. Diabetes Care. 2015; 38(3): 529-537. doi: 10.2337/dc14-2073
78. Joseph JJ, Echouffo-Tcheugui JB, Golden SH, et al. Physical activity, sedentary behaviors and the incidence of type 2 diabetes mellitus: the Multi-Ethnic Study of Atherosclerosis (MESA). BMJ Open Diabetes Research & Care. 2016; 4(1): e000185. doi: 10.1136/bmjdrc-2015-000185
79. Deng MG, Cui HT, Lan YB, et al. Physical activity, sedentary behavior, and the risk of type 2 diabetes: A two-sample Mendelian Randomization analysis in the European population. Frontiers in Endocrinology. 2022; 13. doi: 10.3389/fendo.2022.964132
80. Noh JW, Chang Y, Park M, et al. Self-rated health and the risk of incident type 2 diabetes mellitus: A cohort study. Scientific Reports. 2019; 9(1). doi: 10.1038/s41598-019-40090-y
81. Hayes AJ, Clarke PM, Glasziou PG, et al. Can Self-Rated Health Scores Be Used for Risk Prediction in Patients With Type 2 Diabetes? Diabetes Care. 2008; 31(4): 795-797. doi: 10.2337/dc07-1391