

RESEARCH

Open Access



# Evaluating machine learning-powered classification algorithms which utilize variants in the GCKR gene to predict metabolic syndrome: Tehran Cardio-metabolic Genetics Study

Mahdi Akbarzadeh<sup>1</sup>, Nadia Alipour<sup>2</sup>, Hamed Moheimani<sup>3</sup>, Asieh Sadat Zahedi<sup>4</sup>, Firoozeh Hosseini-Esfahani<sup>5</sup>, Hossein Lanjanian<sup>4</sup>, Fereidoun Azizi<sup>6</sup> and Maryam S. Daneshpour<sup>4\*</sup> 

## Abstract

**Background:** Metabolic syndrome (MetS) is a prevalent multifactorial disorder that can increase the risk of developing diabetes, cardiovascular diseases, and cancer. We aimed to compare different machine learning classification methods in predicting metabolic syndrome status as well as identifying influential genetic or environmental risk factors.

**Methods:** This candidate gene study was conducted on 4756 eligible participants from the Tehran Cardio-metabolic Genetic study (TCGS). We compared predictive models using logistic regression (LR), Random Forest (RF), decision tree (DT), support vector machines (SVM), and discriminant analyses. Demographic and clinical features, as well as variables regarding common GCKR gene polymorphisms, were included in the models. We used a 10-repeated tenfold cross-validation to evaluate model performance.

**Results:** 50.6% of participants had MetS. MetS was significantly associated with age, gender, schooling years, BMI, physical activity, rs780094, and rs780093 ( $P < 0.05$ ) as indicated by LR. RF showed the best performance overall (AUC-ROC = 0.804, AUC-PR = 0.776, and Accuracy = 0.743) and indicated BMI, physical activity, and age to be the most influential model features. According to the DT, a person with BMI < 24 and physical activity < 8.8 possesses a 4% chance for MetS. In contrast, a person with BMI  $\geq 25$ , physical activity < 2.7, and age  $\geq 33$ , has 77% probability of suffering from MetS.

**Conclusion:** Our findings indicated that, on average, machine learning models outperformed conventional statistical approaches for patient classification. These well-performing models may be used to develop future support systems that use a variety of data sources to identify persons at high risk of getting MetS.

**Keywords:** Decision tree, Discriminant analysis, Logistic Regression, Metabolic syndrome, Random Forest, Support vector machines

## Introduction

The metabolic syndrome (MetS) refers to the simultaneous occurrence of a set of interrelated factors (high blood sugar, high blood pressure, dysregulated blood lipids, and

\*Correspondence: [daneshpour@sbmu.ac.ir](mailto:daneshpour@sbmu.ac.ir)

<sup>4</sup> Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran  
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

abdominal obesity) that increases the risk of cardiovascular diseases (CVD), type 2 diabetes (T2D) and different types of cancer [1]. The pathophysiological mechanisms behind the MetS are complex and involve genetic and environmental factors, such as lifestyle, diet, and physical inactivity [2].

The prevalence of the MetS is reported to be about 31% worldwide. While this number varies between genders and ethnicities, MetS is generally more prevalent among men and women of nations with aging populations [3, 4]. In Iran, approximately 33.7% of adults suffer from this syndrome [5]. As the rising prevalence of MetS among the aging Iranian population can lead to higher CVD rates and other devastating diseases, this affliction demands further investigative efforts in all aspects [6–8].

Designing predictive models that can aid in diagnosing patients who are more likely to have MetS, can aid preventive interventions designed to battle this syndrome as well as future cardiovascular complications that it engenders. While researchers have frequently been using clinical or demographic variables in these efforts, developing models that incorporate genetic variables is complex. As MetS is a complicated multifactorial disease, taking advantage of the data on well-researched MetS-associated genes has the potential to provide us with much more powerful predictive tools.

Glucokinase (GCK) enzyme is the primary glucose sensor in the liver and pancreatic cells. It regulates carbohydrate metabolism by adjusting biochemical pathways in glycogen synthesis, gluconeogenesis, and insulin release by pancreatic  $\beta$ -cells. Glucokinase regulatory protein (GKRP) binds to glucokinase and controls its intracellular location and activity. The glucokinase regulator (GCKR) gene resides on the short arm of chromosome 2 (2p23.3-p23.2), contains 19 exons, and encodes GKRP (68 kDa, 625 amino acids) [9, 10]. Genome-wide association studies (GWAS) and multiple candidate gene studies have reported GCKR variants to be associated with several metabolic parameters, including triglyceride (TG) levels [11–16], insulin resistance and fasting plasma glucose (FPG) levels [14, 15, 17] as well as metabolic disorders like T2DM [12, 15, 17], dyslipidemia (high TG and low high-density lipoprotein (HDL) cholesterol levels) [11, 13]. Common functional variants, rs780094, rs780093, and rs1260326, are the most researched genetic variants of the GCKR gene. Minor T-alleles of rs780094 and rs1260326 are linked to hypertriglyceridemia, lower insulin resistance, and plasma glucose levels. While these effects might seem like opposing factors in the development of MetS, some observational studies have found MetS to be more prevalent in individuals with the minor allele of these SNPs [14, 16–18]. Like rs780094, rs780093 is also a common intronic variant in the GCKR gene that

has been associated with polygenic dyslipidemia and high TG levels [19].

In this work, variants in the GCKR gene, as well as clinical and demographic measures, will be utilized to build predictive models for metabolic syndrome. Recently, researchers have used various machine learning algorithms to predict MetS. Methods such as decision tree, Random Forest [20, 21], and support vector machines (SVM) [22], among others, have achieved high performance in evaluations. Each algorithm has its strengths and weaknesses that might suit a particular data and question type. Here, we aimed to compare certain machine learning models (decision tree, Random Forest, support vector machines) with traditional statistical models (logistic regression, linear and quadratic discriminant analysis) developed on data from the participants of the Tehran Cardio-metabolic Genetic Study. We used models to obtain the most critical variables in predicting metabolic syndrome and finding the high-performing ones in classifying individuals regarding MetS.

## Method

### Overview and study population

Subjects for this work were selected from the Tehran Lipid and Glucose Study (TLGS). Research Institute for Endocrine Sciences (RIES) affiliated with Shahid Beheshti University of Medical Sciences, approved the study protocol and initiated TLGS in 1999. It is a dynamic cohort experiment that aims to study the risk and protective factors of non-communicable diseases in the Iranian population. 15,005 people from district 13 of Tehran have been recruited and followed through 6 phases [23]. Tehran cardio-metabolic genetic study (TCGS) is a prospective family-based cohort study within TLGS that aimed to create a comprehensive genome-wide database of the Tehranian population. Participants have been followed every three years, and at each phase, all the participants have signed written consent. Through the six phases of this study, genotype and phenotype data on 13,399 individuals have been gathered. Details on all aspects of this project, including the design and practical methods (phenotyping, genotyping, and quality controls) have been described elsewhere by Azizi F. et al. [24–26].

Of 15,005 participants who entered at the 6 phases of TLGS (1999–2017), 13,399 subjects were genotyped and were included in TCGS. From this population, for this candidate gene study, all people over 18 who were not diagnosed with MetS at the first phase were included. The following were excluded: people with missing genotyping information; participants younger than 19 years old; Participants who were prevalent cases of MetS at the first phase; participants whose baseline or follow-up data were not available; and individuals who did not consent

to participate. Ultimately 4754 eligible participants (2116 men and 2558 women) were selected for this work. A detailed flowchart of patient recruitment can be viewed in Fig. 1.

### Definition of terms

For the purposes of this work, metabolic syndrome (MetS) is defined with the joint interim statement (JIS) criteria [27], that is: the presence of at least 3 of the 5 following metabolic risk factors: (1) Hypertension as DBP  $\geq$  85 and SBP  $\geq$  130 mmHg, or antihypertensive medication; (2) Fasting HDL  $<$  40 mg/dL and  $<$  50 mg/dL in males and females respectively, or under lipid-lowering medication; (3) Fasting serum TG  $\geq$  150 mg/dL or under lipid-lowering medication; (4) Fasting plasma glucose (FPG)  $\geq$  100 mg/dL, or taking diabetes medication; (5) and central obesity (waist circumference (WC)  $\geq$  90 cm for both genders, based on the Iranian National Committee for Obesity guidelines). Based on the JIS criteria, individuals with at least three metabolic risk factors were considered as unhealthy cases. Others with a maximum of two from the mentioned risk factors were deemed

healthy controls. Smoking status was categorized as never, former smoker, current smoker, and second-hand smoker. For Marital status, four categories were defined: single, married, widowed, and divorced.

### Genetic analysis

Genomic DNA samples were extracted from the buffy-coat of venous blood samples using the standard proteinase K/salting out method. For qualitative estimation of the extracted DNA, a Thermo Scientific NanoDrop 1000 Spectrophotometer was used, and samples with low quality and concentration (DNA purification in the range of  $1.7 < A260/A280 < 2$ ) were excluded. DNA samples were genotyped with HumanOmniExpress-24-v1-0 bead chips (containing 649,932 SNP loci with an average mean distance of 4 kb) by deCODE genetics, Inc. (Reykjavik, Iceland) according to the manufacturer's specifications (Illumina Inc., San Diego, CA, USA). The PLINK program (V 1.07) and the R statistical software (V 3.2) used quality control procedures. The genotyping data of GCKR polymorphisms (rs780094, rs1260326, and 780,093) were used for association analysis.

### Statistical analysis

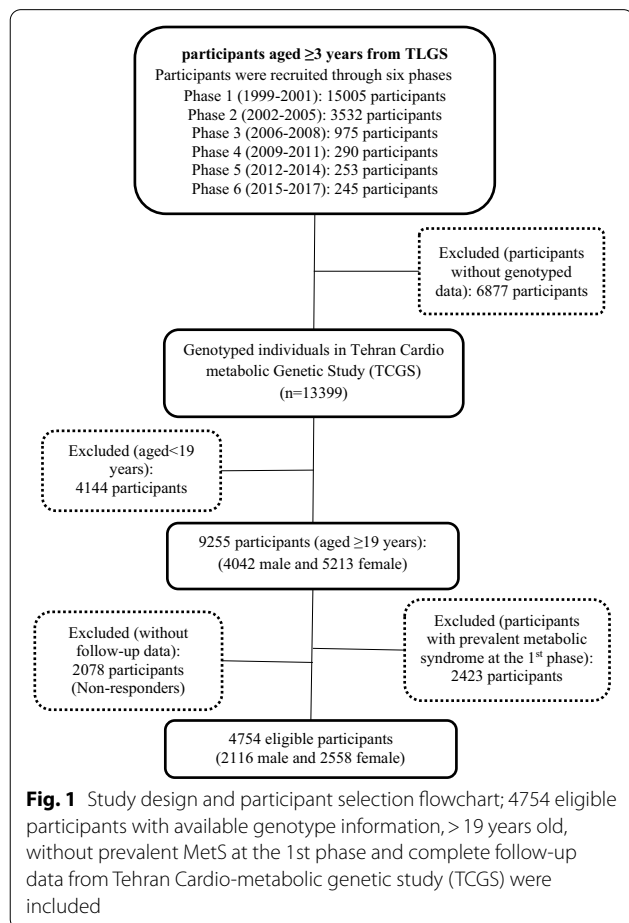
To find the essential predictors associated with metabolic syndrome, we compared classification machine learning (ML) algorithms, including the Random Forest (RF), Decision tree (DT), and Support vector machines (SVM), with three traditional statistical models: Logistic regression (LR), Linear discriminant analysis (LDA), and Quadratic discriminant analysis (QDA). The performance evaluation metrics are also reported by gender. All statistical analytical methods were performed using previously developed "randomForest", "MASS", "PRROC", "rpart", "caret", "e1071" R packages [28–34].

### Logistic regression

Logistic regression (LR), a standard classification method, models the probability of one of the two classes of a dichotomous outcome. Here, the linear combination of predictors is linearly fitted to the response variable's mean with a binomial distribution under the logit link function.

$$\log\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^k \beta_i x_i$$

P is the probability that a person has MetS,  $\alpha$  is the intercept. Xs denote the covariates (age, sex, schooling years, BMI, smoking status, marital status, physical activity, and SNP information of GCKR genotypes), and Bs represent regression coefficients.



### Discriminant analyses (Linear and Quadratic)

Discriminant analysis is one of the oldest classifiers first proposed by Fisher and is currently used in two major frameworks: Linear and Quadratic. These algorithms are based on the Bayes theorem and are different from LR in the classification task. These classifiers model the distribution of the independent variables (X) separately in each response class. They then use the Bayes theorem to estimate the probability of the X values' response levels. While linear discriminant analysis (LDA) computes the discriminant scores by finding the linear combination of independent variables that model and classify the response variable, the quadratic discriminant analysis (QDA) classifies the response variable with a non-linear combination of the predictors [35]. "MASS" package in R software was used to implement discriminant analyses [29].

### Decision tree

A decision tree (DT) is a supervised machine learning method used for regression and classification purposes [28]. DT predicts the target variable's value by learning simple rules represented by a decision tree. It includes three components: nodes, branches, and leaves. This algorithm classifies each sample by sorting them down the tree from the root to some leaf node. Each node in the tree specifies a test of a particular sample attribute, and each branch descending from that node corresponds to one of the possible values for this attribute. Each leaf represents the predicted value of the target variable given the values of the variables defined by the path from the root [36]. "rpart" package in R software was used to implement the decision tree algorithm [33].

### Random forest

Random forest (RF) is an ensemble-based learning algorithm Breiman [39] proposed first. It can be used for classification, regression, and unsupervised learning [28]. This algorithm is a set of non-pruned trees (classification trees based on the decision tree algorithm), and each tree is obtained by a recursive partitioning algorithm [37]. The algorithm for constructing an RF model with  $T$  trees from a dataset with  $n$  observations and  $p$  variables is as follows: (i) By the bootstrap method, a random sample with replacement with  $n$  number of observations is selected. (ii) A tree is created using the recursive partitioning algorithm for each sample. In each node, separation (partitioning) is performed based on a random sample of  $m$  number of predictive variable  $p$ . (iii) The recursive partitioning algorithm continues until the tree reaches its maximum size (i.e. terminal leaf node for each observation) without pruning the tree. (iv) the algorithm then iterates through the samples, and for each bootstrap

sample, steps 1–3 are repeated. The final output will be the mode of classes for classification tasks and the average of predictions for regression analyses [38]. Common choices for  $T$  are 1000 trees and for  $m$  is  $\sqrt{p}$  or  $\log(p)$  [39]. Interpreting the Random Forest model can be challenging, so we need to summarize the information generated using quantitative indicators such as the variable importance (VI). VI is an indicator used to rank the predictor variables based on their influence on the response variable. The most famous indices are the Gini and permutation. The "randomForest" package in R software was used to implement this algorithm [28].

### Support vector machines

A support vector machine is another common supervised learning algorithm proposed by Vapnik to deal with classification and regression analysis [40]. It is mainly used for binary classification problems and applies to linear and non-linear data classification tasks. SVM's goal is to find the best classification function to discriminate between the two classes present in the data set. SVM creates a hyperplane or multiple hyperplanes in a high-dimensional space. The best hyperplane optimally divides the data into different classes with the maximum separation and gap between the classes (highest margin). In its non-linear classification method, SVM utilizes various kernel functions (i.e., linear, polynomial, radial basis, and sigmoid) to estimate and maximize the hyperplane margins. "e1071" package in R software was used to implement the SVM algorithm [32].

### Model assessment (validation and comparison of the models)

To evaluate the model performance more precisely and decrease the potential variance between the estimates, we utilized 10-repeated tenfold cross-validation [41]. This procedure divides the data into 10 subsets, and each subset is used to evaluate the model exclusively trained on the other nine remaining subsets. The estimates of performance obtained from 10 repeated cross-validation are then averaged to get the overall performance indices such as sensitivity (SE), specificity (SP), accuracy (ACC), the area under the receiver operating characteristics curve (AU-ROC) and kappa. It is important to note that each subset's proportion of cases and controls was held the same. Each subset properly represented the main sample and the status of the underlying community.

For each evaluation task, a confusion matrix was drawn. Evaluation metrics were defined as follows: Sensitivity indicates the proportion of patients with MetS that the algorithms correctly classify as MetS positives. Specificity indicates the proportion of healthy subjects that the algorithms correctly classifies as MetS negatives.

Accuracy is the proportion of subjects among all participating individuals who were correctly classified as positives or negatives.

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \text{ Specificity} = \frac{TN}{FP + TN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

The receiver operating characteristic (ROC) curve is another useful indicator of model performance [41]. The X-axis and Y-axis of the ROC curve are sensitivity and 1-specificity, respectively [42]. The area under the ROC curve (AU-ROC) indicates the model's discriminative ability, and its values range from 0.5 to 1. The precision-recall curve can summarize the information prediction performance with a single value as with ROC curves. This summary statistic is referred to as the AUC-PR; area under the (precision-recall) curve. By and large, the higher the AUC-PR score, the better a classifier performs on a particular task. Values closer to one indicate higher model performance. "PRROC" and "caret" R packages were used to obtain relevant performance metrics [30, 31, 34].

## Results

### Study population characteristics

Of 4754 subjects in this study, 54.8% were female, and the participants' mean age was  $36.78 \pm 13.21$  years. Based on the JIS criteria, 2365 (50.6%) participants had MetS. Information on independent variables consisting of age, sex, schooling years, BMI, smoking status, marital status, physical activity, and SNP information of GCKR genotypes are described in Table 1. Here, the univariate p-values, calculated to compare the MetS positive and MetS negative groups on each predictor, are also presented.

For both genders, baseline characteristics and common SNPs of GCKR genotypes for study participants and non-responders of the TCGS population are shown in Table 2. Based on the results, there were no significant differences between responders and non-responders in males and females other than higher BMI and lower physical activity in male non-responders and different smoking and marital status distribution between female responders and non-responders.

We calculated adjusted OR value and its corresponding significance level for each predictor based on the logistic regression model. The results showed that metabolic syndrome was significantly associated with age, gender, schooling years, BMI, physical activity, rs780094, and rs780093 ( $P < 0.05$ ) (Table 3). Males were 2.373 times more at risk to have metabolic syndrome than females.

The odds of developing metabolic syndrome decreases with increasing education years ( $OR = 0.978$ ). On GCKR polymorphisms (rs780094, rs1260326, and 780,093), the results showed that MetS is associated with rs780094 and rs780093 and this relationship is caused by significantly more frequency of minor T alleles in patients with MetS.

### Performance comparison between machine learning algorithms

Table 4 summarizes the classification performance of various machine learning and traditional statistical methods based on the average value of 10-repeated tenfold cross-validation overall and by gender. Overall, the Random Forest showed higher classification accuracy (mean = 0.743) and area under the ROC curve (mean = 0.804) and AUC-PR (mean = 0.776). The decision tree ranked second in overall accuracy (mean = 0.738) with relatively high specificity (mean = 0.804) and AUC-PR (mean = 0.730). By and large, machine learning algorithms provided better accuracy, AUC-ROC, and AUC-PR compared with traditional statistical models. Accuracy, kappa, and AUC-ROC and AUC-PR of the machine learning models were higher overall and in both genders. While linear discriminant analysis (LDA) showed a high sensitivity (mean = 0.915), its specificity was considerably lower (mean = 0.230). Similar to the overall results, LDA provided the highest sensitivity in males (mean = 0.754). But in females, the highest sensitivity was achieved through logistic regression (mean = 0.798).

The importance of variables in the Random Forest model was calculated using the mean decrease Gini and mean decrease accuracy and is shown in Fig. 2. BMI, physical activity, and age were the most influential variables in both indices.

BMI present in the tree root was the most significant decision tree method and acted as the main prognostic factor. A combination of BMI + Physical activity + age is an accurate predictor for MetS. According to the induced decision tree shown in Fig. 3, the probability that an individual with  $BMI < 24$  and physical activity  $< 8.8$  has MetS is a mere 4%. In contrast, there is a 77% probability that a person with  $BMI \geq 25$ , physical activity  $< 2.7$ , and age  $\geq 33$  suffers from MetS.

## Discussion

This study aimed to compare the performance of machine learning-powered classification models, namely, support vector machines (SVM), decision tree (DT), and Random Forest (RF) in predicting metabolic syndrome, to that of three traditional classifiers: logistic regression (LR), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). Through developing such

**Table 1** Comparing independent demographic and genetic predictors of MetS in the healthy and unhealthy groups

Variables	Unhealthy (MetS) (%)	Healthy (No MetS) (%)	P value
Group size (%)	2365(50.6)	2309(49.4)	
Age (mean $\pm$ SD)	40.53 $\pm$ 12.93	33.04 $\pm$ 12.47	<0.001 <sup>a</sup>
Schooling years (mean $\pm$ SD)	9.19 $\pm$ 4.34	10.41 $\pm$ 4.63	<0.001 <sup>a</sup>
BMI (mean $\pm$ SD)	27.08 $\pm$ 4.09	23.8 $\pm$ 3.95	<0.001 <sup>a</sup>
Physical activity (mean $\pm$ SD)	575.16 $\pm$ 923.29	452.48 $\pm$ 808.96	<0.001 <sup>a</sup>
Sex (%)			
Male	1249(52.81)	867 (37.55)	<0.001 <sup>b</sup>
Female	1116 (47.19)	1442 (62.45)	
Smoking status (%)			
Never	1213(51.29)	1323(55.94)	<0.001 <sup>b</sup>
Former smoker	146(6.17)	73(3.09)	
Current smoker	336 (14.21)	259 (10.95)	
Second hand smoker	670 (28.33)	654 (27.65)	
Marital status (%)			
Divorced	24 (1.01)	19 (0.80)	<0.001 <sup>b</sup>
Married	1967 (83.17)	1612 (68.16)	
Single	312(13.19)	652(27.57)	
Widowed	62(2.62)	26 (1.10)	
rs1260326 (%)			
CC	662(27.99)	734 (31.04)	<0.01 <sup>b</sup>
TC	1156(48.88)	1090 (46.09)	
TT	547(23.13)	485 (20.51)	
rs780094 (%)			
CC	675 (28.54)	752 (31.80)	<0.01 <sup>b</sup>
TC	1156 (48.88)	1079 (45.62)	
TT	534 (22.58)	478 (20.21)	
rs780093 (%)			
CC	668(28.25)	735(31.08)	<0.01 <sup>b</sup>
TC	1143(48.33)	1107(46.81)	
TT	554(23.42)	476(20.13)	

MetS Metabolic Syndrome, BMI Body Mass Index, SD Standard Deviation; significant difference were observed in SNP information of GCKR genotypes and independent variables between healthy and unhealthy participants

<sup>a</sup> Student's-t test

<sup>b</sup> chi-square test

models on the eligible participants of the Tehran Cardio-metabolic genetic study (TCGS), we also obtained the most influential predictive features of MetS among clinical and GCKR polymorphism variables.

We found age, gender, schooling years, BMI, physical activity, and genetic variants of rs780094 and rs780093 significant risk factors for predicting metabolic syndrome. Despite their statistical significance, gender, schooling years, rs780094, and rs780093 did not influence the MetS prediction considerably. On the other hand, BMI, physical activity, and age were the most influential predictors of MetS, as indicated by the influence metrics of the Random Forest model. This result is in line with Fuentes et al., which denoted BMI as one of the

anthropometric variables associated with metabolic syndrome and essential for early detection [43].

The single nucleotide polymorphisms that showed significant relationships with metabolic syndrome in our predictive models agree with the findings of previous works that had examined the association between MetS and similar genetic markers [18, 44, 45].

Among classification machine learning methods that included SVM, DT and RF, RF had the best performance in classifying subjects on their MetS outcomes as indicated by the highest accuracy (0.743) as well as area under the receiver operating characteristic curve (AU-ROC) (0.804) and AUC-PR (0.776). This result is similar to the findings in the study by Szabo et al. that

**Table 2** Baseline characteristics of study participants and non-responders by common SNPs of GCKR genotypes

Variables	Male			Female		
	Responders (%)	Non-Responders (%)	P value	Responders (%)	Non-Responders (%)	P value
Group size (%)	2164 (72.28)	830 (27.72)		2590 (68.28)	1203 (31.72)	
Age(mean $\pm$ SD)	39.35 $\pm$ 14.59	41.48 $\pm$ 15.31	0.0614 <sup>1</sup>	36.25 $\pm$ 11.51	37.06 $\pm$ 14.6	0.0649 <sup>a</sup>
Schooling years (mean $\pm$ SD)	10.02 $\pm$ 4.65	9.91 $\pm$ 4.4	0.582 <sup>1</sup>	9.58 $\pm$ 4.4	9.32 $\pm$ 4.36	0.145 <sup>a</sup>
BMI (mean $\pm$ SD)	24.86 $\pm$ 3.87	25.12 $\pm$ 4.38	0.0246 <sup>1</sup>	26.31 $\pm$ 4.68	26.63 $\pm$ 4.92	0.0539 <sup>a</sup>
Physical activity (mean $\pm$ SD)	607.79 $\pm$ 1049.11	524.49 $\pm$ 1034.92	0.0510 <sup>1</sup>	429.78 $\pm$ 696.21	384.48 $\pm$ 682.76	0.0607 <sup>a</sup>
Smoking status (%)						
Never smoker	825 (38.12)	283 (34.10)	0.221 <sup>2</sup>	1732 (66.87)	717 (59.60)	0.002 <sup>b</sup>
Former smoker	219 (10.12)	88 (10.60)		8 (0.31)	13 (1.08)	
Current smoker	547(25.28)	196 (23.61)		59 (2.28)	35 (2.91)	
Second hand	573(26.48)	171 (20.60)		791 (30.54)	371 (30.84)	
Marital status (%)						
Divorced	12 (0.55)	1(0.05)	0.104 <sup>2</sup>	33 (1.27)	18 (1.50)	<0.001 <sup>b</sup>
Married	1628 (75.23)	649(29.99)		2011 (77.64)	954 (79.30)	
Single	520 (24.03)	176 (8.13)		462 (17.84)	155 (12.88)	
Widowed	4 (0.18)	3 (0.14)		84 (3.24)	76 (6.32)	
rs1260326 (%)						
CC	662 (30.59)	265 (31.93)	0.233 <sup>2</sup>	755 (29.15)	370 (30.76)	0.602 <sup>b</sup>
TC	1015 (46.90)	402 (48.43)		1265 (48.84)	574 (47.71)	
TT	487 (22.50)	163 (49.64)		570 (22.01)	259 (21.53)	
rs780094 (%)						
CC	676 (31.24)	251 (30.24)	0.30 <sup>2</sup>	773 (28.30)	358 (29.76)	0.998 <sup>b</sup>
TC	1015 (46.90)	414 (49.88)		1256 (48.49)	584 (48.55)	
TT	473 (21.86)	165 (19.88)		561 (21.66)	261 (21.70)	
rs780093 (%)						
CC	672 (31.05)	179 (21.57)	0.176 <sup>2</sup>	754 (29.11)	280 (23.28)	0.087 <sup>b</sup>
TC	1010 (46.67)	315 (37.95)		1275 (49.23)	404 (33.58)	
TT	482 (22.27)	125 (15.06)		561 (21.66)	214 (17.79)	

*BMI* Body Mass Index; There were no significant differences between responders and non-responders in males and females other than higher BMI and lower physical activity in male non-responders and different smoking and marital status distribution between female responders and non-responders

<sup>a</sup> Student's t-test

<sup>b</sup> chi-square test

applied the Random Forest algorithm for a similar task and calculated the accuracy of this method to be 71.4% [46–48]. Worachartcheewan et al. also implemented a Random Forest model to predict MetS in the Bangkok population and identify the most influential predictors. They found that the Random Forest algorithm predicted MetS status in adults aged 18 to 78 with high accuracy (98.11%) [49].

The decision tree was the second-best performing model, and its calculated measure for accuracy, sensitivity, specificity, AUC-ROC, and AUC-PR were 0.738, 0.667, 0.804, 0.771, and 0.730, respectively. Other works have also implemented the decision tree to detect metabolic syndrome with a sensitivity of 91.6% and specificity of 95.7% [43]. Results obtained from the

decision tree algorithm showed that a combination of BMI, physical activity, and age is an accurate predictor for predicting MetS. This agrees with a previous work by Huang et al. conducted to explore the association between lifestyle variables and metabolic syndrome and found that individuals with BMI > 27 kg/m<sup>2</sup> were predisposed to metabolic syndrome [50]. In another study by Worachartcheewan et al. that used a decision tree to diagnose metabolic syndrome, the results confirmed that BMI  $\geq$  25 is an important feature in diagnosing MetS [20]. In our work, the evaluation metrics for DT were almost similar to RF, and both outperformed the SVM. Karimi-Alavijeh et al. also employed DT and SVM to predict metabolic syndrome. In that investigation, SVM outperformed DT on several performance

**Table 3** Applying logistic regression to assess the significance of relationship between Independent demographic and genetic variables and metabolic syndrome

Variables		B	Odds ratio (OR)	P value
Age		0.025	1.025	<0.001
Gender (female = 0)		0.864	2.373	<0.001
Schooling years		- 0.021	0.978	0.009
BMI		0.207	1.230	<0.001
Physical activity		0.0001	1.000	0.005
Smoking status	Current smoker(reference)			
	Never smoker	0.005	1.005	0.962
	Former smoker	0.072	1.075	0.698
	Second hand	0.646	1.066	0.593
Marital status	Divorced (reference)			
	Married	- 0.087	0.916	0.808
	Single	- 0.198	0.820	0.595
	Widowed	- 0.010	0.989	0.981
rs1260326	CC(reference)			
	TC	0.206	1.229	0.347
	TT	0.472	1.603	0.133
rs780094	CC(reference)			
	TC	0.149	1.161	0.664
	TT	- 1.211	0.298	0.008
rs780093	CC(reference)			
	TC	- 0.122	0.884	0.664
	TT	1.066	2.903	0.002

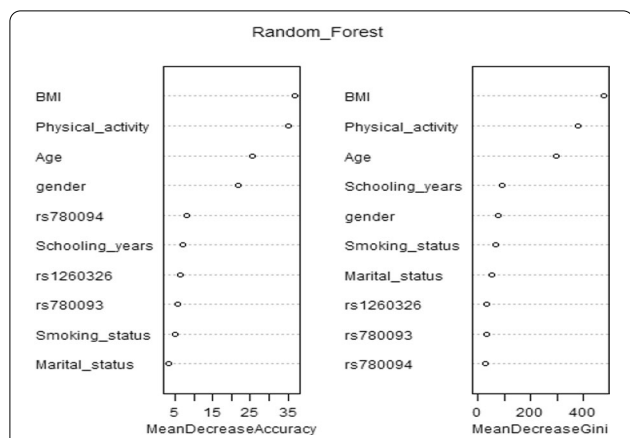
*BMI* Body Mass Index; logistic regression is used to predict the metabolic syndrome status of the participants in TCGs. The metabolic syndrome was significantly associated with age, gender, schooling years, BMI, physical activity, rs780094, and rs780093 ( $P < 0.05$ )

**Table 4** Performances metrics for LR, SVM, DT, RF, LDA, and QDA algorithms

	Models	Accuracy	Sensitivity	Specificity	Kappa	AUC-ROC	AUC-PR
Total	SVM	0.725	0.661	0.785	0.447	0.785	0.761
	DT	0.738	0.667	0.804	0.473	0.771	0.730
	RF	0.743	0.699	0.784	0.484	0.804	0.776
	LR	0.705	0.677	0.732	0.409	0.770	0.748
	LDA	0.562	0.915	0.230	0.141	0.658	0.666
	QDA	0.546	0.492	0.598	0.089	0.563	0.555
	Male	SVM	0.712	0.475	0.870	0.366	0.733
DT		0.735	0.527	0.874	0.421	0.739	0.753
RF		0.729	0.559	0.842	0.415	0.754	0.782
LR		0.711	0.519	0.839	0.373	0.732	0.768
LDA		0.591	0.754	0.482	0.217	0.679	0.734
QDA		0.547	0.394	0.649	0.044	0.531	0.616
Female		SVM	0.733	0.783	0.671	0.456	0.802
	DT	0.748	0.753	0.742	0.492	0.785	0.706
	RF	0.744	0.767	0.715	0.482	0.815	0.754
	LR	0.738	0.798	0.663	0.465	0.803	0.741
	LDA	0.608	0.752	0.427	0.184	0.664	0.617
	QDA	0.635	0.749	0.491	0.245	0.670	0.572

*LR* Logistic Regression, *SVM* support vector machines, *DT* Decision Tree, *RF* Random Forest, *LDA* Linear discriminant analysis, *QDA* Quadratic discriminant analysis, *AUC* Area Under Curve. Machine learning methods outperforms the traditional statistical methods



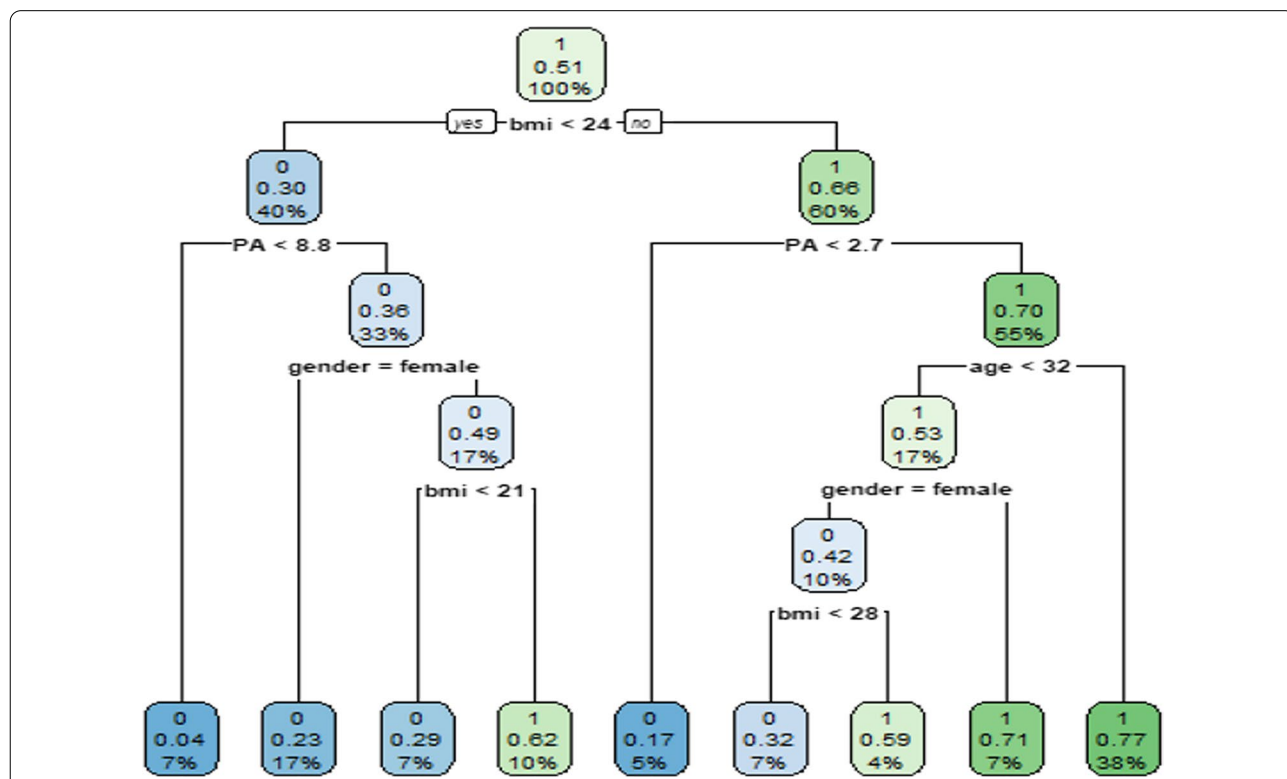


**Fig. 2** Assessing the importance of predictors with Gini and Accuracy importance indices based on the implementation of the random forest model; we confirmed that BMI, physical activity, and age were the most influential variables in MetS prediction

metrics (SVM (DT) model accuracy, sensitivity, and specificity were 0.774 (0.758), 0.74 (0.72), and 0.757 (0.739) [51].

To construct predictive models for MetS, other works have similarly employed various data mining methods

including artificial neural networks (ANN), beside decision tree, Random Forest, support vector machines, principal component analysis (PCA) and association analysis (AA). Their results showed that with an accuracy in the north of 99%, DT outperformed ANN and SVM, which provided lower accuracy metrics [52]. Other investigators have shown DT to be a robust machine learning method for constructing a predictive model of metabolic syndrome with reported accuracies of 73.90% [53] and 71.80% [54]. Lin et al. attempted to identify MetS in patients undergoing treatment with second-generation antipsychotics. They reported that logistic regression model had an accuracy as high as 83.6% and indicated BMI was an important predictor in identifying metabolic syndrome status [54]. This result contrasts with studies that found RF and SVM to be the most accurate classifiers for metabolic syndrome [22, 28, 55]. The complicated and multifactorial nature of metabolic syndrome and the severity of its complications require investigators to put further emphasis on the model sensitivity. While the quadratic discriminant analysis provided very low sensitivity, linear discriminant analysis (LDA) had the highest overall sensitivity. Similar to ours, other investigations have shown that LDA and RF are more sensitive classifiers than SVM, classification tree, and ANN [22, 56–58].



**Fig. 3** Classification decision tree, with probabilities of success for metabolic syndrome shown in each node; A combination of BMI, Physical activity, and age is an accurate predictor for the MetS

Compared with other recent studies conducted to develop predictive models for MetS, this work provides several advantages. It is important to emphasize that metabolic syndrome is a multifactorial disorder in which genetics, environmental factors, and lifestyle habits are all involved in disease pathogenesis. Unlike studies that exclusively use genetic variables, we developed our predictive models using clinically important and genetic information to provide more relevant results. In addition, past modeling efforts have less frequently developed both traditional and machine learning algorithms on big data for MetS prediction. The machine learning models developed through this effort have the advantage of providing good patient classification and indicating the most important risk factors. These models can be the basis of clinical tools that receive genetic and environmental information from patients as inputs and output their chances of having/developing MetS.

On the other hand, we should emphasize that researchers should be cautious when generalizing the results of this effort to other populations that were not represented in our study sample. Moreover, the effect of slight differences between responders and non-responders among participants on the study metrics remains unclear.

## Conclusion

It is essential to focus the resources on individuals who are most likely to develop or be already afflicted with these disorders to improve the potential effects of public health measures in reducing the burden of prevalent diseases such as metabolic syndrome. Traditional statistical models often fail to provide reliable predictive models when facing a multifactorial disorder with many potential independent genetic and environmental risk factors. However, as compared to conventional models, modern machine learning algorithms can enhance predicted accuracy in clinical concerns. Nonetheless, even when combined with genetic information, these models are insufficient for clinical application [59]. The first reason is that the sample size of such studies is insufficient to make a conclusive determination; the second reason is that whole-genome information is required in this regard; additionally, the ancestral discrepancy between populations necessarily requires that these models be considered separately for different ethnic groups [60, 61]. In this work, we compared predictive models for metabolic syndrome using the information on demographic and clinical and genetic data (functional variants of GCKR gene) on patients from TCGS. Our results proved modern methods, particularly Random Forest and decision tree, can provide high performing MetS predicting models that can help reduce

future cardiovascular, cancer, or other related complications when integrated within decision support tools or future investigations.

The study was the first step to predict the phenotypes using the polygenic risk score (PRS) as a modern method for disease prediction. The vital thing in the TCGS is discovering the best prediction model(s) for different diseases, especially MetS, which is multifactorial in terms of the definition and the etiology. Consequently, we decided to test the conventional models versus machine learning methods for known genes in our data to compare them on prediction ability.

## Abbreviations

AA: Association analysis; ACC: Accuracy; ANN: Artificial neural networks; AUC: Area under the curve; AU-ROC: Area under the receiver Operating characteristics; AUC-PR: Area under the precision-recall curve; BMI: Body Mass Index; CVD: Cardiovascular diseases; DBP: Diastolic Blood Pressure; DT: Decision tree; FN: False negative; FP: False positive; FPG: Fasting plasma glucose; GCK: Glucokinase; GCKR: Glucokinase regulator; GKR: Glucokinase regulatory protein; GWAS: Genome-wide association studies; HDL: High-density lipoprotein; JIS: Joint interim statement; LDA: Linear discriminant analysis; LR: Logistic regression; MetS: Metabolic syndrome; ML: Machine learning; OR: Odds ratio; PCA: Principal component analysis; PRS: Polygenic risk score; QDA: Quadratic discriminant analysis; RF: Random forest; RIES: Research Institute for Endocrine Sciences; ROC: Receiver operating characteristic; SBP: Systolic blood pressure; SD: Standard deviation; SE: Sensitivity; SP: Specificity; SVM: Support vector machines; T2D: Type 2 diabetes; TCGS: Tehran cardio-metabolic genetic study; TG: Triglyceride; TLGS: Tehran lipid and glucose study; TN: True negative; TP: True positive; WC: Waist circumference.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12967-022-03349-z>.

**Additional file 1.** Readers can refer to the supplementary material file to see more details of the accuracy indices of the models for each repeat and each fold.

## Acknowledgements

The authors would like to express their gratitude to the staff and participants in the TCGS project and to Sajedeh Masjoodi, who performed quality control on TCGS phenotypes. Special thanks to deCODE genetics, Inc. (Reykjavik, Iceland) for their scientific support.

## Authors' contributions

MA: Conceptualization, Programming, and Software, Formal Analysis, Writing—Original Draft. NA: Formal Analysis, Writing—Editing. Hamed Moheimani: Writing—Review & Editing. ASZ: Data Cleaning. FH-E: Data cleaning. HL: Results confirmation. FA: Supervision, MSD: Supervision.

## Funding

All parts of this research work, design of the study, data collection, analysis, interpretation of data, and manuscript writing were funded by the Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran Iran. The funding body played no role in publication costs.

## Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to containing information that could compromise the privacy of research participants but are available from the corresponding author on reasonable request.

## Declarations

### Ethics approval and consent to participate

The local ethics committee approved this study at Research Institute for Endocrine Sciences; Shahid Beheshti University of Medical Sciences (Research Approval Code:98104 & Research Ethical Code: IR.SBMU.Endocrine.REC.1398.121). In this study, all participants provided written informed consent for participating in the study. The research has been performed in accordance with the Declaration of Helsinki.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Biostatistics, Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. <sup>2</sup>Biostatistics, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. <sup>3</sup>Department of Surgery, University of Pittsburgh, PA, USA. <sup>4</sup>Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. <sup>5</sup>Nutrition and Endocrine Research Centre, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. <sup>6</sup>Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Received: 7 January 2022 Accepted: 11 March 2022

Published online: 09 April 2022

## References

- Kassi E, Pervanidou P, Kaltsas G, Chrousos G. Metabolic syndrome: definitions and controversies. *BMC Med*. 2011;9(1):1–3. <https://doi.org/10.1186/1741-7015-9-48>.
- Cornier MA, Dabelea D, Hernandez TL, Lindstrom RC, Steig AJ, Stob NR, Van Pelt RE, Wang H, Eckel RH. The metabolic syndrome. *Endocr Rev*. 2008;29(7):777–822. <https://doi.org/10.1210/er.2008-0024>.
- Obeidat AA, Ahmad MN, Haddad FH, Azzeh FS. Alarming high prevalence of metabolic syndrome among Jordanian adults. *Pak J Med Sci*. 2015;31(6):1377. <https://doi.org/10.12669/pjms.316.7714>.
- Mehairi AE, Khouri AA, Naqbi MM, Muhairi SJ, Maskari FA, Nagelkerke N, Shah SM. Metabolic syndrome among Emirati adolescents: a school-based study. *PLoS ONE*. 2013;8(2): e56159. <https://doi.org/10.1371/journal.pone.0056159>.
- Nematy M, Ahmadvpour F, Rassouli ZB, Ardabili HM, Azimi-Nezhad M. A review on underlying differences in the prevalence of metabolic syndrome in the Middle East, Europe and North America. *J Mol Genet Med*. 2014;2(s1):019. <https://doi.org/10.4172/1747-0862.S1-019>.
- Shahbazian H, Latifi SM, Jalali MT, Shahbazian H, Amani R, Nikhoo A, Aleali AM. Metabolic syndrome and its correlated factors in an urban population in South West of Iran. *J Diabetes Metab Disord*. 2013;12(1):1–6. <https://doi.org/10.1186/2251-6581-12-11>.
- Karimi F, Jahandideh D, Dabbaghmanesh M, Fattahi M, RANJBAR OG. The prevalence of metabolic syndrome and its components among adults in a rural community, Fars, Iran. *Int Cardiovasc Res J*. 2015;9(2):94–99. <https://www.sid.ir/en/journal/ViewPaper.aspx?id=436592>.
- Frootan M, Mahdavi R, Moradi T, Mobasser M, Farrin N, Ostadrahimi A. Prevalence of metabolic syndrome in an elderly population of Tabriz. *Iran Endocrinol Metabol Syndr*. 2011;1:51.
- Warner JP, Leek JP, Intody S, Markham AF, Bonthron DT. Human glucokinase regulatory protein (GCKR): cDNA and genomic cloning, complete primary structure, and chromosomal localization. *Mamm Genome*. 1995;6(8):532–6. <https://doi.org/10.1007/BF00356171>.
- Veiga-da-Cunha M, Delplanque J, Gillain A, Bonthron DT, Boutin P, Van Schaftingen E, Froguel P. Mutations in the glucokinase regulatory protein gene in 2p23 in obese French caucasians. *Diabetologia*. 2003;46(5):704–11. <https://doi.org/10.1007/s00125-003-1083-y>.
- Shen H, Pollin TI, Damcott CM, McLenithan JC, Mitchell BD, Shuldiner AR. Glucokinase regulatory protein gene polymorphism affects post-prandial lipemic response in a dietary intervention study. *Hum Genet*. 2009;126(4):567. <https://doi.org/10.1007/s00439-009-0700-3>.
- Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007;316(5829):1331–6. <https://doi.org/10.4093/dmj.2014.38.5.375>.
- Weissglas-Volkov D, Aguilar-Salinas CA, Sinsheimer JS, Riba L, Huertas-Vazquez A, Ordoñez-Sánchez ML, Rodríguez-Guillen R, Cantor RM, Tusie-Luna T, Pajukanta P. Investigation of variants identified in caucasian genome-wide association studies for plasma high-density lipoprotein cholesterol and triglycerides levels in Mexican dyslipidemic study samples. *Circ Cardiovasc Genet*. 2010;3(1):31–8. <https://doi.org/10.1161/CIRCGENETICS.109.908004>.
- Orho-Melander M, Melander O, Guiducci C, Perez-Martinez P, Corella D, Roos C, Tewhey R, Rieder MJ, Hall J, Abecasis G, Tai ES. Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes*. 2008;57(11):3112–21. <https://doi.org/10.2337/db08-0516>.
- Sparso T, Andersen G, Nielsen T, Burgdorf KS, Gjesing AP, Nielsen AL, Albrechtsen A, Rasmussen SS, Jørgensen T, Borch-Johnsen K, Sandbaek A. The GCKR rs780094 polymorphism is associated with elevated fasting serum triacylglycerol, reduced fasting and OGTT-related insulinaemia, and reduced risk of type 2 diabetes. *Diabetologia*. 2008;51(1):70–5. <https://doi.org/10.1007/s00125-007-0865-z>.
- Tam CH, Ma RC, So WY, Wang Y, Lam VK, Germer S, Martin M, Chan JC, Ng MC. Interaction effect of genetic polymorphisms in glucokinase (GCK) and glucokinase regulatory protein (GCKR) on metabolic traits in healthy Chinese adults and adolescents. *Diabetes*. 2009;58(3):765–9. <https://doi.org/10.2337/db08-1277>.
- Onuma H, Tabara Y, Kawamoto R, Shimizu I, Kawamura R, Takata Y, Nishida W, Ohashi J, Miki T, Kohara K, Makino H. The GCKR rs780094 polymorphism is associated with susceptibility of type 2 diabetes, reduced fasting plasma glucose levels, increased triglycerides levels and lower HOMA-IR in Japanese population. *J Hum Genet*. 2010;55(9):600–4. <https://doi.org/10.1007/s00125-007-0865-z>.
- Bi M, Kao WH, Boerwinkle E, Hoogveen RC, Rasmussen-Torvik LJ, Astor BC, North KE, Coresh J, Köttgen A. Association of rs780094 in GCKR with metabolic traits and incident diabetes and cardiovascular disease: the ARIC Study. *PLoS ONE*. 2010;5(7): e11690. <https://doi.org/10.1371/journal.pone.0011690>.
- Lian J, Guo J, Chen Z, Jiang Q, Ye H, Huang X, Yang X, Ba Y, Zhou J, Duan S. Positive association between GCKR rs780093 polymorphism and coronary heart disease in the aged Han Chinese. *Dis Markers*. 2013;35(6):863–8. <https://doi.org/10.1155/2013/215407>.
- Worachartcheewan A, Nantasenamat C, Isaranakura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Identification of metabolic syndrome using decision tree analysis. *Diabetes Res Clin Pract*. 2010;90(1):e15–8. <https://doi.org/10.1016/j.diabres.2010.06.009>.
- Babič F, Majnarić L, Lukáčová A, Paralič J, Holzinger A. On patient's characteristics extraction for metabolic syndrome diagnosis: predictive modeling based on machine learning. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer Verlag; 2014. p. 118–32. Doi: [https://doi.org/10.1007/978-3-319-10265-8\\_11](https://doi.org/10.1007/978-3-319-10265-8_11).
- Lehmann C, Koenig T, Jelic V, Prichep L, John RE, Wahlund LO, Dodge Y, Dierks T. Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *J Neurosci Methods*. 2007;161(2):342–50. <https://doi.org/10.1016/j.jneumeth.2006.10.023>.
- Azizi F, Madjid M, Rahmani M, Emami H, Mirmiran P, Hadjipour R. Tehran Lipid and Glucose Study (TLGS): rationale and design. *Iran J Endocrinol Metab*. 2000;2(2):77–86.
- Azizi F. Tehran lipid and glucose study: a national legacy. *Int J Endocrinol Metab*. 2018;16(4 Suppl):84774. <https://doi.org/10.5812/ijem.84774>.
- Daneshpour MS, Fallah MS, Sedaghati-Khayat B, Guity K, Khalili D, Hedayati M, Ebrahimi A, Hajsheikhholeslami F, Mirmiran P, Ramezani Tehrani F,

- Momenan AA, Ghanbarian A, Amouzegar A, Amiri P, Azizi F. Rationale and design of a genetic study on cardiometabolic risk factors: protocol for the Tehran Cardiometabolic Genetic Study (TCGS). *JMIR Res Protoc*. 2017;6(2):e28. <https://doi.org/10.2196/resprot.6050>.
26. Daneshpour MS, Hedayati M, Sedaghati-Khayat B, Guity K, Zarkesh M, Akbarzadeh M, et al. Genetic Identification for non-communicable disease: findings from 20 years of the Tehran Lipid and Glucose Study. *Int J Endocrinol Metab*. 2018;16(4 Suppl):84744. <https://doi.org/10.5812/ijem.84744>.
  27. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, Fruchart JC, James WP, Loria CM, Smith SC Jr. Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity. *Circulation*. 2009;120(16):1640–5. <https://doi.org/10.1161/CIRCULATIONAHA.109.192644>.
  28. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002; 2(3): 18–22. <https://CRAN.R-project.org/doc/Rnews/>.
  29. Venables WN, Ripley BD. *Modern applied statistics with S-PLUS*. Springer Science & Business Media; 2013. <https://doi.org/10.1007/978-1-4757-3121-7>.
  30. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*. 2015;31(15):2595–7. <https://doi.org/10.1093/bioinformatics/btv153>.
  31. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team C. Package 'caret'. *R J*. 2020;20(223):7.
  32. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang CC, Lin CC. e1071: Misc functions of the Department of Statistics (e1071), TU Wien. R package version. 2014;1(3).
  33. R: The R project for statistical computing. [cited 2020 Dec 30]. <https://www.r-project.org/>
  34. Therneau T, Atkinson B, Ripley B. Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone. *Inst Stat Math*. 2015. <https://doi.org/10.1201/9781315139470>.
  35. Huberty CJ. Discriminant analysis. *Rev Educ Res*. 1975;45(4):543–98. <https://doi.org/10.3102/00346543045004543>.
  36. Song YY, Ying LU. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27(2):130. <https://doi.org/10.11919/j.issn.1002-0829.215044>.
  37. Hastie T, Tibshirani R, Friedman J. *Random forests*. In: *The Elements of statistical learning*. Springer series in statistics. Springer, New York, NY; 2009. Doi: [https://doi.org/10.1007/978-0-387-84858-7\\_15](https://doi.org/10.1007/978-0-387-84858-7_15).
  38. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. <https://doi.org/10.1023/A:1010950718922>.
  39. Genuer R, Poggi JM, Tuleau C. Random Forests: some methodological insights. *arXiv preprint arXiv:0811.3619*. 2008 Nov 21.
  40. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–97.
  41. Kuhn M, Johnson K. *Applied predictive modeling*. New York: Springer; 2013.
  42. Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatr*. 2007;96(5):644–7. <https://doi.org/10.1111/j.1651-2227.2006.00178.x>.
  43. Romero-Saldaña M, Fuentes-Jiménez FJ, Vaquero-Abellán M, Álvarez-Fernández C, Molina-Recio G, López-Miranda J. New non-invasive method for early detection of metabolic syndrome in the working population. *Eur J Cardiovasc Nurs*. 2016;15(7):549–58. <https://doi.org/10.1177/1474515115626622>.
  44. Zahedi AS, Sedaghati-Khayat B, Behnami S, Azizi F, Daneshpour MS. Associations of common polymorphisms in GCKR with metabolic syndrome. *Tehran Univ Med J*. 2018;76(7):459–68. <https://doi.org/10.1186/s13098-021-00637-4>.
  45. Mohás M, Kisfali P, Járomi L, Maász A, Fehér E, Csöngéi V, Polgár N, Sáfrány E, Cseh J, Sümegi K, Hettyésy K. GCKR gene functional variants in type 2 diabetes and metabolic syndrome: do the rare variants associate with increased carotid intima-media thickness? *Cardiovasc Diabetol*. 2010;9(1):1–7. <https://doi.org/10.1186/1475-2840-9-79>.
  46. Jamal S, Ali W, Nagpal P, Grover A, Grover S. Predicting phosphorylation sites using machine learning by integrating the sequence, structure, and functional information of proteins. *J Transl Med*. 2021;19(1):1–11. <https://doi.org/10.1186/s12967-021-02851-0>.
  47. Entezari-Maleki R, Rezaei A, Minaei-Bidgoli B. Comparison of classification methods based on the type of attributes and sample size. *J Convergence Inf Technol*. 2009;4(3):94–102. <https://doi.org/10.4156/JCIT.VOL4.ISSUE3.14>.
  48. de Edelenyi FS, Goumidi L, Bertrai S, Phillips C, MacManus R, Roche H, Planells R, Lairon D. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest. *Genes Nutr*. 2008;3(3):173–6. <https://doi.org/10.1007/s12263-008-0097-y>.
  49. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. *ScientificWorldJournal*. 2015;2015: 581501. <https://doi.org/10.1155/2015/581501>.
  50. Huang YC. The application of data mining to explore association rules between metabolic syndrome and lifestyles. *Heal Inf Manag J*. 2013;42(3):29–36. <https://doi.org/10.1177/183335831304200304>.
  51. Karimi-Alavijeh F, Jalili S, Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. *ARYA Atheroscler*. 2016;12:146–52.
  52. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Quantitative population-health relationship (QPHR) for assessing metabolic syndrome. *EXCLI J*. 2013;12:569.
  53. Kim TN, Kim JM, Won JC, Park MS, Lee SK, Yoon SH, Kim HR, Ko KS, Rhee BD. A decision tree-based approach for identifying urban-rural differences in metabolic syndrome risk factors in the adult Korean population. *J Endocrinol Invest*. 2012;35(9):847–52. <https://doi.org/10.3275/8235>.
  54. Miller B, Fridline M, Liu PY, Marino D. Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. *Comput Math Methods Med*. 2014;2014: 242717. <https://doi.org/10.1155/2014/242717>.
  55. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov*. 1998;2(2):121–67.
  56. Meyer D, Leisch F, Hornik K. The support vector machine under test. *Neurocomputing*. 2003;55(1–2):169–86. [https://doi.org/10.1016/S0925-2312\(03\)00431-4](https://doi.org/10.1016/S0925-2312(03)00431-4).
  57. Smith A, Sterba-Boatwright B, Mott J. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Res*. 2010;44(14):4067–76. <https://doi.org/10.1016/j.watres.2010.05.019>.
  58. Statnikov A, Aliferis CF. Are Random Forests better than support vector machines for microarray-based cancer classification? *AMIA Annu Symp Proc*. 2007;11(2007):686–90.
  59. Lawson CE, Marti JM, Radivojevic T, Jonnalagadda SVR, Gentz R, Hillson NJ, et al. Machine learning for metabolic engineering: A review. *Metab Eng*. 2021;1(63):34–60. <https://doi.org/10.1016/j.jymben.2020.10.005>.
  60. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. *Nat Rev Methods Prim* 2021;1(1):1–21. <https://www.nature.com/articles/s43586-021-00056-9>. Doi: <https://doi.org/10.1038/s43586-021-00056-9>
  61. Lanjanian H, Najd Hassan Bonab L, Akbarzadeh M, Moazzam-Jazi M, Zahedi AS, Masjoudi S, et al. Sex, age, and ethnic dependency of lipoprotein variants as the risk factors of ischemic heart disease: a detailed study on the different age-classes and genders in Tehran Cardiometabolic Genetic Study (TCGS). *Biol Sex Differ*. 2022. <https://doi.org/10.1186/s13293-022-00413-7>.

## Publisher's Note

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