



OPEN

Application of generalized linear mixed effects random forest for identifying risk factors of prediabetes in Tehran Lipid and Glucose Study

Maryam Karimi Ghahfarokhi¹, Farid Zayeri¹, Davood Khalili³, Chiara Masci⁴, Ali Sheidaei⁵, Yaser Mehrabi⁶, Fatemeh Masaebi⁷ & Yadollah Mehrabi^{2,3}✉

Prediabetes is a major risk factor for the development of diabetes, defined by blood glucose levels that are elevated but not yet high enough to meet the diagnostic criteria for Diabetes Mellitus. This condition is often clinically “silent” yet it can already lead to negative effects on various organ systems and frequently indicates the impending onset of type 2 diabetes mellitus. This study aimed to compare a traditional statistical model, the Generalized Linear Mixed Model (GLMM), with two tree-based machine learning models, Random Forest (RF) and Generalized Mixed-Effects Random Forest (GMERF), for predicting prediabetes and identifying key risk indicators in longitudinal data. The study sample included 5361 individuals aged over 20 years, focusing on 32 different variables. The target variable was the presence of prediabetes in a longitudinal setting. We applied three models: RF, which is tree-based but does not account for repeated measurements; GLMM, which handles random effects but assumes linear relationships; and GMERF, a hybrid model that incorporates both random effects and the nonlinearity of decision trees. Model performance was evaluated using standard predictive metrics. Among the three models, GMERF achieved the highest predictive performance. The area under the ROC curve was 0.63 for RF, 0.70 for GLMM, and 0.74 for GMERF. In the GMERF model, the top five predictive variables were Waist-to-Hip Ratio (WHR), age, waist circumference, triglyceride level, and Waist-to-Height Ratio (WHtR). WHR was ranked as the most important feature in both the GMERF and RF models. All of these variables, except WHtR, were also found to be significant in the GLMM model. In longitudinal data, there is an inherent dependence between observations collected over time. By incorporating these considerations, models that account for this data structure are better equipped to handle the complexities of longitudinal data, leading to more reliable and accurate predictions.

Keywords Prediabetes, Machine learning, GMERF, GLMM, Longitudinal data

Pre-diabetes is a condition where an individual's blood glucose levels are elevated above normal but not high enough to be diagnosed as diabetes¹. This state indicates an increased risk of developing type 2 diabetes and other diseases, such as cardiovascular diseases, if not addressed². Prediabetes has become a worldwide epidemic³, characterized by various associated complications, and its prevalence is on the rise, notwithstanding its treatable nature^{4,5}. The Expert Committee on Diagnosis and Classification of DM in 1997 initially recognized prediabetes as an intermediate stage, signifying a relatively elevated risk for future diabetes development⁶. It has been

¹Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ²Department of Epidemiology, School of Public Health & Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ³Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ⁴Department of Economics, Management and Quantitative Methods, Università Degli Studi di Milano, Milan, Italy. ⁵Department of Environmental Medicine, Icahn School of Medicine at Mount Sinai, New York, USA. ⁶Statistical Center of Iran, Tehran, Iran. ⁷Department of Biostatistics and Epidemiology, School of Medicine, Zanjan University of Medical Sciences, Zanjan, Iran. ✉email: ymehrabi@gmail.com

documented that around 5–10% of individuals with untreated prediabetes subsequently progress to develop diabetes. With prolonged observation, there is a potential for an increase in this progression rate⁷. This statistic is especially concerning, given the high prevalence of prediabetes among Iranian adults (18.2%)⁸. In Iran, it is predicted that the prevalence of diabetes will increase by 91% by 2035, and the associated economic burden is expected to increase significantly by 2035^{9,10}. Therefore, identifying people with prediabetes provides an opportunity for intervention through lifestyle modification and pharmacological treatments aimed at preventing the progression of diabetes^{11,12}. The primary aim of the present study is to investigate the determinants of prediabetes within an urban population in Tehran, Iran.

Existing prediabetes screening tools, which rely on a limited selection of established risk factors, reportedly failed to identify a significant proportion of undetected individuals with prediabetes¹³. In addition to clinical tools for the diagnosis of pre-diabetes, the scientific community has increasingly focused on the use of powerful statistical models for early and accurate prediction of pre-diabetes. These models can contribute to identifying predisposing factors for pre-diabetes in various communities, particularly in Iranian society, and aid in delaying or preventing diabetes. As a result, they can help reduce mortality and associated complications in the studied population.

This study analyzes longitudinal data from the Tehran Lipid and Glucose Study (TLGS) using advanced statistical models. Traditional methods, such as generalized linear mixed models (GLMM), are commonly applied to such data but have certain limitations^{14–16}. A key limitation of GLMMs is the need for users to pre-specify interactions between predictors and nonlinear relationships with the outcome, features that are often unknown beforehand and difficult to identify in high-dimensional datasets¹⁷. To address these challenges, machine learning (ML) models have been increasingly utilized. Although ML techniques are more complex than conventional statistical methods, they do not rely on assumptions such as linear relationships or normally distributed residuals¹⁸. This flexibility enables data analysts to explore large datasets with many variables more thoroughly and efficiently¹⁹.

Currently, ML methods are being employed for the detection and prediction of pre-diabetes, complications, and disease progression^{20–24}. One of the advanced ML methods used for developing predictive models is the Random Forest (RF) algorithm. The standard RF algorithm, like other ML methods, assumes that observations are independently sampled from the population. Ignoring the dependencies among observations collected over time in longitudinal data can result in biased inferences, such as underestimated standard errors in linear models²⁵. To date, research studies have not adequately emphasized the examination and understanding of the correlations within the data learning methods that have recently been proposed. One advanced technique is generalized mixed effects random forests (GMERF)²⁶. The main concept of integrating random forests into mixed linear models is to replace the fixed effect component with a tree or forest while using random effects to model the dependence structure²⁶. This comprehensive approach enables scientists, especially in areas like healthcare, to investigate intricate questions about treatment efficacy while considering the diverse responses of individual subjects²⁷.

The objective of this research is to utilize a more advanced statistical model to determine the most efficient ML approaches for examining interrelated data. Additionally, it seeks to rigorously evaluate the predictive capabilities of the selected model in accurately classifying new patients at risk of prediabetes. For the first time, we incorporate the longitudinal nature of the Tehran Lipid and Glucose Study (TLGS) data into the RF models. Our objective is to predict prediabetes based on key indicators utilizing the GMERF method. The results will be compared with those obtained from GLMM and a standard RF model. The investigators aim for these findings to help identify individuals at risk of developing prediabetes, thereby contributing to a reduction in the disease's burden on the population.

Materials and methods

Data source

The Tehran Lipid and Glucose Study (TLGS) is the first prospective population-based study in Iran, initiated in 1998, aimed at determining risk factors for non-communicable diseases, including glucose and lipid disorders, obesity, smoking, low physical activity, and hypertension, with the goal of improving lifestyle to prevent these risk factors. This study was conducted on a group of residents aged 3–69 years in District 13 of Tehran. District 13 was chosen as the sampling framework due to the greater stability of its residents, the presence of healthcare centers affiliated with the University of Medical Sciences, and the demographic similarities (age and gender distribution) of its population to that of the broader Tehran population. The participants were invited to the Lipid and Glucose Research Unit through face-to-face contact, where their data were collected through interviews, physical examinations, serum tests, and anthropometry measurements. This study consists of seven phases, with measurements repeated every three years. The design and methodology of the TLGS have been thoroughly detailed in previously published works²⁸.

Initially, 15,011 subjects participated in the Tehran Lipid and Glucose Study. Due to dropouts during the first phase and enrolling new subjects in the second phase, the combined data from both phases were treated as the initial phase of the study. In other words, this combined phase included participants from both phases. Ultimately, the total study population consisted of 12,240 subjects.

Inclusion criteria

The inclusion criteria for this study were: age over 20 years, participants who were healthy in phase one (i.e., without prediabetes, diabetes, or cancer), and individuals who participated in at least two phases of the study. After applying these criteria, 5361 participants remained eligible for the study. Since participants might develop pre-diabetes between phases, the independent risk factors from the previous phase were considered for each outcome phase. For example, the risk factors from the second phase were used for the third phase outcome,

and similarly, up to the seventh phase, where the independent variables from the sixth phase were considered. According to this definition, five phases were identified in total. The next scenario in the study was the possibility that individuals might progress directly from being healthy to developing diabetes without passing through a pre-diabetes stage. This could be due to unrecorded pre-diabetes in previous phases, non-attendance, or developing pre-diabetes in the time interval between phases. Therefore, the initial phase of diabetes onset for these individuals was classified as prediabetes, and any subsequent records for those individuals were removed from the study.

Main outcome and potential predictors

In this study, the primary outcome variable was prediabetes, defined as a binary variable based on fasting blood sugar (FBS) between 100 and 125.9 mg/dL or 2-h post-load blood sugar (BS2HR) between 140 and 199.9 mg/dL, with no use of anti-diabetic medication. In addition, 31 variables were initially treated as the potential factors influencing pre-diabetes. These variables included demographic factors (such as age, gender, marital status, occupation, and education), medical history (including the use of thyroid drugs, antihyperlipidemic drugs, antihypertensive drugs, cardiac drugs, corticosteroids, aspirin, family history of diabetes, family history of cardiovascular diseases (CVD), and history of CVD, as well as smoking and the metabolic equivalent of task (MET). Additionally, medical examination data including pulse rate, systolic blood pressure (SBP), diastolic blood pressure (DBP), and anthropometric measurements such as body mass index (BMI), waist circumference (WC), wrist circumference, hip circumference, waist-to-height ratio (WHtR), and waist-to-hip ratio (WHR) were considered. Blood biomarkers including HDL, total cholesterol (TC), creatinine, triglycerides (TG), Glomerular Filtration Rate (GFR), TC-to-HDL ratio, and TG-to-HDL ratio were also analyzed. In this study, individuals (ID) were used as a clustering variable, or, in other words, as a random effect. Comprehensive information regarding anthropometric and laboratory measurements, along with the techniques used, can be found in earlier published studies²⁹.

Statistical analysis

For each variable in the dataset, the proportion of missing data was below 20%. We assessed the missing data patterns and found no evidence of systematic bias, suggesting that the data were missing at random (MAR). Missing values were imputed using the Multiple Imputation by Chained Equations (MICE) package in R, employing the weighted predictive mean matching (WPMM) method. Five imputed datasets were generated and combined using Rubin's rules to account for imputation uncertainty³⁰. In the quantitative analysis, central tendency and variability measures were computed for numerical variables, while qualitative factors were summarized using frequency distributions. To evaluate model performance, the data was randomly split into 80% training and 20% test sets, ensuring that the outcome variable's distribution remained consistent across both data sets. The data splitting method ensured that all occurrences of a given ID were assigned to either the training or test dataset, not split between them.

In the second step, we employed the '*varSelRF*' and '*VSURF*' R packages to address the problem of feature multicollinearity. This step focused on identifying a definitive set of features most pertinent to predictive modeling, aiming to enhance the effectiveness of our ML techniques. The *varSelRF* method, developed by Diaz-Uriarte et. al., employs a backward elimination approach for the feature selection. This technique systematically removes the least important features while preserving a comparable error rate to that of the full model (including all features)³¹. The *VSURF* is a method that utilizes a stepwise procedure introduced by Genuer et al. This approach also ensures that the error rate remains similar to that of the full model³². We selected these two algorithms because they are robust to multicollinearity, widely validated in biomedical studies, and well-suited to high-dimensional data. Their ensemble nature reduces variance and overfitting, thus improving the generalizability of the models^{31,33}.

Then, the predictor variables were utilized to fit a multivariable random intercept mixed-effects logistic regression (LR) model using the *glmer* package. The model parameters were estimated using the restricted maximum likelihood (REML) method and Z-normalization was finally applied to the numerical variables.

The GMERF is based on a GLMM, where the fixed effects are estimated using an RF, while the dependence structure is preserved through random effects. The RF aims to handle interactions among covariates and addresses highly nonlinear effects.

The severe class imbalance in our longitudinal data, with only 25% of cases belonging to the positive class (those with prediabetes), posed a significant challenge. This imbalance can hinder the model's ability to accurately classify cases, leading to low sensitivity (true positive rate) and potentially low specificity (true negative rate). To address this issue, we adjusted the decision threshold, lowering it to improve sensitivity, though at the expense of specificity. The cutoff point was chosen to ensure that the proportion of instances assigned to the positive class in the training set closely matched the true proportion of the positive class³⁴. Finally, we compared the performance of the GMERF model with the GLMM and RF models in predicting prediabetes. The evaluation was based on various metrics, including accuracy, sensitivity, specificity, and the area under the ROC curve (AUC).

It is crucial to note that the validation parameters were determined using the ten-fold cross-validation technique, tailored to each algorithm type. For the RF algorithm specifically, we fine-tuned several parameters: the total quantity of decision trees in the RF³⁵, the maximum depth of trees (deeper trees can capture more intricate patterns but risk overfitting³⁶ and the number of features considered at each node split³⁵. Other parameters were left at their default settings. The prediction error was quantified as the average absolute difference between predicted and observed values, corresponding to the predictive misclassification rate (PMCR).

Prediction models based on ML were developed using the variables obtained from comparing two feature selection methods. To ensure robust performance and avoid overfitting, we employed the K-fold cross-validation procedure. This approach helps prevent artificially inflated validation metrics by providing a more reliable

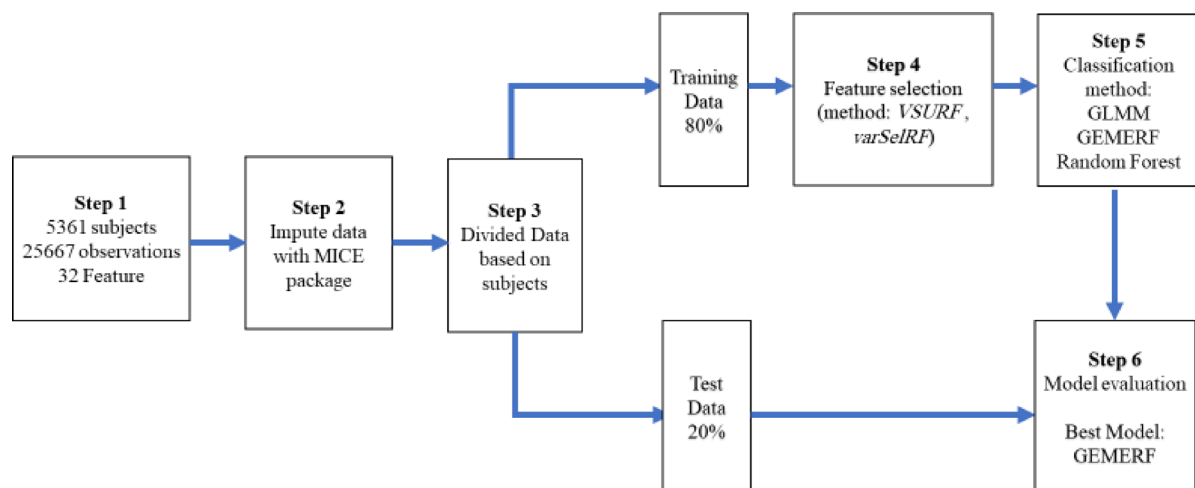


Fig. 1. Flowchart of the study design.

Phase	Gender	No	Age (Mean $\pm$ SD)	Prediabetes No. (%)
1	Male	2306	41.75 $\pm$ 14.71	381 (16.52)
	Female	3055	39.16 $\pm$ 13.09	407 (13.32)
	Total	5361	40.28 $\pm$ 13.87	788 (14.70)
2	Male	2249	45.27 $\pm$ 14.75	612 (27.21)
	Female	2981	42.41 $\pm$ 12.77	646 (21.67)
	Total	5230	43.64 $\pm$ 13.73	1258 (24.05)
3	Male	2219	47.63 $\pm$ 14.25	687 (30.96)
	Female	2934	45.53 $\pm$ 12.08	741 (25.26)
	Total	5153	46.43 $\pm$ 13.49	1482 (27.71)
4	Male	2167	50.56 $\pm$ 14.20	717 (33.09)
	Female	2865	48.39 $\pm$ 12.56	752 (26.25)
	Total	5032	49.32 $\pm$ 13.33	1469 (29.19)
5	Male	2104	53.39 $\pm$ 13.97	705 (33.51)
	Female	2787	51.65 $\pm$ 12.50	756 (27.13)
	Total	4891	52.40 $\pm$ 13.18	1461 (29.87)

Table 1. Percentage of prediabetes for each phase of the study.

estimate of model performance³⁷. Figure 1 presents a detailed overview of the study design framework. All statistical analyses were conducted using the R version 4.4.1, with *P* values less than 0.05 considered statistically significant.

Ethical approval

The research was carried out in alignment with ethical guidelines and national standards for medical research in Iran. All procedures involving human participants in the Tehran Lipid and Glucose Study (TLGS) were approved by the National Research Council of the Islamic Republic of Iran, and written informed consent was obtained from each participant.

Additionally, the study received approval from the Ethics Committee of Shahid Beheshti University of Medical Sciences under approval number IR.SBMU.RETECH.REC.1402.593.

Results

This study analyzed data from a cohort of 5361 individuals. Table 1 presents the distribution of participants by gender, along with the mean (SD) age and the percentage of prediabetes for each phase of the study. In addition, Table 2 displays the general characteristics of the population under study. Based on the varSelRF and VSURF packages, we included the variables identified as important by both methods in the modeling process. Regarding this, 21 variables (features) were selected to include the described models (Fig. 2). The selected features were: age, gender, education, antihyperlipidemic drugs, antihypertensive drugs, SBP, DBP, BMI, WC, wrist circumference, hip circumference, HDL, TC, creatinine, TG, GFR, WHtR, WHR, TC-to-HDL ratio, TG-to-HDL ratio, and Time (Phase).

Variable	Category	$\bar{X} \pm SD$ or %
Gender	Male	43.03
	Female	56.97
Occupation	Employed	58.20
	Unemployed	41.80
Education	Primary	20.69
	Secondary	55.09
	Higher	24.22
Marital status	Married	82.50
	Single/Divorced or Widowed	17.50
Corticosteroids	Yes	1.45
Aspirin use	Yes	6.44
Antihyperlipidemic drugs	Yes	5.15
Cardiac drugs	Yes	2.12
Antihypertensive drugs	Yes	7.58
Thyroid drugs	Yes	4.92
Smoking status	Yes	8.52
CVD history in 1 st degree family	Yes	6.01
Diabetes history in 1 st degree family	Yes	12.13
Physical activity	< 600	74.68
	≥ 600 and < 1500	16.80
	≥ 1500	8.52
History of CVD	Yes	11.42
Age (year)	–	46.28 $\pm$ 14.18
Pulse rate	–	78.24 $\pm$ 10.74
SBP (mmHg)	–	114.24 $\pm$ 17.09
DBP (mmHg)	–	75.35 $\pm$ 10.49
BMI (kg/m ²)	–	27.49 $\pm$ 4.60
WC (cm)	–	92.15 $\pm$ 11.79
Wrist circumference (cm)	–	16.54 $\pm$ 1.48
Hip circumference (cm)	–	100.78 $\pm$ 8.32
HDL (mg/dL)	–	45.41 $\pm$ 12.03
Total cholesterol (mg/dL)	–	190.93 $\pm$ 38.69
Creatinine (mg/dL)	–	1.06 $\pm$ 0.23
Triglyceride (mg/dL)	–	141.30 $\pm$ 82.69
GFR	–	70.93 $\pm$ 15.19

Table 2. General characteristics of the population under study.

In the next step, a multivariate mixed-effects LR model was utilized to identify the risk factors for prediabetes within the study population. Table 3 presents the estimated values obtained from the analysis. The results indicate that the variables age, education, antihyperlipidemic drugs, antihypertensive drugs, SBP, BMI, Waist circumference, Hip circumference, HDL, TC, TG, WHR, TC-to-HDL were significantly linked to the presence of prediabetes in the cohort under investigation ($P < 0.05$).

Additionally, we applied a conventional RF model to the dataset. The variable importance indices obtained from the fitted RF model are displayed in Fig. 3. Based on these findings, it can be inferred that five variables emerged as the most significant predictors of prediabetes: WHR, TG-to-HDL, WHtR, BMI, and TC-to-HDL. Conversely, the five variables with the lowest importance scores were: time, education, antihypertensive drugs, antihyperlipidemic drugs, and gender.

In the third stage of data analysis, we fitted the GMERF model to the data. The most influential predictors identified by the GMERF model were: WHR, age, WC, time (phase), and TG. The least important variables were creatinine, education, antihypertensive drugs, antihyperlipidemic drugs, and gender (Fig. 4).

Figure 6 presents the ranked importance of all 21 predictor variables in both the RF and GMERF models. The figure highlights shifts in variable importance between models using connecting lines, allowing for visual comparison of their relative rankings.

Lastly, we calculated the predictive power indices for the conventional RF, GLMM, and GMERF models (Table 4). The results indicate that the GMERF model outperforms the other models in terms of predictive capability. The RF model exhibited the lowest sensitivity (0.62) and specificity (0.63) values among the fitted models. To visually compare the estimated predictive criteria of these models, we assessed their effectiveness in

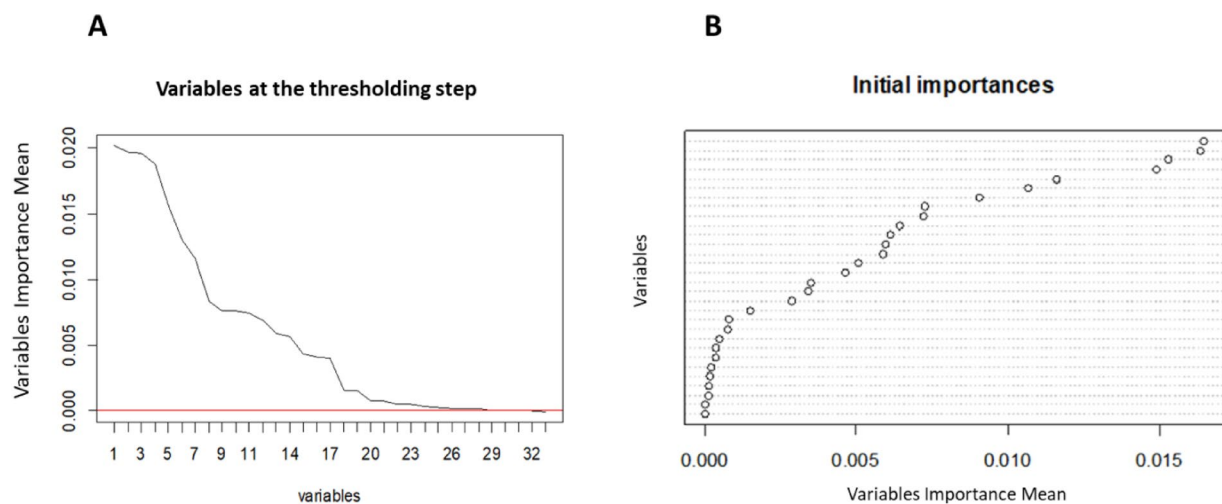


Fig. 2. Feature selection of variables using: (A) VSURF algorithm, (B) varSelRF algorithm.

Variable	Subgroup	Adjusted OR (95%CI)	p value
Gender (Female)		0.93 (0.77,1.13)	0.457
Age		1.25 (1.17,1.34)	<0.001
Education	Primary	-	-
	Secondary	0.98 (0.87,1.09)	0.700
	Higher	0.84 (0.73,0.97)	0.016
Antihyperlipidemic drugs use		1.19 (1.03,1.38)	0.020
Antihypertensive drugs use		1.16 (1.02,1.32)	0.030
SBP		1.08 (1.03,1.14)	0.003
DBP		1.11 (0.94,1.14)	0.660
BMI		1.22 (1.09,1.36)	<0.001
WC		2.00 (1.34,3.55)	0.017
Wrist circumference		1.07 (0.99,1.16)	0.072
Hip circumference		1.66 (1.42,1.86)	0.006
HDL		1.31 (1.69,1.86)	<0.001
TC		1.15 (1.02,1.29)	0.020
Creatinine		1.05 (0.89,1.12)	0.165
TG		1.36 (1.17,1.58)	<0.001
GFR		1.03 (0.93,1.07)	0.340
WHtR		1.03 (0.96,1.29)	0.111
WHR		1.18 (1.07, 1.21)	0.010
TC-to-HDL		1.21 (1.01,1.29)	0.030
TG-to-HDL ratio		1.16 (0.98,1.28)	0.080

Table 3. Results of mixed-effects logistic regression analysis of risk factors for prediabetes in the TLGS. *BMI* body mass index; *WC* waist circumference; *HDL* high-density lipoprotein; *SBP* systolic blood pressure; *DBP* diastolic blood pressure; *GFR* glomerular filtration rate; *TG* triglycerides; *TC* total cholesterol; *WHR* waist-to-hip ratio; *WHtR* waist-to-height ratio.

predicting prediabetes using the ROC curve. In this context, the RF and GMERF models showed, respectively, the lowest and highest area under the ROC curve, with AUC of 0.65 and 0.75. Additionally, the GLMM model demonstrated better performance than the RF model, with an estimated AUC of 0.73 (Fig. 5).

Discussion

Early identification of individuals at risk for prediabetes is crucial for preventing the onset of type 2 diabetes^{38,39}. Despite the benefits of early intervention, no effective strategy currently exists for large-scale prediabetes screening^{40,41}. Our research explored the use of ML approaches to create robust predictive models for diseases. An analysis of predictive power indices showed that the GMERF model outperformed other evaluated models. The GLMM ranked second in performance, yielding results very similar to those of the GMERF model. The

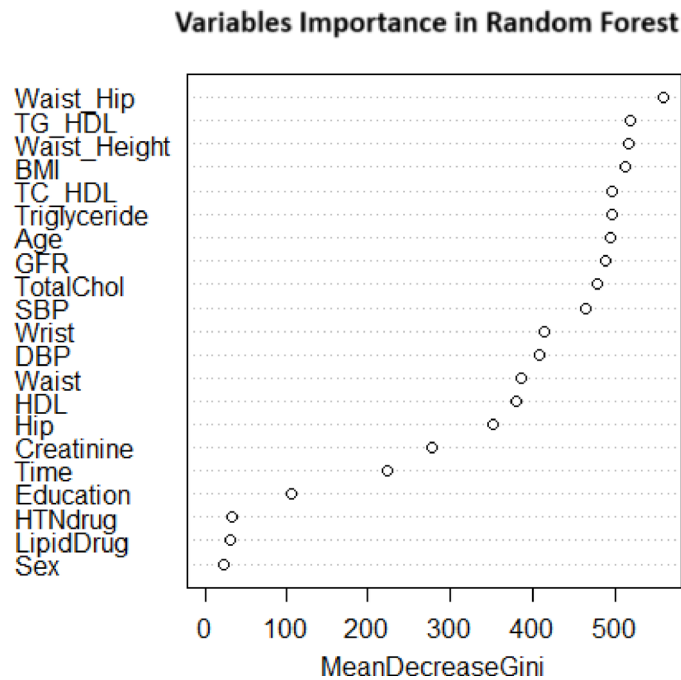


Fig. 3. Importance index of variables in RF. TG_HDL, TG-to-HDL ratio; TC_HDL, TC to HDL ratio.

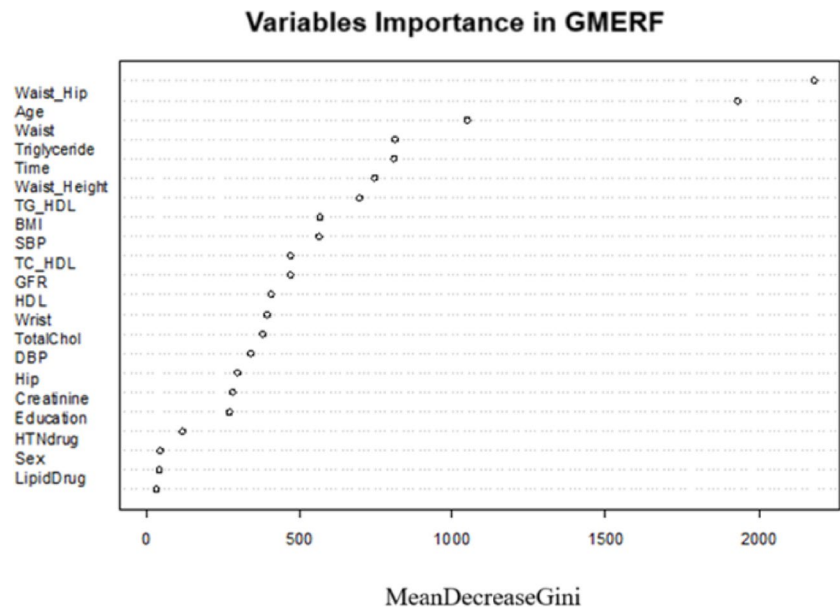


Fig. 4. Importance index of variables in GMERF. TG_HDL, TG-to-HDL ratio; TC_HDL, TC to HDL ratio.

Model	Sensitivity (Recall)	Specificity	Precision	F1-score	Accuracy (95% CI)	AUC (95% CI)
GLMM	0.68	0.70	0.84	0.75	0.68 (0.67, 0.70)	0.73 (0.72,0.75)
RF	0.62	0.63	0.83	0.69	0.62 (0.60,0.63)	0.65 (0.63,0.67)
GMERF	0.70	0.74	0.85	0.77	0.71 (0.70,0.73)	0.75 (0.74,0.77)

Table 4. Comparison of predictive power indices for GLMM, RF, and GMERF.

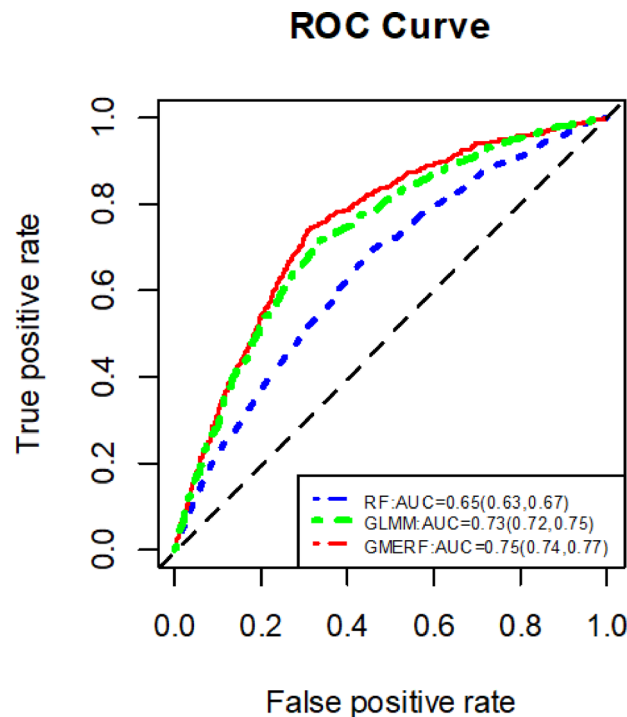


Fig. 5. Receiver operating characteristic (ROC) curve for different machine learning models and GLMM.

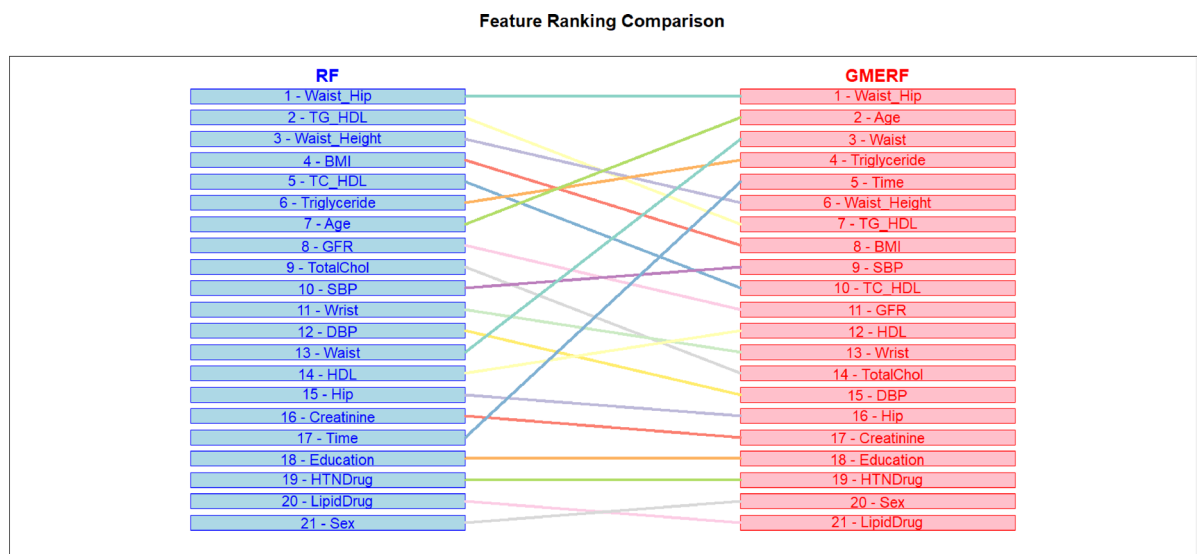


Fig. 6. Variable importance comparison between RF and GMERF models for all 21 predictors.

RF model demonstrated lower accuracy compared to both. This is in agreement with the results of the study conducted by Asadi et al. to diagnose cardiovascular disease (CVD) in clustered data. Among the four statistical models—Generalized Linear Mixed Model tree (GLMMtree), decision tree (DT), GMERF, and RF—the GMERF model showed the best performance, while the RF model had the lowest accuracy⁴² (Fig. 6).

According to the findings from the GMERF model in our study, waist-to-hip ratio and age have the greatest impact on predicting prediabetes. Similarly, in the GLMM model, these variables also had a significant effect on the outcome. A cross-sectional study conducted in Vietnam, which included 2142 individuals aged 30–72 years, found that waist-to-hip ratio and systolic blood pressure are significant predictors of type 2 diabetes risk. The area under the ROC curve for these two risk factors was 0.73 for men and 0.69 for women⁴³. Numerous studies have explored the use of ML for predicting and diagnosing diabetes, presenting various prediction models and comparing their sensitivity and ROC curves^{44–47}. In a cross-sectional study conducted in 2024, a total of 3376 individuals over the age of 30 in Iran, who participated in a diabetes screening program across 16 comprehensive

health service centers, were included. Five different ML algorithms—CatBoost, RF, XGBoost, logistic regression (LR), and artificial neural networks (ANN)—were used to model the dataset. Among these models, CatBoost performed slightly better, achieving an overall AUC of 0.737. Age, WHR, BMI, and SBP were identified as the most important factors for predicting diabetes⁴⁸. A cohort of 36,652 individuals from a rural population in China was analyzed to predict the risk of T2DM using six machine learning algorithms: logistic regression (LR), classification and regression tree (CART), artificial neural network (ANN), support vector machine (SVM), random forest (RF), and gradient boosting machine (GBM). Model performance was evaluated using 30 variables in each algorithm, except for the CART model⁴⁹. In that study, a high urinary glucose level was identified as the most significant variable by all algorithms. However, urinary glucose was not measured in our study, which limits direct comparison. Despite this, obesity indicators such as WHR and WHtR consistently ranked among the most important variables across all ML models in both studies, aligning with our findings. Other significant risk factors identified in the Chinese cohort included gender, age, family history of diabetes, WC, BMI, and SBP. In our analysis, however, gender and family history of diabetes were not among the top predictive variables⁵⁰.

Our study demonstrated that WC is an important additional factor in the prediabetes prediction model. In a study conducted by Hu et al. in 2020 on 22,945 prediabetic individuals who were followed for eight years, it was concluded that those who later developed diabetes experienced a greater annual increase in BMI and WC compared to those who did not develop diabetes⁵¹. Additionally, Alperet et al. demonstrated that for diagnosing diabetes mellitus, central obesity criteria such as WC and WHtR are more effective than BMI in three major Asian ethnic groups (Chinese, Malay, and Asian-Indian)⁵². In Europe, WC has been identified as a stronger predictor of diabetes risk compared to BMI⁵³. Prospective analyses in the Iranian population also revealed comparable associations between BMI and WC in predicting diabetes progression⁵⁴.

In an analytical cross-sectional study conducted in Mashhad, the statistical population was selected using a stratified-cluster sampling method, comprising 8810 individuals aged 35–64 years. LR models were utilized to identify determinants of prediabetes, revealing significant statistical associations between age (OR = 1.055), WC (OR = 1.02), and TG (OR = 1.002) with prediabetes. Furthermore, gender, marital status, and tobacco use did not show significant relationships with prediabetes. The prevalence of prediabetes in this study was 10.2% (885 cases)⁵⁵. In a study conducted in South Africa involving 6442 individuals aged over 15, it was shown that gender and WC have significant relationships with prediabetes⁵⁶, whereas, in our study, the gender variable was not significantly important.

Interestingly, gender did not emerge as a significant predictor of prediabetes in our study, which contrasts with some previous research reporting gender-related differences in metabolic risk^{57–59}. This discrepancy could stem from population-specific factors, such as genetic background, lifestyle behaviors, or cultural influences unique to the TLGS cohort. Additionally, our modeling approach, which adjusted for a comprehensive set of covariates and considered repeated measures, may have reduced residual confounding that could otherwise amplify gender effects. Another possible explanation is that the interaction between gender and other predictors (e.g., age, waist circumference, or lipid profiles) may have moderated the direct effect of gender itself. Further studies are warranted to explore potential gender interactions or to validate this finding in different populations and using alternative modeling strategies.

In our analysis, both triglyceride (TG) levels and time (i.e., study phase) emerged as important predictors of prediabetes, particularly in the GMERF model. The significant role of TG is consistent with previous epidemiological studies demonstrating its association with insulin resistance and impaired glucose metabolism. For instance, a large-scale longitudinal study by Guo et al. in 2024 showed that elevated TG levels were independently associated with increased risk of type 2 diabetes, even after adjusting for other metabolic risk factors⁶⁰. Our findings align with this evidence, as TG was significantly associated with prediabetes across all three modeling approaches (GLMM, RF, GMERF). Similarly, “time” was identified as an important variable in the GMERF model. This reflects the dynamic nature of prediabetes risk, which can evolve over different phases of follow-up due to aging, lifestyle changes, and cumulative exposure to risk factors. In longitudinal or multi-phase cohort studies, time functions as a proxy for these evolving influences. Its predictive importance in our model suggests that incorporating temporal dynamics may improve risk stratification, particularly in settings where data are collected across multiple time points or phases^{61,62}.

This study represents the first application of an ML model that accounts for the correlation structure of the data to the TLGS dataset. A key aspect of our approach is the comprehensive feature selection and model development process. We utilized the power of ML algorithms to systematically assess the significance of these features in predicting prediabetes. Interestingly, our results indicated that the RF model exhibited the lowest accuracy compared to GMERF and GLMM. This discrepancy may arise from the inherent assumptions of the RF algorithm. Like many other machine learning methods, RF assumes that observations are sampled independently from the population, which may not hold true in longitudinal settings. Although RF assumes independent observations and is therefore suboptimal in this context, its widespread use in clinical prediction tasks and decision-support systems makes it a relevant comparator in real-world healthcare applications. In longitudinal data, however, this assumption is often violated due to the inherent dependence between observations collected over time. Despite this limitation, RF was included as a benchmark model in our study to highlight the potential drawbacks of neglecting intra-individual correlations. Its inclusion allowed us to contrast traditional ML methods with those designed to handle clustered or repeated measures data. Ignoring this dependence can result in biased inferences, such as underestimated standard errors in linear models⁶³. This bias may help explain the subpar accuracy observed in the RF model. In contrast, classification methods that properly account for the data structure and effectively handle the correlations from repeated measurements generally exhibit better predictive performance^{64,65}. By incorporating these considerations, GMERF and GLMM models are better equipped to handle the complexities of longitudinal data, leading to more reliable and accurate predictions with

direct relevance to clinical risk stratification and early intervention. Conversely, the GLMM model exhibited accuracy that was nearly comparable to, but slightly lower than, that of GMERF. This difference may arise from the GLMM's reliance on certain assumptions, including a linear relationship between the response and the predictors, as well as a normal distribution of the residuals. Departures from these assumptions may result in inaccurate interpretations, especially in mixed-effects models. In contrast, the GMERF method does not impose these assumptions, allowing for greater flexibility in modeling.

Therefore, the GMERF model holds promise for integration into clinical screening protocols, particularly for identifying individuals at elevated risk of prediabetes in longitudinal cohort settings, where capturing intra-individual variation is essential. Beyond the statistical performance of the models, the clinical implications of our findings are of significant importance. The identification of waist-to-hip ratio (WHR), age, and triglyceride (TG) levels as key predictive factors offers a practical advantage, as these are non-invasive, low-cost, and routinely collected measures in primary care. WHR, in particular, emerged as a dominant predictor, supporting its integration into prediabetes risk assessment tools, especially in resource-limited settings. Age is an easily obtainable demographic factor that can help stratify individuals into risk categories, while TG levels represent a modifiable metabolic risk factor that can guide early lifestyle or pharmacological interventions. By translating these findings into clinical screening protocols and public health strategies, healthcare providers can identify at-risk individuals earlier and more effectively, enabling timely preventive interventions and potentially reducing the burden of diabetes.

Finally, we acknowledge that the TLGS data are based on a Tehran-specific urban population, which may raise concerns regarding the generalizability of the findings to other geographic or ethnic groups. However, previous studies have shown that the TLGS cohort is demographically representative of both the city of Tehran and the broader Iranian population in terms of age and sex distribution^{66,67}. Nevertheless, external validation using independent datasets from other populations is necessary.

Conclusion

This study acknowledges the inherent dependence between observations collected over time, and we account for this data structure in our models. In the current work, we utilized approaches for feature selection and model development. The integrated application of feature selection and machine learning offers a broad array of potential risk factors—such as demographic details, medical history, medical examinations, anthropometric measurements, and blood biomarkers—enabling more accurate predictions of prediabetes. This approach enhances our understanding of potential prediabetes risk ratios at the public health level, facilitating the implementation of more effective diabetes prevention and control strategies. These findings can guide targeted screening and early intervention efforts to delay or prevent progression to type 2 diabetes. In addition, future research should focus on validating these models on external datasets, testing additional machine learning algorithms, and incorporating time-varying predictors to enhance model performance and generalizability.

Study's limitations

This study has some limitations that should be addressed in future research. Although the longitudinal design of the Tehran Lipid and Glucose Study (TLGS) is a key strength, we did not explore whether the importance of predictors changes over time. Analyzing time-dependent variable importance could provide valuable insights into how risk factors evolve, and we suggest that this could be an important direction for future research.

Regarding the GMERF model, while it demonstrated the best predictive performance with an AUC of 0.75, this accuracy is still moderate. There is potential to improve predictive power by exploring more advanced machine learning models, such as XGBoost or neural networks. However, these models do not naturally account for the inherent dependencies in longitudinal data and would require additional adaptations, which were beyond the scope of this study. Future studies could address this limitation by incorporating more complex models designed specifically for longitudinal data, and by considering additional predictors, such as dietary or genetic factors, to further enhance performance.

Although we addressed class imbalance by adjusting the decision threshold, we acknowledge that this approach has limitations. Methods like SMOTE, commonly used for cross-sectional data, are not suitable for longitudinal datasets due to the inherent correlation within subjects over time. Applying SMOTE would violate these dependencies and potentially introduce bias. Therefore, while threshold adjustment helped mitigate imbalance, this approach remains a limitation. Future research could explore imbalance-handling techniques specifically designed for longitudinal data, such as subject-level resampling or model-based.

Data availability

The datasets used and analyzed during the current study are not publicly available due to privacy and security concerns. However, interested researchers can request access to the original raw data by contacting the Endocrine and Metabolic Sciences Research Institute at Shahid Beheshti University of Medical Sciences via email at dkhalili@endocrine.ac.ir.

Received: 10 January 2025; Accepted: 20 October 2025

Published online: 21 November 2025

References

1. Grundy, S. M. Pre-diabetes, metabolic syndrome, and cardiovascular risk. *J. Am. Coll. Cardiol.* **59**(7), 635–643 (2012).
2. Ford, E. S., Zhao, G. & Li, C. Pre-diabetes and the risk for cardiovascular disease: A systematic review of the evidence. *J. Am. Coll. Cardiol.* **55**(13), 1310–1317 (2010).

3. Huang, Y., Cai, X., Mai, W., Li, M. & Hu, Y. Association between prediabetes and risk of cardiovascular disease and all cause mortality: Systematic review and meta-analysis. *BMJ*. **355** (2016).
4. Huang, Y. et al. Prediabetes and the risk of cancer: A meta-analysis. *Diabetologia* **57**, 2261–2269 (2014).
5. Edwards, C. M. & Cusi, K. Prediabetes: A worldwide epidemic. *Endocrinol. Metab. Clin.* **45**(4), 751–764 (2016).
6. Punthakee, Z., Goldenberg, R. & Katz, P. Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. *Can. J. Diabetes* **42**, S10–S15 (2018).
7. Nathan, D. M. et al. Impaired fasting glucose and impaired glucose tolerance: Implications for care. *Diabetes Care* **30**(3), 753–759 (2007).
8. Hadaegh, F. et al. High prevalence of undiagnosed diabetes and abnormal glucose tolerance in the Iranian urban population: Tehran Lipid and Glucose Study. *BMC Public Health* **8**, 1–7 (2008).
9. Javanbakht, M., Mashayekhi, A., Baradaran, H. R., Haghdoust, A. & Afshin, A. Projection of diabetes population size and associated economic burden through 2030 in Iran: Evidence from micro-simulation Markov model and Bayesian meta-analysis. *PLoS ONE* **10**(7), e0132505 (2015).
10. Moosazadeh, M. et al. Enrolment phase results of the Tabari cohort study: Comparing family history, lipids and anthropometric profiles among diabetic patients. *Osong Public Health Res. Perspect.* **10**(5), 289 (2019).
11. Group NDD. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes*. **28**(12), 1039–57 (1979).
12. Inzucchi, S. E. Diagnosis of diabetes. *N. Engl. J. Med.* **367**(6), 542–550 (2012).
13. Dall, T. M. et al. Detecting type 2 diabetes and prediabetes among asymptomatic adults in the United States: Modeling American diabetes association versus US preventive services task force diabetes screening guidelines. *Popul. Health Metr.* **12**, 1–14 (2014).
14. Yari-Boroujeni, R. et al. The association between leisure-time physical activity and blood pressure changes from adolescence to young adulthood: Tehran Lipid and Glucose Study. *Sci. Rep.* **13**(1), 20965 (2023).
15. Sedaghat, Z. et al. The effect of obesity phenotype changes on cardiovascular outcomes in adults older than 40 years in the prospective cohort of the Tehran lipids and glucose study (TLGS): Joint model of longitudinal and time-to-event data. *BMC Public Health* **24**(1), 1126 (2024).
16. Mehrahi, Y., Sarbakhsh, P., Houwing-Duistermaat, J. J., Zayeri, F. & Daneshpour, M. S. Assessment of SNP interactions affecting total cholesterol over time using logic mixed model: TLGS study. *Gene, Cell Tissue*. **2**(1), (2015).
17. Speiser, J. L. et al. BiMM tree: A decision tree method for modeling clustered and longitudinal binary outcomes. *Commun. Stat. Simul. Comput.* **49**(4), 1004–1023 (2020).
18. Fokkema, M., Edbrooke-Childs, J. & Wolpert, M. Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychother. Res.* **31**(3), 329–341 (2021).
19. Alkhamis, M. A., Al Jarallah, M., Attur, S. & Zubaid, M. Interpretable machine learning models for predicting in-hospital and 30 days adverse events in acute coronary syndrome patients in Kuwait. *Sci. Rep.* **14**(1), 1243 (2024).
20. Cahn, A. et al. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes Metab. Res. Rev.* **36**(2), e3252 (2020).
21. Yang, H. et al. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Inf. Fus.* **75**, 140–149 (2021).
22. Dagliati, A. et al. Machine learning methods to predict diabetes complications. *J. Diabetes Sci. Technol.* **12**(2), 295–302 (2018).
23. Ravaut, M. et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. *JAMA Netw. Open* **4**(5), e2111315 (2021).
24. Rabhi, S. et al. Temporal deep learning framework for retinopathy prediction in patients with type 1 diabetes. *Artif. Intell. Med.* **133**, 102408 (2022).
25. Raudenbush, S. W. Hierarchical linear models: Applications and data analysis methods. *Advanced Quantitative Techniques in the Social Sciences Series/SAGE* (2002).
26. Pellagatti, M., Masci, C., Ieva, F. & Paganoni, A. M. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat. Anal. Data Min. ASA Data Sci. J.* **14**(3), 241–257 (2021).
27. Fokkema, M., Smits, N., Zeileis, A., Hothorn, T. & Kelderman, H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav. Res. Methods* **50**, 2016–2034 (2018).
28. Azizi, F. et al. Tehran Lipid and Glucose Study (TLGS): Rationale and design. *Iran. J. Endocrinol. Metab.* **2**(2), 77–86 (2000).
29. Azizi, F. et al. Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials* **10**, 1–15 (2009).
30. Van Buuren, S. Flexible imputation of missing data: CRC press (2018).
31. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinf.* **7**, 1–13 (2006).
32. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. VSURF: An R package for variable selection using random forests. *R J.* **7**(2), 19–33 (2015).
33. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recogn. Lett.* **31**(14), 2225–2236 (2010).
34. Asadi, F. et al. Identifying risk indicators of cardiovascular disease in Fasa cohort study (FACS): An application of generalized linear mixed-model tree. *Arch. Iran. Med.* **27**(5), 239 (2024).
35. Rezaei, N. & Jabbari, P. Immunoinformatics of Cancers: Practical machine learning approaches using R: Academic Press (2022).
36. Duroux, R. & Scornet, E. Impact of subsampling and tree depth on random forests. *ESAIM Probab. Stat.* **22**, 96–128 (2018).
37. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
38. Federation, I. D. IDF Diabetes Atlas Brussels. Belgium: international diabetes federation (2021).
39. Bommer, C. et al. The global economic burden of diabetes in adults aged 20–79 years: A cost-of-illness study. *Lancet Diabetes Endocrinol.* **5**(6), 423–430 (2017).
40. Li, G. et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: A 20-year follow-up study. *The Lancet* **371**(9626), 1783–1789 (2008).
41. Gillies, C. L. et al. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: Systematic review and meta-analysis. *BMJ* **334**(7588), 299 (2007).
42. Asadi, F. et al. Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms. *Sci. Rep.* **14**(1), 22230 (2024).
43. Ta, M., Nguyen, K., Nguyen, N., Campbell, L. & Nguyen, T. Identification of undiagnosed type 2 diabetes by systolic blood pressure and waist-to-hip ratio. *Diabetologia* **53**, 2139–2146 (2010).
44. Saxena, R., Sharma, S. K., Gupta, M. & Sampada, G. A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Comput. Intell. Neurosci.* **2022**(1), 3820360 (2022).
45. Olisah, C. C., Smith, L. & Smith, M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput. Methods Programs Biomed.* **220**, 106773 (2022).
46. Llahi, O. & Rista, A., editors. Prediction and detection of diabetes using machine learning. RTA-CSIT (2021).
47. Lai, H., Huang, H., Keshavjee, K., Guergachi, A. & Gao, X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr. Disord.* **19**, 1–9 (2019).

48. Shojae-Mend, H., Velayati, F., Tayefi, B. & Babaee, E. Prediction of diabetes using data mining and machine learning algorithms: A cross-sectional study. *Healthcare Inf. Res.* **30**(1), 73–82 (2024).
49. Zhang, L., Wang, Y., Niu, M., Wang, C. & Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan rural cohort study. *Sci. Rep.* **10**(1), 4406 (2020).
50. Mahtab, N., Farzad, H., Mohsen, B. & Nakisa, D. The 10-year trend of adult diabetes, prediabetes and associated risk factors in Tehran: Phases 1 and 4 of Tehran Lipid and Glucose Study. *Diabetes Metab. Syndr.* **11**(3), 183–187 (2017).
51. Hu, H. et al. Trajectories of body mass index and waist circumference before the onset of diabetes among people with prediabetes. *Clin. Nutr.* **39**(9), 2881–2888 (2020).
52. Alperet, D. J., Lim, W. Y., Mok-Kwee Heng, D., Ma, S. & van Dam, R. M. Optimal anthropometric measures and thresholds to identify undiagnosed type 2 diabetes in three major Asian ethnic groups. *Obesity* **24**(10), 2185–2193 (2016).
53. Consortium, I. Long-term risk of incident type 2 diabetes and measures of overall and regional obesity: The EPIC-InterAct case-cohort study. *PLoS Med.* **9**(6), e1001230 (2012).
54. Janghorbani, M. & Amini, M. Comparison of body mass index with abdominal obesity indicators and waist-to-stature ratio for prediction of type 2 diabetes: The Isfahan diabetes prevention study. *Obes. Res. Clin. Pract.* **4**(1), e25–e32 (2010).
55. Navipour, E., Esmaily, H. & Ghayourmobarhan, M. Identify determinative factors the occurrence of pre-diabetes using logistic regression model in Mashhad. *J. Sabzevar Univ. Med. Sci.* **25**(5), 619–628 (2018).
56. Grundlingh, N., Zewotir, T. T., Roberts, D. J. & Manda, S. Assessment of prevalence and risk factors of diabetes and pre-diabetes in South Africa. *J. Health Popul. Nutr.* **41**(1), 7 (2022).
57. Kautzky-Willer, A., Harreiter, J. & Pacini, G. Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. *Endocr. Rev.* **37**(3), 278–316 (2016).
58. Wild, S., Roglic, G., Green, A., Sicree, R. & King, H. Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care* **27**(5), 1047–1053 (2004).
59. Vatcheva, K. P., Fisher-Hoch, S. P., Reininger, B. M. & McCormick, J. B. Sex and age differences in prevalence and risk factors for prediabetes in Mexican-Americans. *Diabetes Res. Clin. Pract.* **159**, 107950 (2020).
60. Guo, R., Wei, L., Cao, Y. & Zhao, W. Normal triglyceride concentration and the risk of diabetes mellitus type 2 in the general population of China. *Front. Endocrinol.* **15**, 1330650 (2024).
61. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. Applied longitudinal analysis: John Wiley & Sons (2012).
62. Verbeke, G. Linear mixed models for longitudinal data. Linear mixed models in practice: A SAS-oriented approach: Springer (2000). p. 63–153.
63. Bryk, A. S. & Raudenbush, S. W. Hierarchical linear models: applications and data analysis methods: Sage Publications, Inc (1992).
64. Mangino, A. A. & Finch, W. H. Prediction with mixed effects models: A Monte Carlo simulation study. *Educ. Psychol. Measur.* **81**(6), 1118–1142 (2021).
65. Hu, J. & Szymczak, S. A review on longitudinal data analysis with random forest. *Brief. Bioinf.* **24**(2), bbad002 (2023).
66. Azizi, F. et al. Cardiovascular risk factors in an Iranian urban population: Tehran Lipid and Glucose Study (phase 1). *Soz. Präventivmed.* **47**(6), 408–426 (2002).
67. Hadaegh, F. et al. High prevalence of undiagnosed diabetes and abnormal glucose tolerance in the Iranian urban population: Tehran Lipid and Glucose Study. *BMC Public Health* **8**(1), 176 (2008).

Acknowledgements

We would like to express our gratitude to the Endocrine and Metabolic Sciences Research Institute at Shahid Beheshti University of Medical Sciences for providing the data for this study. This paper was extracted from the PhD thesis of the first author, Department of Biostatistics, School of Allied Medical Sciences, and was approved with No43013765 by the Deputy for Research and Technology, Shahid Beheshti University of Medical Sciences.

Author contributions

Maryam Karimi: methodology, investigation, conceptualization, validation, formal analysis, writing original draft preparation. Yadollah Mehrabi: supervise, conceptualize, validate, edit, and review. Chiara Masci, Farid Zayeri, Yaser Mehrali and Ali Sheidaie: methodology, conceptualization. Davood Khalili: data curation, reviewing, and editing. Fatemeh Masebi: reviewing and editing.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Additional information

Correspondence and requests for materials should be addressed to Y.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025