scientific reports

Check for updates

OPEN

Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms

Fariba Asadi[1], Reza Homayounfar[2], Yaser Mehrali[3], Chiara Masci[4], Samaneh Talebi[1] & Farid Zayeri[5]✉

Cardiovascular disease (CVD) can often lead to serious consequences such as death or disability. This study aims to identify a tree-based machine learning method with the best performance criteria for the detection of CVD. This study analyzed data collected from 9,499 participants, with a focus on 38 different variables. The target variable was the presence of cardiovascular disease (CVD) and the villages were considered as the cluster variable. The standard tree, random forest, Generalized Linear Mixed Model tree (GLMM tree), and Generalized Mixed Effect random forest (GMERF) were fitted to the data and the estimated prediction power indices were compared to identify the best approach. According to the analysis of important variables in all models, five variables (age, LDL, history of cardiac disease in first-degree relatives, physical activity level, and presence of hypertension) were identified as the most influential in predicting CVD. Fitting the decision tree, random forest, GLMM tree, and GMERF, respectively, resulted in an area under the ROC curve of 0.56, 0.73, 0.78, and 0.80. The GMERF model demonstrated the best predictive performance among the fitted models based on evaluation criteria. Regarding the clustered structure of the data, using relevant machine-learning approaches that account for this clustering may result in more accurate predicting indices and targeted prevention frameworks.

Keywords Cardiovascular disease, Machine learning, GMERF, GLMM Tree, Clustering data

Cardiovascular disease (CVD) can often lead to serious consequences such as death or disability[1]. According to the published reports, CVD is responsible for about 30% of all deaths globally, while it accounts for 46% of deaths in our country, Iran.[2] If the current trend continues, it is predicted that approximately 23.6 million people will die from this disease in 2030[3]. Furthermore, The economic costs of CVD are estimated to be around four hundred billion dollars throughout the world[4]. Similar to other non-communicable diseases, various factors influence the risk of cardiovascular disease[5]. Apparently, some of these factors such as age, gender, and inheritance, are beyond the control of the individuals, while factors like smoking, obesity, physical inactivity as well as higher levels of blood pressure, blood sugar, blood fat, and stress could be considered as the factors which can be managed by the individuals[5–8].

In recent decades, data analysts have employed various statistical techniques to predict and identify effective indicators of non-communicable diseases[9,10]. In healthcare research, well-established statistical methods such as logistic regression (LR), Fisher's discriminant analysis[11], and the area under the curve (AUC) method have been extensively utilized to identify factors linked to different diseases and categorize subjects accordingly. However, the landscape of these statistical methods has undergone a significant transformation, driven by the exponential growth of data and enhanced access to detailed medical information[6].

Machine learning (ML) algorithms, which fall under the broader field of artificial intelligence, employ different statistical, probabilistic, and optimization techniques to learn from previous experiences and identify helpful patterns in unstructured, complex, and big data sets[12]. Nowadays, there are numerous machine learning techniques available, such as support vector machines (SVM)[13], decision trees (DT)[14], random forests (RF)[15], and K-nearest neighbor (KNN)[16]algorithms, each with its unique advantages and limitations making them suitable for different types of problems and datasets[17]. The choice of the most appropriate machine learning

[1]Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [2]Food Technology Research Institute, Faculty of Nutrition Sciences and Food Technology, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [3]Statistical Center of Iran, Tehran, Iran. [4]MOX—Department of Mathematics, Politecnico Di Milano, Milan, Italy. [5]Proteomics Research Center, Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Qods Square, Darband Street, Tehran, Iran. ✉email: fzayeri@gmail.com

method often depends on the specific requirements of the task at hand, the characteristics of the data, and the desired model performance[18]. Therefore, Ml techniques are continuously evolving and improving over time, which makes them more sophisticated and effective, with researchers and data analysts working to enhance their performance[19].

The diagnosis of CVD in its initial stages and providing timely treatment remains a major challenge for cardiologists[20]. Detecting cardiovascular disease at early stages can provide the opportunity to treat patients and save numerous lives[21]. ML methods can assist cardiologists in predicting diseases at an early stage, allowing them to provide appropriate treatment for patients[17]. Literature review in this field shows that numerous studies have previously explored the application of machine learning in predicting cardiovascular disease, for instance, studies conducted by Marbaniang et al.[22], Baghdadi et al.[23], Swathy et al.[24], and Pal et al.[25]. In the study by Subramani et al.[26] seven different machine learning models were evaluated—logistic regression (LR), RF, DT, KNN, naive Bayes (NB), SVM, and extreme gradient boosting (XGB). The evaluation was performed on a dataset consisting of 1,190 samples. The results showed that for predicting the risk of CVD, the penalized logistic regression model outperformed the other machine learning approaches, similar to the performance of the SVM model. The KNN and RF models were ranked the next best. Ogunpola et al.[27] evaluated seven traditional machine learning models—KNN, SVM, LR, Convolutional Neural Network (CNN), Gradient Boost, XGBoost, and RF—on two datasets: the Cardiovascular Heart Disease Dataset and the Heart Disease Cleveland Dataset, each containing 1000 samples. When compared, the XGBoost model demonstrated the best performance, achieving an accuracy of 98.50% and a precision of 99.14%, outperforming the other six machine learning approaches. The study conducted by Uddin et al.[28] employed a variety of ML algorithms to detect the presence of cardiac abnormalities. The algorithms utilized in the study included DT, AdaBoost Classifier (AB), Extra Trees Classifier (ET), SVM, Gradient Boost, MLP, XGB, RF, KNN, and LR. The researchers combined three different datasets with training and testing. The experimental results showed that, among the tested algorithms, the Decision Tree algorithm achieved the highest accuracy at 99.16%, outperforming the other ML techniques.

Typically, traditional statistical models such as linear models—the term "traditional models" refers to well-established models—generally assume a linear relationship between the risk factors and outcomes[29,30]. In addition, these models tend to oversimplify complex relationships, which can include nonlinear interactions[29,31,32]. In contrast, machine learning (ML) models do not carry assumptions as traditional statistical models, like linear relationship or normally distributed residuals[33]This property enables the data analysts to thoroughly and effectively examine and analyze extensive amounts of data with numerous variables, and even medical images and signals[34]. The sole fundamental assumption in these machine learning models is the independence of the data[35]. The research studies that have been carried out up to this point have not placed sufficient emphasis or focus on examining and understanding the correlation between the data[36–38]. When data are collected over time (longitudinally) or as clusters, many ML methods may not be accurate or efficient due to the correlation within such data structures[35,39]. To resolve the challenge of correlated data, the integration of statistical and machine learning methods has recently been suggested . generalized mixed effects random forests[40] and generalized linear mixed-model tree (GLMM tree)[41] are some of these advanced techniques. These approaches combine the strengths of GLMMs to handle data with correlated structures and non-normal distributions[42], along with the ability of tree-based machine learning to discovery and pattern recognition in the data[43]. This integrated methodology empowers researchers, particularly in fields like medicine, to tackle complex questions that involve evaluating the effectiveness of treatments while accounting for variability among individual patients[44]. Studies in this field show that utilizing classification techniques that match the data structure and appropriately handle the correlation arising from repeated measurements can lead to improved predictive performance remarkably[35,45].

Regarding the significant burden of CVD in our country, Iran, and considering the clustered (correlated) nature of the available data from a cohort study in Fars Province, this study aims to utilize more complex statistical models to identify the most effective machine learning model for analyzing correlated data and rigorously evaluate the chosen model's predictive capabilities for true classification of new patients at risk of CVD. By adopting more advanced modeling methodologies that are aligned with the data structure, researchers and policymakers will be better equipped to analyze the CVD landscape, identify critical patterns and trends, and inform the development of targeted strategies to address this major public health challenge.

Materials and methods

Data source

The data used in this study were from the baseline phase of the Fasa Cohort Study (FACS)[46]. The FACS was designed to examine and evaluate the health conditions and risk factors that contribute to the increased vulnerability of rural inhabitants to non-communicable diseases (NCDs) in Fasa, located in Fars Province, southern parts of Iran. The cohort consisted of 10,118 individuals aged 35 to 70 from 29 villages in the Sheshdeh and Qara Balag areas (suburbs of Fasa City). The FACS was conducted in accordance with the standards of the ethics committee of Fasa University of Medical Sciences. Informed consent was obtained from all participants before they took part in the study. More detailed information about the design of the FACS and its participants can be found elsewhere[47]. In this study, the villages (rural regions) were considered as clusters and incorporated into the model as a random intercept.

Main outcome and potential predictors

In this study, 38 different variables were considered as the potential effective factors of CVD including demographic variables (such as age, gender, and marital status), lifestyle habits (including night sleep hour, morning wake-up hour,and dietary energy intake(Energy)), socio-economic status (wealth score index (WSI)), anthropometric measurements (BMI, waist-to-height ratio), metabolic equivalent of task (MET), dietary patterns (Dietary Inflammatory Index(DII)), blood biomarkers (LDL, HDLC, PLT, SGOT, SGPT, ALP, GGT,

cholesterol, glucose, triglyceride (TG)), and medical history (having diabetes, hypertension, thyroid problems, chronic headaches, depression, fatty liver) as well as family health history, smoking, alcohol consumption, and tobacco use. In addition, the main outcome variable was defined as the presence of cardiovascular disease (CVD) at the baseline phase of the FACS study. Subjects with heart failure (HF) or ischemic heart disease (IHD) were classified as those with CVD. In this study, villages were considered as a cluster variable or, in other words, a random effect.

Statistical analysis

First, individuals with 50% or more missing data in the input variables removed and the analysis focused on a dataset comprising 9499 participants. In this dataset, the proportion of missing data for each variable was less than 1%, and these missing values were filled in through a single imputation (mean imputation for numerical features and mode imputation for categorical features) process. Then the Z-score normalization was applied to the numerical variables. In order to evaluate the performance of the machine learning models, the data was divided into 80% training set and 20% test set. This data split was performed in a stratified manner, ensuring the distribution of the outcome variable was preserved across the training and test sets (the proportion of CVD in the total data was 0.11).

Next, we employed the 'Boruta' R package to address the issue of feature multicollinearity by selecting a final set of features that were most relevant to the subsequent predictive modeling, with the aim of enhancing the performance of the machine learning (ML) approaches. Boruta implements a feature selection method based on random forests. It identifies relevant features by comparing the importance of original attributes to the importance that could be achieved randomly, estimated using their permuted copies (shadows). It iteratively eliminates features that a statistical test shows to be less relevant than random probes[48].

Then, four different tree-based ML models including RF, decision tree (DT), GLMM tree, and GLMM RF were fitted to the data using functions randomForest, rpart, glmertree, and gmerf, respectively. A random intercept was calculated for villages, utilizing Default settings of restricted maximum likelihood (REML) estimation for the GLMM tree and GLMM RF. The tree-based methods have the capability to model complex, high-order interactions between the input variables. Finally, to compare the ability of these models to predict the disease, we assessed and compared predictive criteria such as accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve.

It is noteworthy to mention that the ten-fold cross-validation method was also used to estimate the validation parameters (based on the algorithm type) such as the number of trees (the total count of decision trees in Random Forest[49]), maximum tree depth—a deeper tree can capture more complex patterns but may lead to overfitting[50]-, number of features at each split (The number of features considered when making a split at each node in the tree[49]) and evaluate the predictive performance of each algorithm. Default settings were used for other parameters. Prediction error was quantified as the mean absolute difference between predicted and observed response variable values (predictive misclassification rate (PMCR)). Machine learning predictive models were developed using the features outlined in Supplementary Table S1 online. The K-fold cross-validation (CV) procedure can help us to prevent overfitting and artificially inflated validation metrics[51]. All statistical analyses were conducted utilizing the RStudio software (version 2023.06.0). The default significance level (α) for testing parameter instability across all trees was established at 0.05. Figure 1 illustrates the overall framework of the study design.

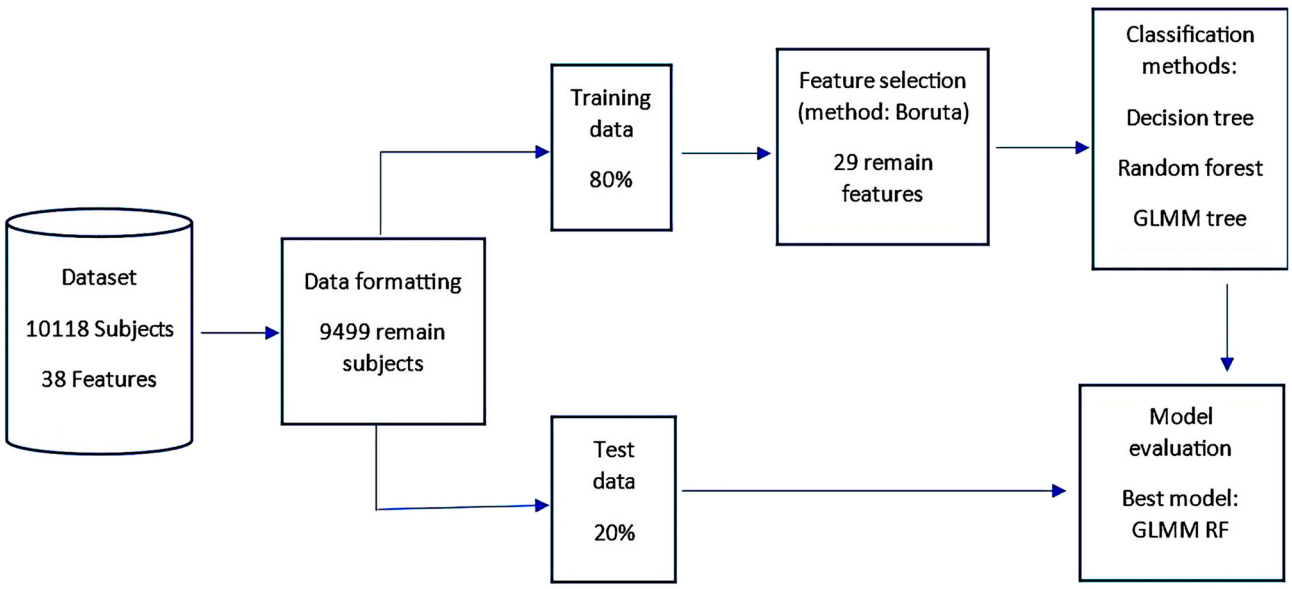

Fig. 1. The framework of the study design.

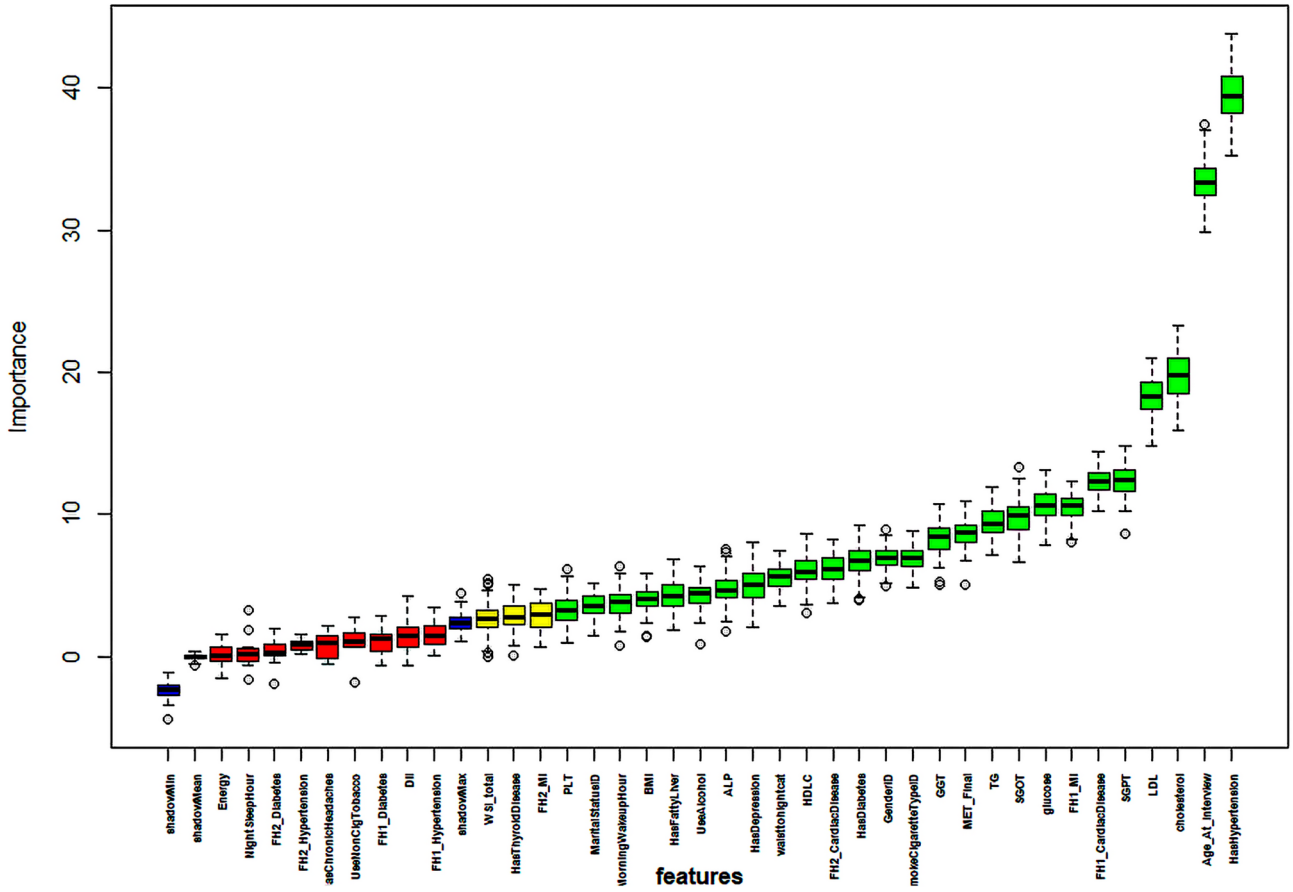

Fig. 2. Feature selection using the Boruta algorithm.

Model	Sensitivity	Specificity	Accuracy (0.95%CI)	AUC (0.95%CI)
Tree	0.83	0.59	0.81 (0.79,0.83)	0.56 (0.54,0.59)
RF	0.70	0.75	0.70(0.68,0.72)	0.73 (0.69,0.76)
GLMM Tree	0.74	0.73	0.74 (0.72,0.76)	0.78 (0.75,0.82)
GMERF	0.75	0.72	0.75 (0.73,0.77)	0.80 (0.77,0.84)

Table 1. Comparison of Predictive Power Indices for RF, Decision Tree, GLMMRF, and GLMM Tree Models.

Results

In this study, the data from a total of 9,499 individuals, including 4199 (44.20%) male and 5300 (55.80%.) female participants, was analyzed. The mean ± SD age of the participants was 48.95 ± 9.47 years. Table S1 online shows more details about the characteristics of the participants in the presence of CVD.

In the first step, we used the 'Boruta' R package to select the final set of features for predicting CVD. The obtained findings are presented in Fig. 2. Based on this figure, the following features were removed in the first stage: dietary energy intake, night sleep duration, history of diabetes and hypertension in second-degree relatives (FH2_diabetes and FH2_hypertension), history of diabetes and hypertension in first-degree relatives (FH1_diabetes and FH1_hypertension), smoke use, DII, and chronic headaches. Then, the remaining features were entered into the random forest model. Three features WSI, FH2_MI, and thyroid problems in Fig. 2 were identified as tentative features, which were removed one by one from the final model to improve the decision-making process and enhance the predictive power of the model. The predictive power indices were then calculated each time a feature was removed. Ultimately, all three features were found to cause an increase in the predictive power indices, and therefore, they were retained in the final model. Consequently, the final model includes 29 features such as age, gender, marital status, WSI, BMI, waist-to-height ratio, morning wake-up hour, MET, LDL, HDLC, SGOT, SGPT, ALP, GGT, PLT, alcohol consumption, smoking, cholesterol, glucose, triglyceride, medical history (history of diabetes, hypertension, thyroid problems, depression, and fatty liver), as well as the history of cardiac disease and MI in first and second-degree relatives.

Comparing the estimated predictive power indices in Table 1 shows that the gmerf model has resulted in the best performance among the evaluated models (AUC (0.95% CI) = 0.80 (0.77,0.84)). The glmm tree model was

ranked second, with a very small difference in performance compared to the gmerf model. While the tree model had the highest accuracy equal to 0.81, there was an imbalance between its sensitivity and specificity, so this model showed the lowest specificity value (0.59) among the other models.

To compare the estimated predictive criteria of the fitted models more visually, we also evaluated the power of these models for predicting the outcome (presence of CVD) using the ROC curve. As shown in Fig. 3, the tree model had the lowest value of the area under ROC curve (with AUC of 0.55), while the gmerf model had the highest AUC value of 0.80. Furthermore, the glmm tree model outperformed the random forest (RF) model.

According to the output of important variables displayed in Fig. 4, one can conclude that the five variables identified as the most important indicators of CVD were age, LDL, family history of cardiac disease in first-degree relatives (FH1-cardiac disease), physical activity level (MET-Final), and presence of hypertension. In contrast, the variables that were recognized as less important in at least two of the models were alcohol use, waist-to-height ratio, family history of myocardial infarction in second-degree relatives (FH2-MI), presence of thyroid disease, gender, and presence of fatty liver.

Finally, the features of hypertension and age were recognized as the variables with the most importance in the GLMM tree model. After pruning the tree model, the variables that were identified as less important, similar to other models, were alcohol use, waist-to-height ratio, family history of myocardial infarction in second-degree relatives (FH2-MI), presence of thyroid disease, gender, and presence of fatty liver. These less important variables were ultimately removed from the final tree model (Fig. 5). Based on Fig. 5, patients can be classified into one of the terminal nodes according to their characteristics. For instance, terminal node 28 shows that individuals with the criteria "Hypertension = 1", "Age < 49.99", or "standard age < 0.11", the observed proportion of cases with CVD is approximately 0.20.

Discussion

Regarding the high burden and huge cost of CVD imposed on all communities in recent decades, the effective prediction and management of cardiovascular disease is of paramount importance in improving public health outcomes. Timely diagnosis of the condition can significantly improve patient outcomes by saving lives, minimizing disease complications, and reducing mortality rates. In this context, machine learning models have emerged as valuable tools that can facilitate the early detection and prediction of various cardiac disorders. In this study, we explored the use of machine learning techniques to develop robust predictive models for this disease. Predictive power indices enable a systematic and data-driven comparison of the results obtained from different models. This information can guide us to select the most appropriate modeling approach for a given problem or dataset. In the present study, the results from comparing the predictive power indices showed that the gmerf (generalized mixed effect random forest) model exhibited the best performance among the evaluated models. The GLMM tree was ranked as the second best performing model with results very close to the GMERF results, coming before the RF and DT models in terms of its ranking or performance. In a study by Bernaich, the researcher also compared three statistical models—generalized linear mixed-effects models (GLMMs), GLMM trees, and RF—to analyze the occurrence of filled and unfilled pauses across different varieties of English. They found that GLMMs and GLMM trees performed similarly in predicting pause length, while RF was less accurate in this field[52]. This aligns with our results and indicates that the GLMM tree outperforms the RF. This finding is also aligned with the previously reported results in another study by Asadi et al.[6]. When we have multiple observations or measurements taken from the same source or subject, mixed-effects models are able to explicitly account for and systematically remove the effects of the relatedness or non-independence of the data. These types of machine learning models are designed to handle the fact that observations from the same source are likely to be more similar to each other compared to those from different sources. By accounting for

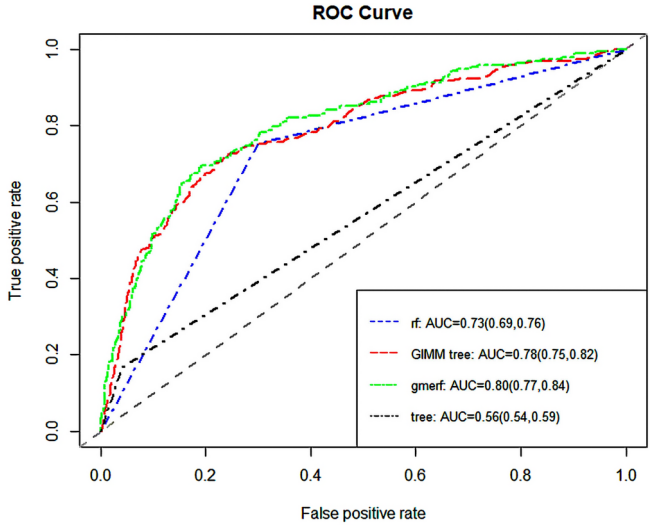

Fig. 3. Receiver Operating Characteristic (ROC) Curve for Different Machine Learning Models.

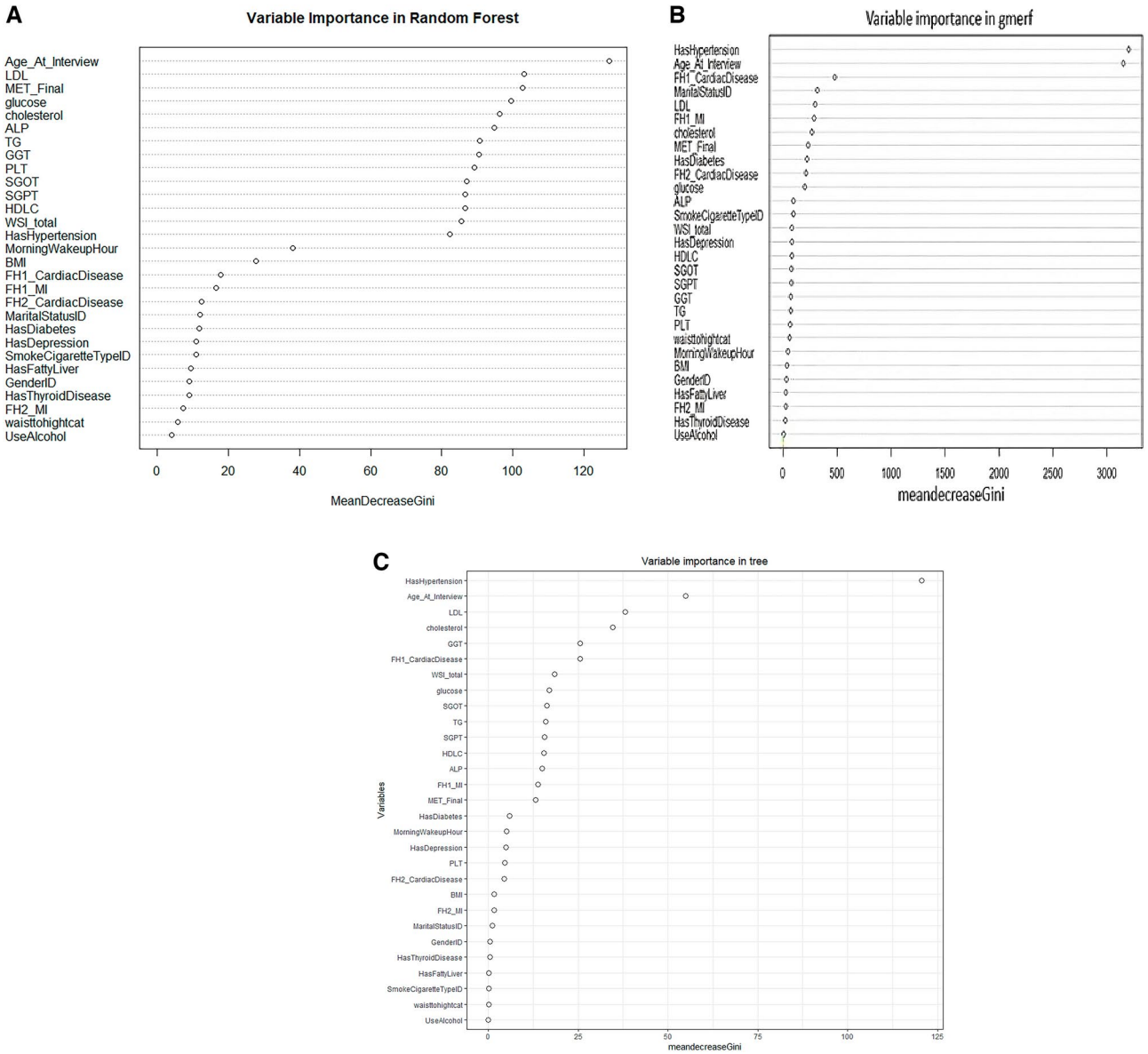

Fig. 4. Importance index of variables: (**A**): Random Forest Model, (**B**): Gmerf Model, (**C**): Tree Model.

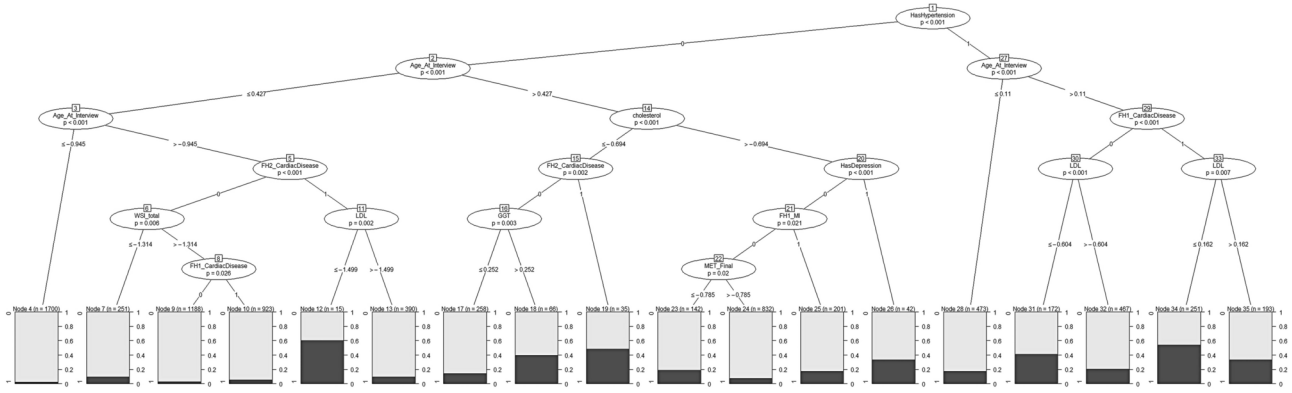

Fig. 5. GLMM tree for predicting cardiovascular disease (CVD), where black and gray nodes represent individuals with and without CVD.

this relatedness, mixed-effects models can provide more accurate and reliable results compared to models that do not consider the non-independent nature of the data[53–55]. Similarly, the strong performance of mixed-effect machine learning techniques is consistent with the growing body of literature to account for hierarchical data structures and improve predictive accuracy[33,40,41].

Interestingly, the tree model had an imbalance between its sensitivity and specificity, with the tree model having the lowest specificity value (0.59) among the other models. This suggests that the DT approach may lead to a tendency to misclassify certain cases and subsequently a relatively lower ability to correctly identify negative instances. This discrepancy between the accuracy and the balance of sensitivity and specificity is important when selecting the optimal model for a given application. The emphasis on evaluating not only the overall accuracy but also the balance of sensitivity and specificity underscores the importance of comprehensive model assessment to ensure the selection of the most appropriate tool for the available data. Furthermore, the tree model exhibited the lowest area under the ROC curve, with an AUC value of 0.55. This suggests that the tree-based approach had relatively weaker discriminative power in differentiating between positive and negative cases compared to the other models. In contrast, the gmerf model demonstrated the highest AUC value of 0.80. This indicates that the gmerf model had the strongest predictive capability and was able to most effectively capture the underlying patterns in the data to accurately classify the study participants.

The findings from the present study indicated that features of age, LDL, family history of cardiac disease, physical activity level, and presence of hypertension are the most influential predictors of CVD.. In 2019, Alizadehsani et al. conducted a review study on the diagnosis of coronary artery disease using machine learning techniques with 67 different datasets. The findings from this study showed that variables of age, type of chest pain, gender, total cholesterol, ST depression, hypertension, and maximum heart rate were the common effective predictors of CAD in machine learning algorithms. The algorithms that utilized these key features achieved high accuracy in predicting the risk of cardiovascular disease[56]. In 2017, Weng et al. explored whether machine learning could improve cardiovascular risk prediction using routine clinical data. The study identified the variables age, gender, and smoking as top-ranked risk indicators in all four machine-learning algorithms including LR, RF, Gradient Boosting, and ANN[57]. It seems that the discrepancy between our study and the study conducted by Weng could be attributed to the differences in the input features in the machine learning algorithms.

Our findings indicated that LDL is a significant indicator in CVD, aligning with prior research in this field[25,58,59]. In 2022, Pal et al. used a dataset from the University of California Irvine Repository to develop machine-learning models for detecting CVD. This dataset consisted of 303 patients between the ages of 29 and 77 years. The researchers indicated that increasing a person's cholesterol level leads to a rise in blood pressure, and subsequently the rise in blood pressure increases the risk of CVD[25]. In another study by Shuaib M. Abdullah et al., involving a low 10-year risk cohort of 36,375 participants, the results indicated that LDL-C and non–HDL-C levels of $\geq$ 160 mg/dL were independently associated with a 50% to 80% increase in the relative risk of cardiovascular disease mortality[60]. These results underscore the importance of monitoring and managing LDL levels to mitigate cardiovascular risk. Age and blood pressure were also significant factors of CVD in our research. The AHA 2019 Heart Disease and Stroke Statistical Update reported that the incidence of CVD among individuals aged 60 to 79 years was 77.2% for males and 78.2% for females[61]. In 2020, a prospective study in China, with 71,245 participants found that as the age of hypertension onset increased, the risks of adverse outcomes gradually decreased. For those under 45, the hazard ratio for CVD was 2.26 (1.19 to 4.30); for ages 45 to 54, the ratios were 1.62 (1.24 to 2.12)[59]. In another study by Fuchs and Whelton, the authors noted that adults without CVD, who at lower systolic (120–139 mmHg) and diastolic (80–89 mmHg) levels are at more heightened risk of developing higher blood pressure during short follow-up periods[62]. While age is an uncontrollable factor, focusing on early blood pressure detection and ongoing monitoring can help address high blood pressure challenges and lower cardiovascular disease prevalence across all age groups. In addition to, both family history of heart disease in first-degree relatives and the level of physical activity were important factors associated with the occurrence of CVD. The 2023 study with a follow-up period of 15 years, found that having a parental history of cardiovascular disease (CVD) increased an individual's risk of developing future CVD by 1.7 times among offspring[63]. On the other hand, a meta-analysis of 33 studies found that adults engaging in physical activity at least half the level recommended by the "2008 Physical Activity Guidelines for Americans" had a 14% lower risk of coronary heart disease[64]. According to the ACC/AHA guidelines, even if adults are unable to meet the recommended levels of physical activity, any amount of moderate-to-vigorous exercise can still reduce their cardiovascular disease risk[65]. These findings emphasize the importance of a holistic prevention strategy that takes into account both genetic and lifestyle factors for the effective prevention and management of CVD.

The findings from the final GLMM tree also indicate that several key variables contribute to the risk score for CVD. These variables include blood pressure, age, LDL cholesterol levels, family history of myocardial infarction (MI) and cardiac disease, as well as MET (Metabolic Equivalent of Task), and WSI (Weight Status Index). Many of these factors are also commonly recognized in traditional cardiovascular risk assessment studies, such as the Framingham[66], CUORE project[67], and Reynolds Risk Score[68]. The unique capability of the Generalized Linear Mixed Model (GLMM) tree to account for complex interactions and nonlinear relationships among these variables provides a more detailed and nuanced understanding of heart disease within the FACS population.

One of the key strengths of the current work is utilizing a comprehensive approach for feature selection and model development. In this context, we tried to explore a wide range of potential risk factors, including demographic characteristics, clinical biomarkers, lifestyle factors, and family history to predict CVD more accurately. By leveraging the power of machine learning algorithms, the relative importance of these features was systematically evaluated in predicting CVD. Our findings in the present study can help healthcare professionals to focus on targeted screening, risk stratification, and personalized preventive strategies. Furthermore, the development of accurate machine learning-based predictive models will be a supplement to the clinical portfolio,

enhancing human-led decision-making and clinical practices and reducing the financial burden on patients and the healthcare system.

Data availability

The data sets used in this study are not publicly accessible due to valid privacy and security reasons. Nevertheless, the original raw data can be obtained by contacting the corresponding author upon a reasonable request.

Received: 19 July 2024; Accepted: 10 September 2024
Published online: 27 September 2024

References

1. Joseph, P. *et al.* Reducing the Global Burden of Cardiovascular Disease, Part 1. *Circ. Res.* **121**, 677–694. https://doi.org/10.1161/CIRCRESAHA.117.308903 (2017).
2. Sooki, Z., Sharifi, K., Tagharrobi, Z. & Nematian, F. The effect of cognitive - behavioral intervention therapy on anxiety of cardiovascular patients: A systematic review and meta-analysis study. *Feyz Med. Sci. J.* **24**, 462–472 (2020).
3. Hazavehei, S. M. M., Shahabadi, S. & Hashemi, S. Z. The role of health education in reducing cardiovascular diseases risk factors: a systematic review. *J. Knowl. Health* **9**, 30–42 (2014).
4. Shamsi, A. & Ebadi, A. Risk factors of cardiovascular diseases in elderly people. *Critical Care Nursing* **3**, 189–194 (2011).
5. Dizdarevic-Bostandzic, A. *et al.* Cardiovascular risk factors in patients with poorly controlled diabetes mellitus. *Med. Arch.* **72**, 13–16. https://doi.org/10.5455/medarh.2018.72.13-16 (2018).
6. 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**, 239–247. https://doi.org/10.34172/aim.2024.35 (2024).
7. Koolaji, S. *et al.* A 30-year trend of ischemic heart disease burden in a developing country; a systematic analysis of the global burden of disease study 2019 in Iran. *Int. J. Cardiol.* **379**, 127–133. https://doi.org/10.1016/j.ijcard.2023.03.012 (2023).
8. Pepera, G., Tribali, M.-S., Batalik, L., Petrov, I. & Papathanasiou, J. Epidemiology, risk factors and prognosis of cardiovascular disease in the Coronavirus Disease 2019 (COVID-19) pandemic era: A systematic review. *Rev. Cardiovasc. Med.* **23**, 28. https://doi.org/10.31083/j.rcm2301028 (2022).
9. Kontis, V. *et al.* Contribution of six risk factors to achieving the 25×25 non-communicable disease mortality reduction target: A modelling study. *Lancet* **384**, 427–437. https://doi.org/10.1016/S0140-6736(14)60616-4 (2014).
10. Wang, Y. & Wang, J. Modelling and prediction of global non-communicable diseases. *BMC Public Health* **20**(822), 1–13. https://doi.org/10.1186/s12889-020-08890-4 (2020).
11. Johnson, R. A. & Wichern, D. W. *Applied multivariate statistical analysis*. (Upper saddle River, 2002).
12. Afshari, S. S., Enayatollahi, F., Xu, X. & Liang, X. Machine learning-based methods in structural reliability analysis: A review. *Reliabil. Eng. Syst. Safety*. https://doi.org/10.1016/j.ress.2021.108223 (2022).
13. Yan, H. *et al.* Least squares twin bounded support vector machines based on L1-norm distance metric for classification. *Pattern Recogn.* **74**, 434–447 (2018).
14. Aworski, M., Duda, P. & Rutkowski, L. New splitting criteria for decision trees in stationary data streams. *IEEE Trans. Neural Netw. Learn Syst.* **29**, 2516–2529 (2018).
15. Simon, S. M., Glaum, P. & Valdovinos, F. S. Interpreting random forest analysis of ecological models to move from prediction to explanation. *Sci. Rep.* **13**, 1–13. https://doi.org/10.1038/s41598-023-30313-8 (2023).
16. Zhang, S., Cheng, D., Deng, Z., Zong, M. & Deng, X. A novel K-NN algorithm with data driven k parameter computation. *Pattern Recogn. Lett.* **109**, 44–54 (2018).
17. Pal, M., Parija, S., Panda, G., Dhama, K. & Mohapatra, R. K. Risk prediction of cardiovascular disease using machine learning classifiers. *Open Med. (Wars).* **17**(1), 1100–1113. https://doi.org/10.1515/med-2022-0508.PMID:35799599;PMCID:PMC9206502 (2022).
18. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 1–21. https://doi.org/10.1007/s42979-021-00592-x (2021).
19. Jin, H., Zhang, E. & Espinosa, H. D. Recent advances and applications of machine learning in experimental solid mechanics: A review. *Appl. Mech. Rev.* https://doi.org/10.1115/1.4062966 (2023).
20. Wang, Y.-R. *et al.* Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. *Nat. Med.* **30**, 1471–1480. https://doi.org/10.1038/s41591-024-02971-2 (2024).
21. Celermajer, D. S., Chow, C. K., Marijon, E., Anstey, N. M. & Woo, K. S. Cardiovascular Disease in the Developing World. *J. Am. Coll. Cardiol.* **60**, 1207–1216. https://doi.org/10.1016/j.jacc.2012.03.074 (2012).
22. Marbaniang, I. A., Choudhury, N. A. & Moulik, S. *IEEE 17th India council international conference (INDICON).* 1–6 (IEEE). (2020).
23. Baghdadi, N. A. *et al.* Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *J. Big Data* **10**, 144 (2023).
24. Swathy, M. & Saruladha, K. A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *ICT Express* **8**, 109–116 (2022).
25. Pal, M., Parija, S., Panda, G., Dhama, K. & Mohapatra, R. K. Risk prediction of cardiovascular disease using machine learning classifiers. *Open Med.* **17**, 1100–1113 (2022).
26. Subramani, S. *et al.* Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Front. Med.* **10**, 1150933. https://doi.org/10.3389/fmed.2023.1150933 (2023).
27. Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M. & Qasem, S. N. Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics* **14**, 144. https://doi.org/10.3390/diagnostics14020144 (2024).
28. Uddin, K. M. M., Ripa, R., Yeasmin, N., Biswas, N. & Dey, S. K. Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset. *Intelligence-Based Med.* **7**, 100100. https://doi.org/10.1016/j.ibmed.2023.100100 (2023).
29. Ley, C. *et al.* Machine learning and conventional statistics: Making sense of the differences. *Knee Surg. Sports Traumatol. Arthrosc.* **30**, 753–757. https://doi.org/10.1007/s00167-022-06896-6 (2022).
30. Kim, H.-Y. Statistical notes for clinical researchers: simple linear regression 3–residual analysis. *Restorat. Dent. Endodon.* https://doi.org/10.5395/rde.2019.44.e26 (2019).
31. Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009).
32. Pal, S. C., Ruidas, D., Saha, A., Islam, A. R. M. T. & Chowdhuri, I. Application of novel data-mining technique-based nitrate concentration susceptibility prediction approach for coastal aquifers in India. *J. Cleaner Prod.* **346**, 131205 (2022).
33. 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**, 329–341. https://doi.org/10.1080/10503307.2020.1785037 (2021).
34. 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**, 1243 (2024).

35. Jianchang, H. & Silke, S. A review on longitudinal data analysis with random forest. *Briefings Bioinform.* https://doi.org/10.1093/bib/bbad002 (2023).
36. Gould, M. K., Huang, B. Z., Tammemagi, M. C., Kinar, Y. & Shiff, R. Machine learning for early lung cancer identification using routine clinical and laboratory data. *Am. J. Respir. Crit. Care Med.* **204**, 445–453. https://doi.org/10.1164/rccm.202007-2791OC (2021).
37. You, J., Guo, Y. & Kang, J. Development of machine learning-based models to predict 10-year risk of cardiovascular disease: A prospective cohort study. *Stroke Vasc. Neurol.* **8**, 475–485. https://doi.org/10.1136/svn-2023-002332 (2023).
38. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. & Van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One* **14**, e0213653 (2019).
39. Athey, S. *The economics of artificial intelligence: An agenda* (University of Chicago Press, 2019).
40. Pellagatti, M., Masci, C., Ieva, F. & Paganoni, A. M. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat. Anal.* **14**, 241–257. https://doi.org/10.1002/sam.11505 (2021).
41. Hajjem, A., Larocque, D. & Bellavance, F. Generalized mixed effects regression trees. *Statist. Probabil. Lett.* **126**, 114–118. https://doi.org/10.1016/j.spl.2017.02.033 (2017).
42. Salinas Ruíz, J., Montesinos López, O. A., Hernández Ramírez, G. & Crossa Hiriart, J. *Generalized Linear Mixed Models with Applications in Agriculture and Biology* (Springer, 2023).
43. Jiryaei Sharahi, Z., Zare Mehrjerdi, Y., Owlia, M. S. & Abessi, M. Machine learning decision tree based on regression in data mining to extract more knowledge. *J. Indus. Eng. Manag. Stud.* **9**, 86–112. https://doi.org/10.22116/jiems.2022.327172.1474 (2022).
44. 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).
45. Mangino, A. A. & Holmes, F. W. Prediction with mixed effects models: A Monte Carlo simulation study. *Educ. Psychol. Meas.* **81**(6), 1118–1142 (2021).
46. Homayounfar, R. *et al.* Cohort Profile: The Fasa Adults Cohort Study (FACS): A prospective study of non-communicable diseases risks. *Int. J. Epidemiol.* **52**, e172–e178. https://doi.org/10.1093/ije/dyac241 (2023).
47. Farjam, M. *et al.* A cohort study protocol to analyze the predisposing factors to common chronic non-communicable diseases in rural areas: Fasa Cohort Study. *BMC Public Health* **16**, 1–8. https://doi.org/10.1186/s12889-016-3760-z (2016).
48. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Statist. Softw.* https://doi.org/10.18637/jss.v036.i11 (2010).
49. Rezaei, N. & Jabbari, P. *Immunoinformatics of Cancers, Practical Machine Learning Approaches Using R* (eds Nima Rezaei & Parnian Jabbari) Ch. 11, 169–179 (2022).
50. Duroux, R. & Scornet, E. Impact of subsampling and tree depth on random forests. *ESAIM PS* **22**, 96–128 (2018).
51. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
52. Bernaisch, T. Comparing generalised linear mixed effects models, generalised linear mixed-effects model trees and random forests. In *Data and methods in corpus linguistics: Comparative approaches* (ed. Bernaisch, T.) (Cambridge University Press, 2022).
53. Fokkema, M. & Zeileis, A. Subgroup detection in linear growth curve models with generalized linear mixed model (GLMM) trees. *Behav. Res. Methods*. https://doi.org/10.48550/arXiv.2309.05862 (2023).
54. Moscatelli, A., Mezzetti, M. & Lacquaniti, F. Modeling psychophysical data at the population-level: The generalized linear mixed model. *J. Vis.* **12**, 26–26 (2012).
55. Fallahzadeh, H. & Asadi, F. Generalized linear mixed models: Introduction, estimation methods and their application in medical. *Studies* **14**, 33–39 (2019).
56. Alizadehsani, R. *et al.* Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Comput. Biol. Med.* **111**, 1–14 (2019).
57. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS One* **12**, e0174944 (2017).
58. Rodgers, J. L. *et al.* Cardiovascular risks associated with gender and aging. *J. Cardiovasc. Dev. Dis.* **6**, 19. https://doi.org/10.3390/jcdd6020019 (2019).
59. Wang, C. *et al.* Association of age of onset of hypertension with cardiovascular diseases and mortality. *J. Am. Coll. Cardiol.* **75**, 2921–2930. https://doi.org/10.1016/j.jacc.2020.04.038 (2020).
60. Abdullah, S. M. *et al.* Long-term association of low-density lipoprotein cholesterol with cardiovascular mortality in individuals at low 10-year risk of atherosclerotic cardiovascular disease. *Circulation* **138**, 2315–2325. https://doi.org/10.1161/CIRCULATIONAHA.118.034273 (2018).
61. Benjamin, E. J. *et al.* Heart Disease and Stroke Statistics—2019 Update: A report from the American heart association. *Circulation* **139**, e56–e528. https://doi.org/10.1161/CIR.0000000000000659 (2019).
62. Fuchs, F. D. & Whelton, P. K. High blood pressure and cardiovascular disease. *Hypertension* **75**, 285–292. https://doi.org/10.1161/HYPERTENSIONAHA.119.14240 (2020).
63. Taylor, C. N. *et al.* Family history of modifiable risk factors and association with future cardiovascular disease. *J. Am. Heart Assoc.* https://doi.org/10.1161/JAHA.122.027881 (2023).
64. Sattelmair, J. *et al.* Dose response between physical activity and risk of coronary heart disease: A meta-analysis. *Circulation* **124**, 789–795. https://doi.org/10.1161/CIRCULATIONAHA.110.010710 (2011).
65. Arnett, D. K. *et al.* 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation* **140**, e596–e646. https://doi.org/10.1161/CIR.0000000000000678 (2019).
66. D'Agostino, R. B. *et al.* General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation* https://doi.org/10.1161/CIRCULATIONAHA.107.699579 (2008).
67. Palmieri, L. *et al.* CUORE project: implementation of the 10-year risk score *Eur. J. Cardiovasc. Prev. Rehabil.* **18**, 642–649. https://doi.org/10.1177/1741826710389925 (2011).
68. Ridker, P. M., Buring, J. E., Rifai, N. & Cook, N. R. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score. *JAMA* **297**, 611–619. https://doi.org/10.1001/jama.297.6.611 (2007).

Acknowledgements

The authors express their gratitude towards all individuals who patiently contributed to this study.

Author contributions

F.A.: Supervision, methodology, investigation, conceptualization, validation, data curation,formal analysis, writing—original draft preparation. F.Z.: Supervision, conceptualization, validation, reviewing and editing. reviewing. Y.M. and CH.M.: Methodology, conceptualization. R.H and S.T.: Conceptualization, data curation, reviewing and editing.

Competing interests
The authors declare no competing interests.

Ethical approval
This study was approved by the Ethics Committee of Shahid Beheshti University of Medical Sciences (approval number: IR.SBMU.RETECH.REC.1402.137).

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-72819-9.

Correspondence and requests for materials should be addressed to F.Z.

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.