BMJ Open

Learning from the machine: is diabetes in adults predicted by lifestyle variables? A retrospective predictive modelling study of NHANES 2007–2018

Efrain Riveros Perez , Bibiana Avella-Molano

To cite: 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. doi:10.1136/bmjopen-2024-096595

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-096595).

Received 14 November 2024
Accepted 07 March 2025

Augusta University Medical College of Georgia, Augusta, Georgia, USA

Correspondence to
Dr Efrain Riveros Perez; efrainriveros@gmail.com

ABSTRACT

Objectives This study aimed to compare the performance of five machine learning algorithms to predict diabetes mellitus based on lifestyle factors (diet and physical activity).

Design Retrospective cross-sectional predictive modelling study.

Setting This study was conducted using publicly available data from the National Health and Nutrition Examination Survey (NHANES), a nationally representative survey designed to assess the health and nutritional status of the US population.

Participants We analysed data from 29 509 non-pregnant adults who participated in NHANES between 2007 and 2018.

Primary and secondary outcome measures The primary outcome was the prediction of type 2 diabetes mellitus (T2DM) by self-reported responses based on machine learning models. The performance of five machine learning algorithms (logistic regression, support vector machine, random forest, XGBoost and CatBoost) was evaluated using accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and the area under the receiver operating characteristic curve (AUC). The secondary outcome measures were feature importance and model performance comparison.

Results XGBoost exhibited the highest overall predictive performance (AUC 0.8168), followed by random forest and logistic regression (AUCs around 0.79). In terms of accuracy, logistic regression, XGBoost and random forest performed similarly at approximately 85%. While most models demonstrated high specificity (>97%), the SVM stood out for having the highest sensitivity (58.57%), although with a lower accuracy (62.44%). This trade-off underscores the strength of SVM in identifying more true-positive cases, though at the cost of lower overall classification precision. The random forest model, despite having lower sensitivity (7.15%), provided one of the most balanced performances in terms of specificity and interpretability.

Conclusion The results support the use of machine learning models, particularly XGBoost, for early identification of individuals at risk for T2DM. Despite their limited sensitivity, the high specificity and accuracy underscore these models' potential for non-invasive risk assessment. This study is innovative in its integration of machine learning algorithms to predict type 2 diabetes based solely on non-invasive, easily accessible lifestyle and anthropometric variables, demonstrating the potential of data-driven models for early risk assessment without requiring laboratory tests. Despite the lower sensitivity observed in most models, their high specificity makes them valuable for early screening in clinical and public health settings, where they can be complemented with follow-up assessments or ensemble approaches that optimise the balance between sensitivity and specificity for improved risk stratification.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Large, nationally representative data set enhancing generalisability.
⇒ Robust comparison of multiple machine learning algorithms for type 2 diabetes mellitus.
⇒ Use of cross-validation to minimise overfitting and increase model reliability.
⇒ Diabetes status based on self-reported data, which may introduce recall bias.
⇒ The cross-sectional study design limits inferences about causality.

INTRODUCTION

Diabetes mellitus is a metabolic disorder with a rising prevalence across the globe, affecting around half a billion people, with the number expected to increase by 51% by 2045.[1] In the USA, 9.4% of the population has diabetes, with 90–95% of cases classified as type 2 diabetes mellitus (T2DM).[2] The impact of this disorder on quality of life is widely recognised, as well as its economic burden, with an estimated total cost of $245 billion resulting from direct costs, chronic disability, absenteeism and premature mortality.[3 4]

Development of T2DM has been linked to anthropometric and lifestyle factors, such as body habitus, sedentarism, smoking, unhealthy diet and high alcohol consumption.[5–9] Most of these factors are modifiable through lifestyle changes. Furthermore, since the progression of T2DM is slow and silent,

early detection and identification of individuals at high risk are of utmost importance to implement lifestyle strategies aimed at preventing the development and progression of the disease.

Machine learning algorithms have been used as predictive tools for a wide variety of health problems, including T2DM.[10–12] Most of these models focus on predictive accuracy; however, there is significant variability among studies, partly due to the use of different linear and non-linear models that are difficult to compare side by side because of differences in populations and study designs. Current T2DM risk screening methods often rely on laboratory-based tests, such as fasting glucose and HbA1c, which, while effective, can be resource-intensive, impractical for large-scale screening and may miss early-stage or undiagnosed cases in asymptomatic individuals. We aim to compare the predictive performance of five machine learning algorithms applied to the same population. We used data from the National Health and Nutrition Examination Survey (NHANES) collected between 2007 and 2018 to train and test the following machine learning methods: logistic regression, XGBoost, CatBoost, support vector machine (SVM) and random forest. The importance of this research lies in its potential to bridge the gap between data-driven analytics and clinical application.

METHODS

Study population and variables

We used publicly available data from the NHANES 2007–2018, a nationally representative survey designed to assess the nutritional and health status of children and adults in the USA. The data set encompasses four primary categories: Demographics Data, Dietary Data, Examination Data and Questionnaire Data. Our analysis focused on non-pregnant adults.[13] As a self-reported and cross-sectional data set, its limitation for causal inference must be acknowledged. The NHANES data set used in this study provides a comprehensive collection of demographic, anthropometric, lifestyle and nutritional variables through structured household interviews, physical examinations and dietary recall assessments. Trained personnel conducted standardised measurements for anthropometric variables such as waist circumference and body mass index (BMI), while dietary and lifestyle factors, including alcohol intake, physical activity and macronutrient consumption, were obtained via validated survey instruments. This structured and standardised approach ensures data quality and consistency across survey cycles. After excluding entries with missing values, our final sample consisted of 29 509 respondents and 71 predictors, covering demographic, anthropometric, lifestyle and dietary factors. A detailed description of the predictors influencing the prevalence of diabetes mellitus is provided in table 1. The outcome variable was based on self-reported diagnoses of diabetes mellitus as recorded in the questionnaires. Diabetes status was based on the answer to the question, 'Doctor told you have diabetes?' We considered having diabetes when the answer was either 'yes' or 'borderline'.

Machine learning models

Five supervised machine learning algorithms were evaluated for predictive accuracy with respect to the risk of T2DM. In supervised learning methods, an observed outcome variable is used to train the model to predict future observations. We used logistic regression, SVM, random forest, XGBoost and CatBoost models. These models were selected based on their complementary strengths in classification tasks, balancing model interpretability, predictive accuracy and robustness in handling non-linear relationships within the NHANES data set. Hyperparameter tuning was conducted using grid search and 10-fold cross-validation to optimise model performance. For tree-based models (random forest, XGBoost, CatBoost), parameters such as learning rate, tree depth and regularisation terms were optimised, while for SVM, kernel type and cost parameter were selected based on performance metrics. The following is a brief description of the algorithms:

- Logistic regression is a statistical method that relates the logit transformation of a binary-dependent variable to one or more predictor variables. The coefficients assigned to predictors quantify the magnitude and direction of the relationship to the outcome after minimising the error between the model and the observed outcome.[14]
- SVM is a classifier that separates classes by a line or a plane serving as a boundary. Optimisation ensures that the distance of the boundary separation is maximised. SVM can learn non-linear decision surfaces and perform well in the presence of multiple predictors. Compared with simpler linear models such as logistic regression, it requires more computational complexity.[15]
- Random forest is an ensemble model based on the construction of multiple random decision trees using a bagging technique. Each individual tree predicts the outcome with the maximum possible purity, whereas the average prediction yielded by all the trees determines the global prediction, which reduces variance. In addition, the model ranks the predictors in order of importance.[16]
- XGBoost (extreme gradient boosting) is a powerful ensemble statistical learning method that combines the predictions of multiple individually trained classifiers. XGBoost uses decision trees that are built sequentially to correct errors, in contrast to random forest, which creates decision trees independently and combines their outputs. Some hyperparameters, such as the number of decision trees and their depth, are optimised through cross-validation before training the model.[17]
- CatBoost is a member of the family of gradient boosted decision trees. It is an ensemble technique whose strength lies in its versatility to be used on a

Table 1 List of variables and their meaning

RIAGENDR	Gender	RIDAGEYR	Age	RIDRETH1	Race	DMDEDUC2	Education level
DR1IKCAL	Energy (kcal)	DR1IPROT	Protein (g)	DR1ICARB	Carbohydrate (g)	DR1ISUGR	Total sugars (g)
DR1IFIBE	Dietary fibre (g)	DR1ITFAT	Total fat (g)	DR1ISFAT	Total saturated fatty acids (g)	DR1IMFAT	Total monounsaturated fatty acids (g)
DR1IPFAT	Total polyunsaturated fatty acids (g)	DR1ICHOL	Cholesterol (mg)	DR1IATOC	Vitamin E as α-tocopherol (mg)	DR1IATOA	Added α-tocopherol (vitamin E) (mg)
DR1IRET	Retinol (µg)	DR1IVARA	Vitamin A, RAE (µg)	DR1IACAR	α-Carotene (µg)	DR1IBCAR	β-Carotene (µg)
DR1ICRYP	β-Cryptoxanthin (µg)	DR1ILYCO	Lycopene (µg)	DR1ILZ	Lutein+zeaxanthin (µg)	DR1IVB1	Thiamin (vitamin B1) (mg)
DR1IVB2	Riboflavin (vitamin B2) (mg)	DR1INIAC	Niacin (mg)	DR1IVB6	Vitamin B6 (mg)	DR1IFOLA	Total folate (µg)
DR1IFA	Folic acid (µg)	DR1IFF	Food folate (µg)	DR1IFDFE	Folate, DFE (µg)	DR1ICHL	Total choline (mg)
DR1IVB12	Vitamin B12 (µg)	DR1IB12A	Added vitamin B12 (µg)	DR1IVC	Vitamin C (mg)	DR1IVD	Vitamin D (D2+D3) (µg)
DR1IVK	Vitamin K (µg)	DR1ICALC	Calcium (mg)	DR1IPHOS	Phosphorus (mg)	DR1IMAGN	Magnesium (mg)
DR1IIRON	Iron (mg)	DR1IZINC	Zinc (mg)	DR1ICOPP	Copper (mg)	DR1ISODI	Sodium (mg)
DR1IPOTA	Potassium (mg)	DR1ISELE	Selenium (µg)	DR1ICAFF	Caffeine (mg)	DR1ITHEO	Theobromine (mg)
DR1IALCO	Alcohol (g)	DR1IS040	SFA 4:0 (butanoic) (g)	DR1IS060	SFA 6:0 (hexanoic) (g)	DR1IS080	SFA 8:0 (octanoic) (g)
DR1IS100	SFA 10:0 (decanoic) (g)	DR1IS120	SFA 12:0 (dodecanoic) (g)	DR1IS140	SFA 14:0 (tetradecanoic) (g)	DR1IS160	SFA 16:0 (hexadecanoic) (g)
DR1IS180	SFA 18:0 (octadecanoic) (g)	DR1IM161	MFA 16:1 (hexadecenoic) (g)	DR1IM181	MFA 18:1 (octadecenoic) (g)	DR1IM201	MFA 20:1 (eicosenoic) (g)
DR1IM221	MFA 22:1 (docosenoic) (g)	DR1IP182	PFA 18:2 (octadecadienoic) (g)	DR1IP183	PFA 18:3 (octadecatrienoic) (g)	DR1IP184	PFA 18:4 (octadecatetraenoic) (g)
DR1IP204	PFA 20:4 (eicosatetraenoic) (g)	DR1IP205	PFA 20:5 (eicosapentaenoic) (g)	DR1IP225	PFA 22:5 (docosapentaenoic) (g)	DR1IP226	PFA 22:6 (docosahexaenoic) (g)
BMXBMI	Body mass index (kg/m^2)	BMXWAIST	Waist circumference (cm)	PAD680	Minutes of sedentary activity		

MFA, monounsaturated fatty acid; PFA, polyunsaturated fatty acid; SFA, saturated fatty acid.

variety of data sets. The two main innovations introduced by this algorithm are the implementation of ordered boosting and the ability to process categorical features.[18]

Sample weights and validation

Weights provided by NHANES were used for demographic variables to account for the survey design and sample representativeness of the population. Specifically, NHANES provides weights to prevent oversampling of certain demographic groups and adjust for non-response. The weights were incorporated into all the models to ensure generalisability of the results.

For validation and training of the models, the data set was first split into training and test sets to evaluate predictive performance. In general, 80% of the observations were used for training purposes. We employed 10-fold cross-validation to optimise the models and prevent overfitting, ensuring robust performance across different data subsets. The test set was used for final evaluation, providing an unbiased measure of accuracy, sensitivity and specificity.

Feature selection

The predictors identified as statistically significant in the logistic regression model with forward selection were included in all the models. The Akaike information criterion value helped us determine the set of final features. Reducing the number of predictors in the final models minimised the computational resources needed for analysis to make them more interpretable without sacrificing predictive power.

Performance metrics

The performance of the models was compared using several metrics: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and area under the receiving operating curve (ROC). Accuracy measures the proportion of correct predictions across all classes, providing a global sense of model performance. Sensitivity assesses the ability of the model to correctly identify positive cases of T2DM, whereas specificity measures the model's ability to identify negative cases. PPV is a measure of precision, and in conjunction with NPV provides the proportion of correct predictions (positive and negative). Finally, the area under the curve (AUC) evaluates the model's discrimination ability across different thresholds. To balance model sensitivity and specificity, we evaluated multiple decision thresholds using Youden's index and precision-recall curve analysis. The optimal threshold was selected based on maximising sensitivity while maintaining clinically acceptable specificity for diabetes screening. The 10-fold cross-validation errors indicate that logistic regression (0.1095) had the lowest error, while XGBoost (0.1903) exhibited the highest. The results suggest that while tree-based models can capture complex interactions, they may also introduce additional variance, leading to slightly higher error rates compared with traditional statistical methods.

Statistical analysis

Data are presented as mean±SD for continuous variables and frequency (%) for categorical variables. We assessed the difference between diabetes and non-diabetes for each categorical variable with the χ^2 test. Continuous variables were evaluated for normality with the Kolmogorov-Smirnov test. For normally distributed variables, we used Student's t-test, whereas for non-normally distributed data, we employed the Mann-Whitney U test for comparison between diabetes and non-diabetes. Statistical significance was set at $p<0.05$. Statistical analyses were performed using the R package (V.4.2.2, https://www.r-project.org/).

Patient and public involvement

Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of the study.

RESULTS

Characteristics of the study population

A total of 29 509 respondents from the NHANES survey between 2007 and 2018 were included in this study. The mean age of the participants was 49.32 years, and 14 408 participants were male (48.8%). A detailed description of participant characteristics is shown in table 2. Except for other races, the proportion of ethnic groups was higher among non-diabetic participants. On the other hand, a low educational level was more prevalent among non-diabetics, whereas the proportion of highly educated participants was higher in the group of diabetics.

Correlation of variables with clinical outcome

A total of 23 608 participants were assigned to the training cohort (80% of the total population). In the training cohort, univariate analyses followed by a logistic regression with forward variable selection were used to identify 21 significant predictors: age, race, waist circumference, BMI, daily intake of sugar, protein, alcohol, carbohydrate, fibre, calcium, phosphorus, potassium, vitamin A, folate, vitamin B1, vitamin B2, niacin, choline and fats (saturated, monounsaturated and polyunsaturated). This list of predictors was used for all the machine learning models.

Development and validation of predictive models

Based on the predictors identified through forward logistic regression, five machine learning algorithms were used: logistic regression, SVM, random forest, XGBoost and CatBoost. Table 3 shows the ORs and CIs for the forward logistic regression model. First, we employed 10-fold cross-validation to determine the error rate for the models, and based on the predictions, we constructed contingency tables to calculate the predictive performance indicators (table 4). For internal validation, 10-fold cross-validation was used to calculate errors for logistic

Table 2 Baseline characteristics of participants between diabetics and non-diabetics

Variable	Diabetics (mean±SD or %)	Non-diabetics (mean±SD or %)	P value
Age (years)	61 (13.1)	47 (17.5)	<0.0001
Gender	Male: 2481 (45%)	Male: 12 499 (48%)	<0.001
Race			
Hispanic	3805 (13%)	819 (17%)	<0.001
Non-Hispanic white	11 096 (39%)	1717 (36%)	0.75
Non-Hispanic black	2695 (9%)	510 (10%)	<0.001
Asian	5349 (19%)	1283 (27%)	<0.001
Other/mixed	3015 (20%)	498 (10%)	0.04
Education			
Less than high school	5902 (23%)	1601 (33%)	<0.001
High school graduate	5967 (23%)	1092 (22%)	0.11
Some college/higher	14 091 (54%)	2134 (45%)	<0.001
Waist circumference (cm)	109.17 (16.07)	97.84 (15.87)	<0.0001
Sugar intake (g)	89.12 (64.90)	115.60 (80.05)	<0.0001
Alcohol (g)	5.17 (20.68)	10.93 (29.66)	<0.001
Dodecanoic acid (g)	0.74 (1.23)	0.82 (1.49)	0.003
Choline (mg)	322.01 (192.87)	335.63 (203.53)	0.001
Fibre (g)	16.44 (10.15)	16.82 (10.68)	0.02
Calcium (mg)	848.31 (518.57)	933.14 (593.89)	<0.0001
Copper (mg)	1.19 (1.13)	1.25 (0.94)	0.003

regression, SVM, random forest, XGBoost and CatBoost, with values 0.1095, 0.1505, 0.1511, 0.1903 and 0.1473, respectively. Among the models, the XGBoost algorithm exhibited the best prediction performance with an AUC of 0.8168. Additionally, the models were evaluated by calculating sensitivity, specificity and predictive values (table 4). Figure 1 shows the ROC for the five models. XGBoost was identified as the model with the highest AUC, coupled with high specificity. However, random forest and logistic regression exhibited a slightly lower AUC and similar sensitivity and accuracy compared with XGBoost, with an advantage in terms of interpretability. All models showed low sensitivity ranging between 7.15% and 15.34%. The SVM algorithm had slightly better sensitivity, although with the lowest accuracy. The CatBoost algorithm exhibited the lowest AUC.

Table 3 Forward logistic regression: association of covariates and diabetes mellitus

Variable	OR	95% CI lower	95% CI upper
Age	1.045	1.042	1.047
Race	1.040	1.008	1.074
Education level	0.980	0.964	0.995
Waist circumference	1.049	1.044	1.054
Total sugar	0.993	0.992	0.995
Alcohol	0.992	0.990	0.994
Sodium	0.991	0.981	1.020
Fibre	1.014	1.080	1.020
Eicosatetraenoic acid	1.992	1.475	2.691
Butanoic acid	0.769	0.693	0.854
Dodecanoic acid	1.029	1.003	1.056
Body mass index	0.983	0.971	0.994
Zinc	0.994	0.987	1.002

Model interpretation

This study compared five machine learning algorithms (figure 2). Although the XGBoost slightly outperformed logistic regression and random forest, the latter two exhibited a high specificity and accuracy. Considering that random forest stands as an easy-to-interpret model, we identified the ranking of feature importance. Demographic and anthropometric variables such as age, waist circumference and BMI were the most important features associated with the risk of T2DM, followed by a group of dietary factors, including daily intake of sugar, carbohydrates, protein, fibre, calcium, phosphorus, α-tocopherol and various vitamins (figure 3). To enhance model interpretability, we conducted a Shapley Additive Explanations (SHAP) analysis using the 'shapviz' package in R. SHAP values were computed for a subset of 500 observations to quantify the contribution of each predictor variable to the model's classification of T2DM. The SHAP summary and importance plots were used to identify the key features influencing the model's decision-making

Table 4 Model performance measures for five machine algorithms

Model	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)	AUC
LR	11.2	98.14	52	86	84.96	0.798
SVM	58.57	83.7	95.1	26.9	62.44	0.781
RF	7.15	98.86	52.89	85.63	84.95	0.793
XB	11.62	98.08	52	86.12	84.97	0.8168
CB	15.34	97.09	47.67	86.94	85.08	0.5622

AUC, area under the curve; CB, Cat Boosting; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine; XGB, XG Boosting.

process, providing insights into variable impact beyond traditional statistical coefficients (figure 4). The analysis revealed that age (RIDAGEYR) and waist circumference (BMXWAIST) were the most influential predictors in the XGBoost model, with higher SHAP values indicating a strong positive contribution to diabetes risk classification. Other significant variables included dietary sugar intake (DR1ISUGR) and alcohol consumption (DR1IALCO), reinforcing the role of lifestyle factors in diabetes prediction. Interestingly, some variables with high importance in traditional regression models, such as BMI (BMXBMI), had comparatively lower SHAP values, suggesting that waist circumference may be a stronger predictor within the machine learning framework.

DISCUSSION

In this study, we developed and tested five machine learning algorithms to predict T2DM based on demographic, anthropometric, lifestyle and nutritional variables. We used the NHANES database from 2007 to 2018 to train the models and evaluated their accuracy and predictive performance on a test subset of data. The highest AUC for ROC corresponded to the XGBoost model, while the CatBoost algorithm exhibited the lowest AUC. Specificity was greater than 97% for all models except for SVM (83.7%). The lowest sensitivity was observed in the random forest model (7.15%), and the highest level occurred with SVM (58.57%). Regarding accuracy, the SVM had the lowest value at 62.44%. In contrast, the other four models had comparable accuracy, around 85%. The random forest model, although not the most accurate, exhibited a performance level similar to XGBoost and helped identify the most important features (age, waist circumference, BMI and specific dietary variables).

T2DM is a chronic metabolic disease with multiorgan deleterious consequences. Lifestyle risk factor modification plays a significant role in the prevention and management of this metabolic disorder. On the other hand, public health efforts addressing early identification of populations at risk of developing T2DM have been

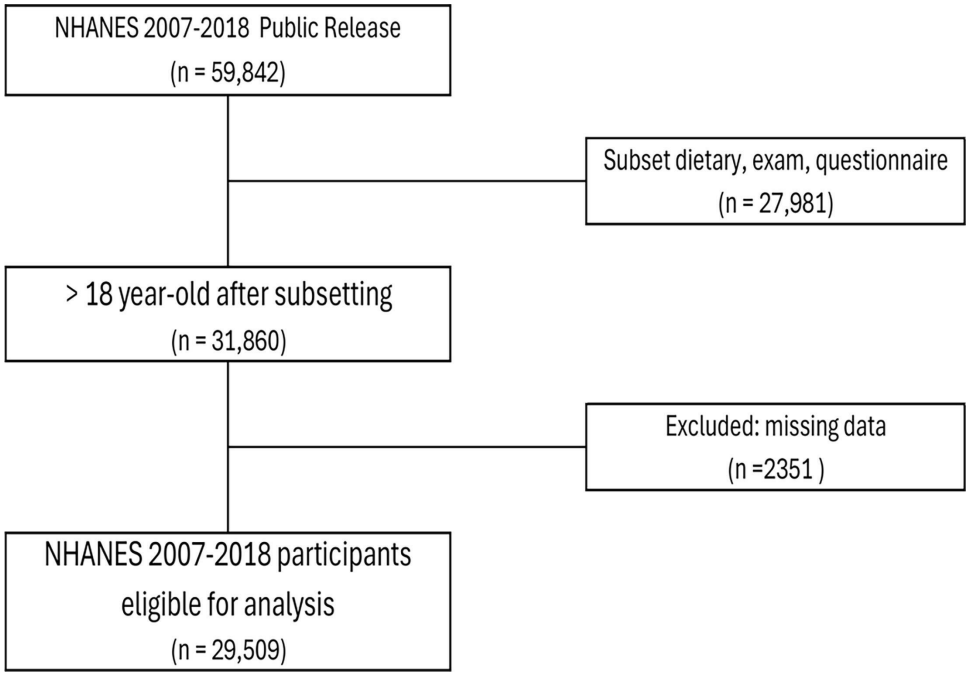

Figure 1 Flow diagram of the participants included in the analysis. NHANES, National Health and Nutrition Examination Survey.

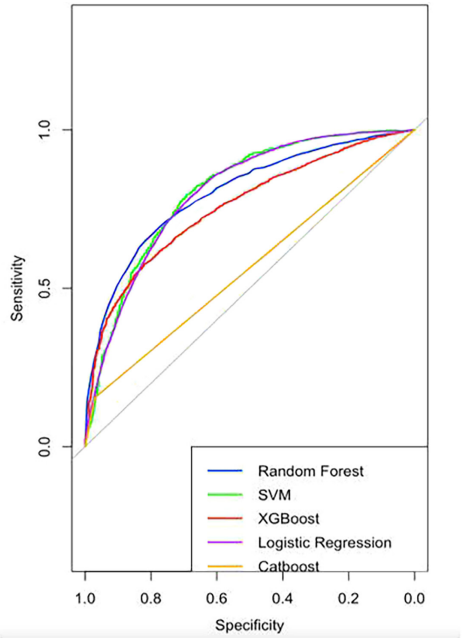

Figure 2 Combined receiver operating characteristic curves for all models. SVM, support vector machine; XGB, XG Boosting.

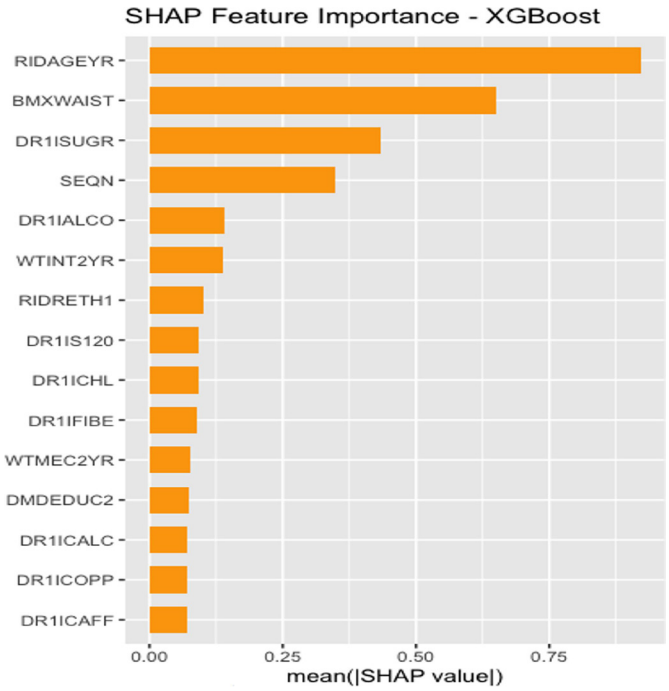

Figure 4 Shapley Additive Explanations analysis. SHAP, Shapley Additive Explanations; XGB, XG Boosting.

reported in the last few decades. Given the multifactorial nature of the disease in addition to the contribution of modifiable risk factors to its development and progression, incorporation of lifestyle factors in early prediction may prove beneficial.

Machine learning, as a subdomain of artificial intelligence, uses algorithms to explore and learn from data to identify patterns and associations and use them for predictive purposes. Several authors have explored the clinical application of machine learning models. Wilson *et al* reported that a logistic regression model based on simple clinical predictors (family history, obesity and metabolic

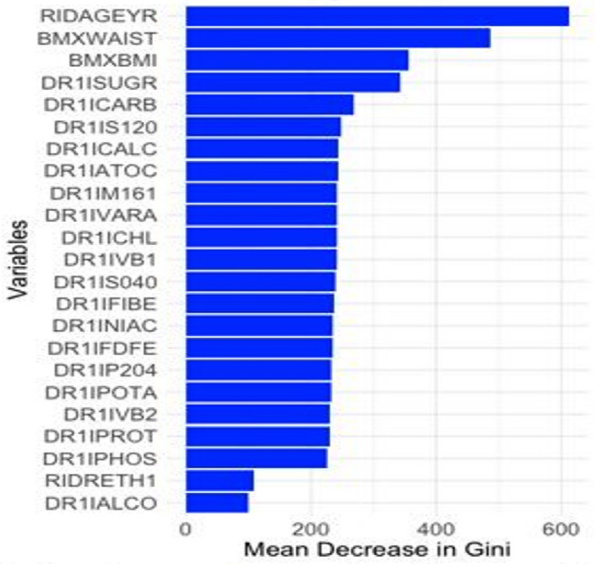

Figure 3 Predictor variables order of importance. SHAP, Shapley Additive Explanations.

syndrome traits) had an acceptable predictive accuracy compared with models incorporating more sophisticated laboratory measurements, with an AUC for ROC of 0.85. Considering that this study was limited to a 99% white population, Mashayehi *et al* applied Wilson's simple clinical model to a Canadian population, reporting an AUC of 0.83, indicating that adding ethnic diversity to the sample decreases the predictive accuracy.

Our study found that one of the models (XGBoost) has an AUC of 0.8168, which can still be considered excellent for the capability to discriminate diabetics from non-diabetics. In addition, the XGBoost, CatBoost, logistic regression and random forest exhibited specificities greater than 97% and accuracy around 85%. Our model employed a sample that is representative of the population of the USA, thanks to the sample weights assigned for analysis. Among the differences between our study and the two mentioned above is the use of dietary variables. We may argue that since T2DM is strongly associated with nutritional lifestyle factors, developing a model with such factors may help fine-tune the predictive accuracy of a model.

XGBoost exhibited the best performance, and SVM showed the lowest predictive accuracy. Our results agree with a study by Kushwaha *et al* that revealed the highest AUC for XGBoost and the lowest for SVM in children and adolescents. The predictors in that study included glycosylated haemoglobin.[19] Other studies have evaluated machine learning algorithms with varying results. Khanam *et al* used the Pima Indians Diabetes database, consisting of 768 observations. The list of predictors included plasma glucose tolerance test, blood pressure, triceps skinfold thickness and insulin levels in addition

to demographic variables. They found an accuracy of 77–78% for SVM and logistic regression, judged by them as good, and a complex neural network model with accuracy of 88%.[20] Our results show accuracy of 85% for all the models (except for SVM). The accuracy of our models was positively impacted by the sample size of more than 29 000 subjects. Adjusting the classification threshold from the default 0.5 to the optimal value identified by Youden's index improved sensitivity while maintaining a reasonable trade-off with specificity. This approach enhances the real-world applicability of the model by making it more effective for early diabetes risk identification, particularly in population-wide screening programmes.

There is a direct influence of lifestyle and nutritional factors on the development and progression of T2DM.[21] The machine learning algorithms for T2DM focusing on nutritional factors are scarce. De Silva *et al* studied a cohort of 6673 NHANES participants between 2013 and 2016 to develop 12 models, with the best AUC (74.9–75.7%). The authors evaluated 114 potential nutritional factors, 13 behavioural and 12 socio-economic variables. The best models classified 39 markers of T2DM.[22] Diagnosis of T2DM was established by a positive answer to the question, 'Have you ever been told by a doctor that you have diabetes?' in both their study and our analysis. We employed this criterion as an indicator of T2DM because self-reported physician diagnoses capture a clinical judgement that integrates patient history and laboratory results. Self-reported data can reflect a diagnosis that is clinically significant, even in the presence of fluctuating laboratory values. Compared with the AUC reported by De Silva, the AUC was higher for all the models except for CatBoost. The sensitivity of our models was significantly inferior to their models; however, the AUC for ROC, specificity and accuracy were higher. This indicates a strong discriminatory ability for T2DM.

Despite the low sensitivity, the high AUC illustrates the trade-offs between false negatives and false positives. For instance, our XGBoost model is excellent at identifying patients at high risk of developing T2DM at the cost of missing some patients at risk. In this context, both De Silva's and our results identified that the variables with the highest importance are waist circumference, BMI and age. We may argue that the recommendation would be to start lifestyle/nutritional intervention in those identified at risk by the model and follow-up closely those with alterations in those three variables. Furthermore, those with abnormalities in the three top variables could also receive the intervention, considering the favourable risk-to-benefit ratio associated with lifestyle changes. Although the models evaluated in our study showed limited sensitivity, the easy-to-obtain information about the predictors and the high specificity, accuracy and overall performance is sufficiently high to be of clinical use. None of the predictors requires blood sampling or imaging techniques and can be followed up with simple surveys. Adjusting classification thresholds may offer a practical approach to improving sensitivity while maintaining an acceptable trade-off with specificity. Given that AUC remains constant across different thresholds, selecting a lower decision threshold could enhance the detection of true-positive cases, making the model more applicable for screening purposes while allowing clinicians to refine risk stratification based on specific population needs.

Some specific anthropometric and nutritional variables stood out as significant to predict T2DM. Of note, we showed a positive association for waist circumference and a negative association for BMI. A systematic review reported that each 10 cm increase in waist circumference represented a relative risk of 1.61 (1.52–1.70), whereas a five-unit increase in BMI had a relative risk of 1.61 (1.52–1.70).[23] In this study, the association was positive for both variables. Siren *et al* recognised that a waist circumference >94 cm in middle-aged men predicted T2DM with 84% sensitivity and 78% specificity.[24] Han *et al* identified an 'obesity paradox' in T2DM, where high BMI was associated with lower mortality. They postulate measurement error and the presence of confounding factors as the cause for this phenomenon.[25] On the other hand, Feller *et al* evaluated a cohort of the European Prospective Investigation into Cancer and Nutrition study. Although they found a positive association for both waist circumference and BMI, waist circumference was a strong predictor in those with low or normal weight and BMI.[26] Our results, although paradoxical at first sight, highlight the importance of waist circumference as a predictor of T2DM that may be less sensitive to changes in dietary interventions. This means that since our study is cross-sectional, some patients already diagnosed with T2DM could be on dietary regimes that lower their BMI; however, the waist circumference was not affected in the same way, making this predictor more robust for prediction and follow-up. A similar situation can be observed for added sugars (negative association) and carbohydrate intake (positive association). It is well recognised that high carbohydrate intake is a risk factor for T2DM; however, patients with T2DM tend to reduce the consumption of added sugars but not necessarily the total intake of carbohydrates.[27] The SHAP analysis provides valuable insights into feature importance, demonstrating that age and waist circumference play a dominant role in diabetes risk prediction. These findings align with established clinical knowledge but further highlight the advantage of machine learning in detecting non-linear interactions among predictors. Notably, the identification of dietary sugar intake and alcohol consumption as key contributors suggests that integrating modifiable lifestyle factors into diabetes screening tools could improve early risk assessment. The discrepancy between BMI and waist circumference underscores the need for reassessing standard anthropometric measures in predictive models. These results emphasise the potential of SHAP analysis in enhancing model transparency, supporting its use in clinical and public health applications for individualised risk stratification. Finally, the SHAP analysis suggests that machine learning can uncover complex, non-linear relationships in T2DM risk

factors that may not be captured in traditional statistical approaches, enhancing the potential for more precise, personalised chronic disease screening strategies.

Our results show that calcium and phosphorus intake are positively associated with T2DM. Other studies have found that serum concentration of these elements are associated with the disease independently of glycaemia and insulin dynamics.[28] Although the mechanism for the association is not well known, high calcium-phosphate product levels may underlie peripheral tissue insulin resistance.[29] It is recognised that the consumption of saturated fats increases the risk of T2DM, which was corroborated by our study.[30] However, recent studies have found contradictory results.[31]

The role of polyunsaturated fatty acids (PUFA) is even less clear. We found a positive association between PUFA and T2DM. Hu *et al*, in a dose-response meta-analysis of cohort studies, showed in subgroup analyses a positive association in Europe and Australia and a negative association in Asia.[32] Our study includes the population in the USA, whose diverse genetic background may play a role in the positive associations found.

The relationship between protein intake and T2DM risk has been a subject of significant investigation, with varying conclusions. High protein diets, especially those rich in animal protein, have been associated with an increased risk of T2DM in multiple studies. For example, research shows that a higher intake of red and processed meats can elevate the risk, likely due to the insulin resistance induced by certain amino acids present in animal proteins.[33] Evidence also supports the role of total protein intake in weight management, which is a crucial factor in T2DM prevention. High-protein diets can aid in weight loss and improve metabolic markers like fasting glucose and glycated haemoglobin (HbA1c) levels. This is especially relevant since weight loss is a key strategy in managing and preventing T2DM. However, the type of protein consumed plays a critical role, with plant proteins showing more favourable outcomes compared with animal proteins.[34]

Our findings align with existing literature, such as the study by Ma *et al*, which demonstrated that moderate alcohol consumption, particularly with meals, is associated with a lower risk of T2DM.[35] Similarly, a nationwide cohort study highlighted the increased risk of T2DM associated with repeated binge drinking.[36] On the other hand, a meta-analysis supports a dose-response relationship between alcohol and T2DM, suggesting a nuanced interaction between consumption patterns and T2DM risk.[37] Overall, these findings suggest that while alcohol consumption in moderation might reduce T2DM risk, other dietary factors and consumption patterns, such as binge drinking, play a critical role in modulating this risk. Beyond alcohol consumption, the growing body of research comparing machine learning and traditional statistical methods suggests that advanced predictive models may better capture the intricate web of lifestyle and metabolic interactions contributing to diabetes onset.[38] In particular, community-based follow-up data have proven to enhance risk prediction, emphasising the importance of long-term dietary behaviours in metabolic health.[39] Additionally, emerging evidence suggests that T2DM may not only be influenced by dietary factors but also by infectious disease interactions, with data mining approaches revealing significant associations between diabetes and SARS-CoV-2 outcomes.[40] These findings reinforce the multifactorial nature of T2DM risk, where dietary patterns, disease susceptibility and predictive modelling advancements converge to shape our understanding and future prevention efforts.

Our study has limitations.[1] Recall bias is a concern given that our data were extracted from NHANES data sets, and the information was self-reported in its entirety.[2 41] The diagnosis of T2DM was based solely on the answer to a single question, without laboratory results for confirmation.[3] The use of NHANES data is limited to the population of the USA, restricting the generalisability of our results.[4] This is a cross-sectional study that did not take into consideration the progression of the disease.[5] Causality cannot be assessed due to the cross-sectional nature of the study; moreover, reverse causality is possible for some predictors.

CONCLUSION

Our study demonstrates the potential of machine learning models in predicting T2DM using non-invasive demographic, anthropometric, lifestyle and dietary factors. Among the five models tested, logistic regression exhibited the lowest cross-validation error, XGBoost demonstrated the highest predictive performance, consistent with prior studies that use this algorithm for complex data sets, while SVM showed the highest sensitivity, suggesting that model selection should be tailored to specific screening objectives. SHAP analysis identified age, waist circumference and dietary sugar intake as the most influential predictors, reinforcing the importance of targeted interventions focusing on central obesity and dietary modifications.

From a clinical perspective, integrating machine learning-based risk prediction tools into routine screening could enable earlier identification of high-risk individuals, allowing for timely lifestyle interventions and personalised treatment strategies. These findings suggest that future research should explore real-world validation of these models in diverse populations and assess their effectiveness in guiding clinical decision-making. Additionally, refining machine learning approaches through threshold optimisation and hybrid modelling techniques could further enhance predictive accuracy and clinical applicability.

Ultimately, this study highlights the value of artificial intelligence in chronic disease screening and prevention, offering a data-driven framework to support precision medicine initiatives in diabetes care.

Contributors ERP is the guarantor and contributed to concept idea, data analysis and manuscript construction. BA-M contributed to article construction, edits and revision.

Funding This study was funded by the authors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The National Center for Health Statistics Research Ethics Review Board approved NHANES. Written informed consent was obtained from all adult participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The data used in this research are publicly available (NHANES).

ORCID iD

Efrain Riveros Perez http://orcid.org/0000-0002-3874-5783

REFERENCES

1 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.
2 Cannon A, Handelsman Y, Heile M, *et al*. Burden of Illness in Type 2 Diabetes Mellitus. *J Manag Care Spec Pharm* 2018;24:S5–13.
3 Polonsky WH. Emotional and quality-of-life aspects of diabetes management. *Curr Diab Rep* 2002;2:153–9.
4 American Diabetes Association. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care* 2013;36:1033–46.
5 Gray N, Picone G, Sloan F, *et al*. Relation between BMI and diabetes mellitus and its complications among US older adults. *South Med J* 2015;108:29–36.
6 Hamilton MT, Hamilton DG, Zderic TW. Sedentary behavior as a mediator of type 2 diabetes. *Med Sport Sci* 2014;60:11–26.
7 Park SE, Seo MH, Cho J-H, *et al*. Dose-Dependent Effect of Smoking on Risk of Diabetes Remains after Smoking Cessation: A Nationwide Population-Based Cohort Study in Korea. *Diabetes Metab J* 2021;45:539–46.
8 Kalandarova M, Ahmad I, Aung TNN, *et al*. Association Between Dietary Habits and Type 2 Diabetes Mellitus in Thai Adults: A Case-Control Study. *Diabetes Metab Syndr Obes* 2024;17:1143–55.
9 Kim SJ, Kim DJ. Alcoholism and diabetes mellitus. *Diabetes Metab J* 2012;36:108–15.
10 Qin Y, Wu J, Xiao W, *et al*. Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. *Int J Environ Res Public Health* 2022;19:15027.
11 Lugner M, Rawshani A, Helleryd E, *et al*. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Sci Rep* 2024;14:2102.
12 Lv K, Cui C, Fan R, *et al*. Detection of diabetic patients in people with normal fasting glucose using machine learning. *BMC Med* 2023;21:342.
13 Zipf G, Chiappa M, Porter KS, *et al*. National health and nutrition examination survey: plan and operations, 1999-2010. 1999.
14 Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb)* 2014;24:12–8.
15 Rodríguez-Pérez R, Bajorath J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J Comput Aided Mol Des* 2022;36:355–62.
16 Rigatti SJ. Random Forest. *J Insur Med* 2017;47:31–9.
17 Moore A, Bell M. XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. *Clin Med Insights Cardiol* 2022;16:11795468221133611.
18 Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data* 2020;7:94.
19 Kushwaha S, Srivastava R, Jain R, *et al*. Harnessing machine learning models for non-invasive pre-diabetes screening in children and adolescents. *Comput Methods Programs Biomed* 2022;226:::.
20 Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *ICT Express* 2021;7:432–9.
21 Guo Y, Huang Z, Sang D, *et al*. The Role of Nutrition in the Prevention and Intervention of Type 2 Diabetes. *Front Bioeng Biotechnol* 2020;8:575442.
22 De Silva K, Lim S, Mousa A, *et al*. Nutritional markers of undiagnosed type 2 diabetes in adults: Findings of a machine learning analysis with external validation and benchmarking. *PLoS ONE* 2021;16:e0250832.
23 Jayedi A, Soltani S, Motlagh SZ-T, *et al*. Anthropometric and adiposity indicators and risk of type 2 diabetes: systematic review and dose-response meta-analysis of cohort studies. *BMJ* 2022;376:e067516.
24 Siren R, Eriksson JG, Vanhanen H. Waist circumference a good indicator of future risk for type 2 diabetes and cardiovascular disease. *BMC Public Health* 2012;12:631.
25 Han SJ, Boyko EJ. The Evidence for an Obesity Paradox in Type 2 Diabetes Mellitus. *Diabetes Metab J* 2018;42:179–87.
26 Feller S, Boeing H, Pischon T. Body mass index, waist circumference, and the risk of type 2 diabetes mellitus: implications for routine clinical practice. *Dtsch Arztebl Int* 2010;107:470–6.
27 Veit M, van Asten R, Olie A, *et al*. The role of dietary sugars, overweight, and obesity in type 2 diabetes mellitus: a narrative review. *Eur J Clin Nutr* 2022;76:1497–501.
28 Lorenzo C, Hanley AJ, Rewers MJ, *et al*. Calcium and phosphate concentrations and future development of type 2 diabetes: the Insulin Resistance Atherosclerosis Study. *Diabetologia* 2014;57:1366–74.
29 Akter S, Eguchi M, Kochi T, *et al*. Association of Serum Calcium and Phosphate Concentrations with Glucose Metabolism Markers: The Furukawa Nutrition and Health Study. *Nutrients* 2020;12:2344.
30 Risérus U, Willett WC, Hu FB. Dietary fats and prevention of type 2 diabetes. *Prog Lipid Res* 2009;48:44–51.
31 Gaeini Z, Bahadoran Z, Mirmiran P. Saturated Fatty Acid Intake and Risk of Type 2 Diabetes: An Updated Systematic Review and Dose-Response Meta-Analysis of Cohort Studies. *Adv Nutr* 2022;13:2125–35.
32 Hu M, Fang Z, Zhang T, *et al*. Polyunsaturated fatty acid intake and incidence of type 2 diabetes in adults: a dose response meta-analysis of cohort studies. *Diabetol Metab Syndr* 2022;14:34.
33 Pfeiffer AFH, Pedersen E, Schwab U, *et al*. The Effects of Different Quantities and Qualities of Protein Intake in People with Diabetes Mellitus. *Nutrients* 2020;12:365.
34 Sluik D, Brouwer-Brolsma EM, Berendsen AAM, *et al*. Protein intake and the incidence of pre-diabetes and diabetes in 4 population-based studies: the PREVIEW project. *Am J Clin Nutr* 2019;109:1310–8.
35 Ma H, Wang X, Li X, *et al*. Moderate alcohol drinking with meals is related to lower incidence of type 2 diabetes. *Am J Clin Nutr* 2022;116:1507–14.
36 Choi JW, Han E, Kim TH. Risk of Hypertension and Type 2 Diabetes in Relation to Changes in Alcohol Consumption: A Nationwide Cohort Study. *Int J Environ Res Public Health* 2022;19:4941.
37 Baliunas DO, Taylor BJ, Irving H, *et al*. Alcohol as a risk factor for type 2 diabetes: A systematic review and meta-analysis. *Diabetes Care* 2009;32:2123–32.
38 Choi SG, Oh M, Park DH, *et al*. Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. *Sci Rep* 2023;13:13101.
39 Jiang L, Xia Z, Zhu R, *et al*. Diabetes risk prediction model based on community follow-up data using machine learning. *Prev Med Rep* 2023;35:102358.
40 Ghazizadeh H, Shakour N, Ghoflchi S, *et al*. Use of data mining approaches to explore the association between type 2 diabetes mellitus with SARS-CoV-2. *BMC Pulm Med* 2023;23:203.
41 Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). Department of health and human services, centers for disease control and prevention, 2007-2018. Hyattsville, MD: U.S National Health and Nutrition Examination Survey Data; 2018. Available: https://wwwn.cdc.gov/nchs/nhanes/