Evaluating Feature Selection Methods and Feature Contributions for Cardiovascular Disease Risk Prediction

Suraiya Akhter[1*], John H. Miller[2]

[1] School of Business and Technology, Emporia State University, Emporia, Kansas, USA

[2] School of Engineering and Applied Sciences, Washington State University, Richland, Washington, USA

[*]Corresponding author

Email: sakhter1@emporia.edu

Abstract

Cardiovascular disease (CVD) remains the foremost contributor to global illness and death, underscoring the critical need for effective tools that can predict risk at early stages to support preventive care and timely clinical decisions. With the growing complexity of healthcare data, machine learning has shown considerable promise in extracting insights that enhance medical decision-making. Nonetheless, the effectiveness and clarity of machine learning models largely rely on the relevance and quality of input features. In this work, we explored and compared three distinct feature selection strategies—Alternating Decision Tree (ADT)-based analysis, Cross-Validated Feature Evaluation (CVFE), and Hypergraph-Based Feature Evaluation (HFE)—to isolate the most predictive clinical variables for assessing CVD risk. Our analysis utilized data from the National Health and Nutrition Examination Survey (NHANES), administered by the National Center for Health Statistics under the Centers for Disease Control and Prevention (CDC), encompassing demographic, clinical, laboratory, and survey data collected across the U.S. from August 2021 through August 2023. Distinct sets of features obtained through the selection techniques were used to develop eXtreme Gradient Boosting (XGBoost) models, which were then assessed for predictive effectiveness. Among the approaches, the HFE method achieved the most accurate results, reaching 75% accuracy and an AUC of 0.7857, outperforming the alternatives. To improve clarity and understand the model's decision-making, SHapley Additive exPlanations (SHAP) was utilized to interpret the influence of each feature in the top-performing model. This study highlights the importance of strategic feature selection in refining predictive accuracy and interpretability, offering a practical data-centric approach that could aid clinicians in evaluating cardiovascular risk and tailoring preventive care.

Introduction

Globally, cardiovascular diseases (CVDs) remain the primary cause of death, responsible for around 17.9 million fatalities each year, which equates to nearly one-third of all deaths worldwide [1]. This broad category includes multiple heart and circulatory system disorders such as coronary artery disease, stroke, peripheral artery disease, rheumatic heart conditions, and congenital cardiovascular defects [2, 3]. Among these, coronary artery disease constitutes the majority, accounting for approximately 64% of CVD occurrences [2]. These statistics highlight the critical need for robust early detection methods and preventive healthcare measures to reduce the global impact of CVD. CVDs result from a multifaceted interaction between controllable and uncontrollable risk factors. Lifestyle-related elements such as elevated blood cholesterol, diabetes, excess body weight, tobacco use, and lack of physical activity are among the key modifiable influences [4, 5]. In contrast, unchangeable determinants include

advancing age, biological sex, and racial or ethnic background [6-9]. The widespread adoption of unhealthy behaviors in today's society has intensified these risks [10, 11]. Therefore, identifying high-risk individuals with precision and at an early stage is essential to enable preventive actions, slow disease progression, and decrease mortality rates.

Traditional clinical risk assessment tools—such as the Framingham Risk Score, SCORE risk charts, and the REGICOR risk score—have been widely used in practice [12-14]. However, these models often rely on a limited set of features and assume linear relationships between predictors and outcomes, potentially oversimplifying the true complexity of cardiovascular risk. Moreover, these models can be population-specific and may lack generalizability across diverse demographic and clinical settings. To address these challenges, machine learning has gained prominence as an advanced approach for improving the accuracy of cardiovascular risk assessment. Machine learning methods are capable of modeling complex, non-linear interactions among diverse features, offering greater flexibility and predictive accuracy than traditional statistical techniques [15-18]. Frequently applied machine learning algorithms in the prediction of cardiovascular diseases encompass decision tree models, support vector machines (SVM), logistic regression, *k*-nearest neighbors (KNN), random forest classifiers, gradient boosting methods, XGBoost, and advanced deep learning architectures like convolutional neural networks (CNNs) [2, 7, 19-23].

While personalized predictive modeling has advanced, significant difficulties remain in fully grasping the intricate relationships between various contributing factors, their differing effects across population subgroups, and how to tailor the most effective treatment plans for each individual. A broad range of socio-demographic, behavioral, and clinical factors contribute to variability in CVD outcomes, such as gender, age, race/Hispanic origin, education level, blood pressure readings, body mass index, waist circumference, lifestyle and physical activity indicators, sleep behavior, diabetes status, socioeconomic indicators, lipid profiles, glycemic control, inflammation markers, and smoking history. The complexity and interplay of these variables highlight the need for data-driven models capable of capturing these nuances. We hypothesize that a machine learning method can be harnessed to uncover and rank the most influential factors in predicting CVD risk, offering an objective and individualized framework for risk evaluation. Such a model has the potential to assist healthcare professionals in selecting the most suitable, patient-specific treatment plans to improve CVD outcomes.

For our study, we utilized data from the National Health and Nutrition Examination Survey (NHANES), a publicly accessible dataset curated by the Centers for Disease Control and Prevention (CDC). We developed a baseline predictive pipeline for CVD risk prediction using a machine learning algorithm and integrated robust feature selection techniques. By

evaluating feature contributions, we aim to guide the collection and prioritization of essential clinical and lifestyle data for optimizing risk prediction and treatment decisions. This approach is intended to support evidence-based decision-making by health professionals and improve treatment outcomes for patients at risk of CVD.

Materials and methods

Figure 1 outlines the full workflow of our proposed methodology. The process begins with collecting data from individuals classified as either at risk for cardiovascular disease (CVD) or not at risk. From this data, a pool of candidate features is generated. We then apply multiple feature evaluation techniques—including the Alternating Decision Tree (ADT) [24, 25], Cross-Validated Feature Evaluation (CVFE) [26], and Hypergraph-Based Feature Evaluation (HFE) [27]—to eliminate features with low relevance or minimal impact. The refined feature subsets are subsequently utilized to train an XGBoost model, which is evaluated for its predictive performance.

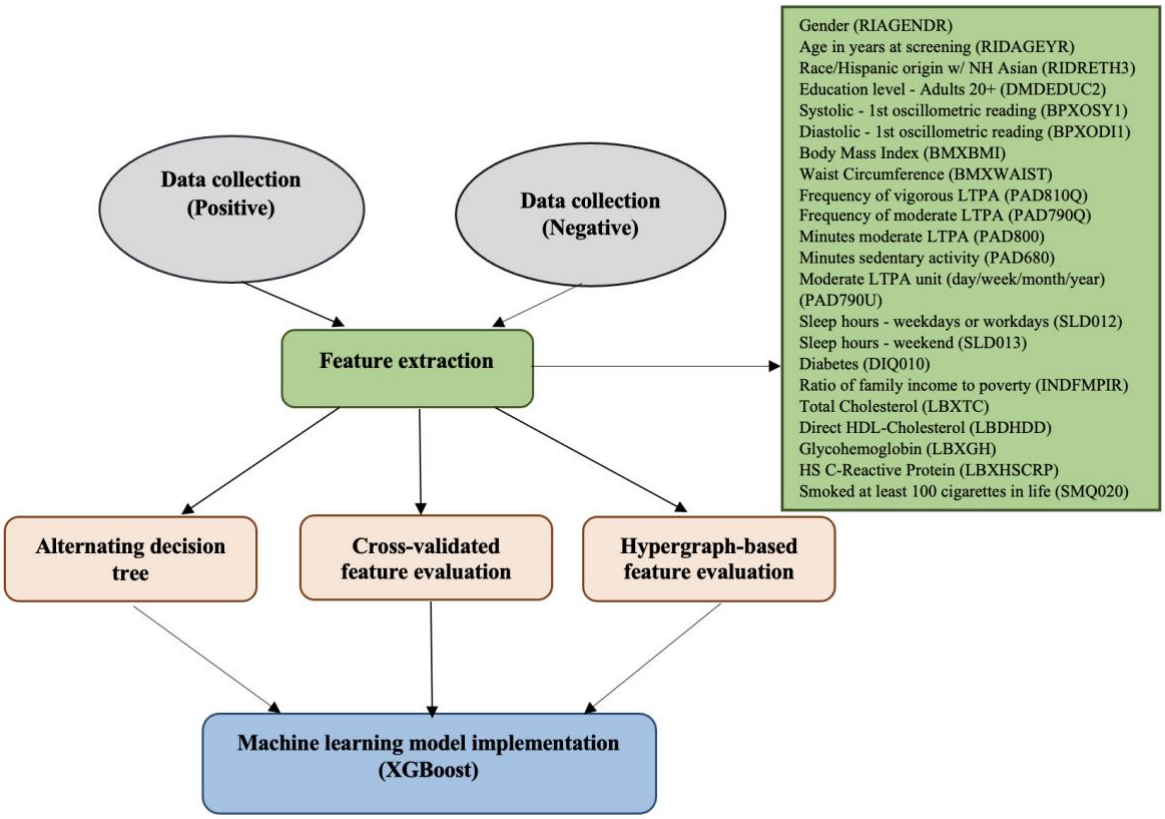


Figure 1 - Overview of the procedure for detecting cardiovascular risk.

Data and features

The data for this study was sourced from the National Health and Nutrition Examination Survey (NHANES), a publicly available program overseen by the National Center for Health Statistics and the Centers for Disease Control and Prevention (CDC) [28]. The final dataset consisted of 341 individuals identified as having cardiovascular disease (CVD) risk and 3,230 individuals without such risk. To mitigate class imbalance, a random sampling strategy was applied to downsample the non-CVD group to 341 cases, achieving a balanced dataset. To construct the model, the dataset was partitioned such that 80% was used for training purposes, while the remaining 20% was reserved for evaluation on a test set. Table 1 presents the selected variables, their categorical groupings, demographic distributions, and corresponding *p*-values, with a threshold of $p < 0.05$ denoting statistical significance. The outcome variable, CVD risk, was defined such that individuals with a reported history of stroke, heart attack, coronary heart disease, or heart failure were classified as at risk for CVD. A total of 22 features were initially included as candidate predictors and are listed in Supplementary Table S1 (**Supplementary Material**).

Table 1 – Summary of the demographic profile and clinical features of the study population.

Characteristics		All subjects (N = 642)	CVD risk (N = 321)	No CVD risk (N = 321)	p-value
SMQ020 - Smoked at least 100 cigarettes in life, N (%)	1 (Yes)	344 (50.44)	218 (63.93)	126 (36.95)	<0.0001[a]
	2 (No)	338 (49.56%)	123 (36.07)	215 (63.05)	
PAD810Q - Frequency of vigorous LTPA, Median (IQR)		5.39760534693403e-79 (5.39760534693403e-79, 2.0)	5.39760534693403e-79 (5.39760534693403e-79, 2.0)	1.0 (5.39760534693403e-79, 2.0)	<0.0001[b]
PAD790U - Moderate LTPA unit, N (%)	1 (Day)	95 (13.93)	59 (17.3)	36 (10.56)	0.0559[a]
	2 (Week)	520 (76.25)	248 (72.73)	272 (79.77)	
	3 (Month)	56 (8.21)	27 (7.92)	29 (8.5)	
	4 (Year)	11 (1.61)	7 (2.05)	4 (1.17)	
PAD800 - Minutes moderate LTPA, Median (IQR)		45 (30, 60)	45 (30, 60)	60 (30, 60)	<0.0001[b]
PAD680 - Minutes sedentary activity, Median (IQR)		300 (195, 480)	300 (240, 480)	300 (180, 480)	<0.0001[b]
	3 to 13.5 (Range of Values)	677 (99.27)	336 (98.53)	341 (100.00)	0.0806[a]

SLD012 - Sleep hours - weekdays or workdays, N (%)	2 (Less than 3 hours)	3 (0.44)	3 (0.88)	0 (0.00)	
	14 (14 hours or more)	2 (0.29)	2 (0.59)	0 (0.00)	
SLD013 - Sleep hours – weekends, N (%)	3 to 13.5 (Range of Values)	679 (99.56)	339 (99.41)	340 (99.71)	0.6061[a]
	2 (Less than 3 hours)	1 (0.15)	1 (0.29)	0 (0.00)	
	14 (14 hours or more)	2 (0.29)	1 (0.29)	1 (0.29)	
DIQ010 - Doctor told you have diabetes N (%)	1 (Yes)	132 (19.35)	98 (28.74)	34 (9.97)	<0.001[a]
	2 (No)	520 (76.25)	223 (65.4)	297 (87.1)	
	3 (Borderline)	30 (4.4)	20 (5.87)	10 (2.93)	
RIDAGEYR - Age in years at screening, N (%)	0 to 79 (Range of Values)	625 (91.64)	294 (86.22)	331 (97.07)	<0.001[a]
	80 (80 years of age and over)	57 (8.36)	47 (13.78)	10 (2.93)	
RIAGENDR – Gender, N (%)	1 (Male)	367 (53.81)	211 (61.88)	156 (45.75)	<0.001[a]
	2 (Female)	315 (46.19)	130 (38.12)	185 (54.25)	
RIDRETH3 - Race/Hispanic origin w/ NH Asian, N (%)	1 (Mexican American)	28 (4.11)	8 (2.35)	20 (5.87)	0.0153[a]
	2 (Other Hispanic)	62 (9.09)	27 (7.92)	35 (10.26)	
	3 (Non-Hispanic White)	438 (64.22)	224 (65.69)	214 (62.76)	
	4 (Non-Hispanic Black)	74 (10.85)	38 (11.14)	36 (10.56)	
	6 (Non-Hispanic Asian)	36 (5.28)	14 (4.11)	22 (6.45)	
	7 (Other Race - Including Multi-Racial)	44 (6.45)	30 (8.8)	14 (4.11)	
DMDEDUC2 - Education level - Adults 20+, N (%)	1 (Less than 9th grade)	30 (4.4)	22 (6.45)	8 (2.35)	<0.001[a]
	2 (9-11th grade (Includes 12th grade with no diploma))	49 (7.18)	35 (10.26)	14 (4.11)	
	3 (High school graduate/GED or equivalent)	136 (19.94)	78 (22.87)	58 (17.01)	
	4 (Some college or AA degree)	220 (32.26)	112 (32.84)	108 (31.67)	

	5 (College graduate or above)	247 (36.22)	94 (27.57)	153 (44.87)	
INDFMPIR - Ratio of family income to poverty, N (%)	0 to 4.99 (Range of Values)	509 (74.63)	272 (79.77)	237 (69.50)	0.0028[a]
	5 (Value greater than or equal to 5.00)	173 (25.37)	69 (20.23)	104 (30.50)	
BMXBMI - Body Mass Index (kg/m**2), Median (IQR)		28.50 (24.92, 33.18)	28.90 (25.40, 33.30)	28.10 (24.40, 33.10)	<0.001[b]
BMXWAIST - Waist Circumference (cm), Median (IQR)		100.70 (90.53, 111.70)	104.10 (93.90, 115.80)	98.00 (87.60, 109.20)	<0.001[b]
BPXOSY1 - Systolic - oscillometric reading, Median (IQR)		122.00 (112.00, 135.00)	127.00 (114.00, 140.00)	119.00 (111.00, 130.00)	<0.001[b]
BPXODI1 - Diastolic - oscillometric reading, Median (IQR)		75.00 (67.00, 82.00)	74.00 (66.00, 82.00)	75.00 (68.00, 83.00)	<0.001[b]
LBXTC - Total Cholesterol (mg/dL), Median (IQR)		176.00 (148.00, 210.00)	158.00 (134.00, 189.00)	194.00 (168.00, 223.00)	<0.001[b]
LBDHDD - Direct HDL-Cholesterol (mg/dL) - Total Cholesterol (mg/dL), Median (IQR)		51.00 (43.00, 61.00)	49.00 (41.00, 59.00)	53.00 (44.00, 64.00)	<0.001[b]
LBXGH - Glycohemoglobin (%), Median (IQR)		5.60 (5.30, 6.00)	5.80 (5.50, 6.40)	5.50 (5.20, 5.70)	<0.001[b]
LBXHSCRP - HS C-Reactive Protein (mg/L), Median (IQR)		1.62 (0.80, 4.09)	1.64 (0.84, 4.26)	1.61 (0.76, 3.92)	<0.001[b]

[a] Chi-Squared test

[b] Wilcoxon Two-Sample test

N = Number of individuals, LTPA = Leisure-Time Physical Activity, IQR = Interquartile range, 25% and 75%

Assessment of predictive features

To build an effective predictive model, it is vital to discard non-contributory features during the preprocessing stage. In our study, we independently applied three feature reduction techniques—Alternating Decision Tree (ADT) [24, 25], Cross-Validated Feature Evaluation (CVFE) [26], and Hypergraph-Based Feature Evaluation (HFE) [27]—to refine the original feature set by eliminating attributes with limited predictive value. The ADT approach merges the straightforward, interpretable form of a conventional decision tree with the performance enhancements derived from boosting algorithms. This technique structures its model using decision tree stumps, which are foundational units commonly associated with boosting. One of the notable advantages of ADT is its flexible branching structure; unlike traditional trees with mutually exclusive paths, ADT allows overlapping routes, enabling multiple decision paths to contribute simultaneously to a prediction. The structure begins with a prediction node that assigns a numerical score, followed by layers of decision nodes that contain conditions used to evaluate input features. These layers alternate in a pattern—prediction nodes followed by decision nodes and vice versa. Decision nodes apply specific logical criteria, while prediction nodes assign fixed numeric contributions to the outcome. Importantly, prediction nodes appear at both the starting point (root) and terminal ends (leaves) of the tree, underscoring the distinctive, layered composition and operational logic that sets ADT apart from conventional decision tree models.

The ADT constructs a series of classification rules, each composed of three main elements: a prerequisite condition, a logical condition, and a pair of numerical scores. The logical condition takes the form of a predicate expressed as "feature <operator> threshold," while the prerequisite is a compound logical statement formed by combining multiple such conditions using conjunctions. These rules are evaluated hierarchically using nested "if" statements, and their respective scores are used to compute the final prediction for a data sample. The procedure initiates with a root rule defined by unconditional logic—both the prerequisite and condition are set to "true"—and corresponding scores are computed using the weights assigned to training instances. Initially, each training sample is assigned an equal weight of $\frac{1}{t}$, where t denotes the total number of training examples. During the training process, the ADT algorithm repeatedly generates new rules by identifying the most effective pair of prerequisite and condition that minimizes a specific objective function, denoted as z. This function evaluates the discriminatory power of a rule based on its ability to separate positive and negative classes effectively. For every new rule created, updated scores are computed through a boosting-based mechanism. Training sample weights are also revised in accordance with the rule's classification accuracy on each example, emphasizing incorrectly predicted instances to refine future splits. This iterative process continues until a predetermined stopping criterion is satisfied—such as reaching the maximum number of iterations or

when further performance gains become negligible. The resulting rule set defines an alternating decision tree, where each prediction node holds a scalar value, and the tree's topology is dictated by the prerequisite logic embedded within the constructed rules. Only a portion of the total feature set is used in the final ADT, reflecting the most relevant variables identified through this process. For implementation, we evaluated 50 randomly selected values for B, representing the number of boosting cycles. The final ADT model—shown in Supplementary Figure S1 (**Supplementary Material**)—demonstrates the decision structure derived from 22 candidate features. Ultimately, 14 features were retained through the ADT-based selection process, as presented in Supplementary Table S2 (**Supplementary Material**).

In addition, we incorporated both the Cross-Validated Feature Evaluation (CVFE) approach and a hypergraph-based technique to refine the initial collection of features. The CVFE process is illustrated in Figure 2. Initially, the dataset was randomly divided into c distinct subsets. For each subset, we applied the XGBoost algorithm to determine the most influential features, optimizing model hyperparameters through a grid search procedure. Feature selection was conducted independently for every subset, followed by the construction of an intersected feature set comprising features common to all subsets. This entire process was repeated e times, resulting in e intersected feature sets. Subsequently, any feature appearing in at least $(p \times 100)\%$ of these intersected sets was incorporated into the final selected feature list. Table 2 details the quantity of features obtained from CVFE under various parameter settings of c, e, and p. A comprehensive list of the selected key features can be found in Supplementary Tables S3-S6 included in the **Supplementary Material**.

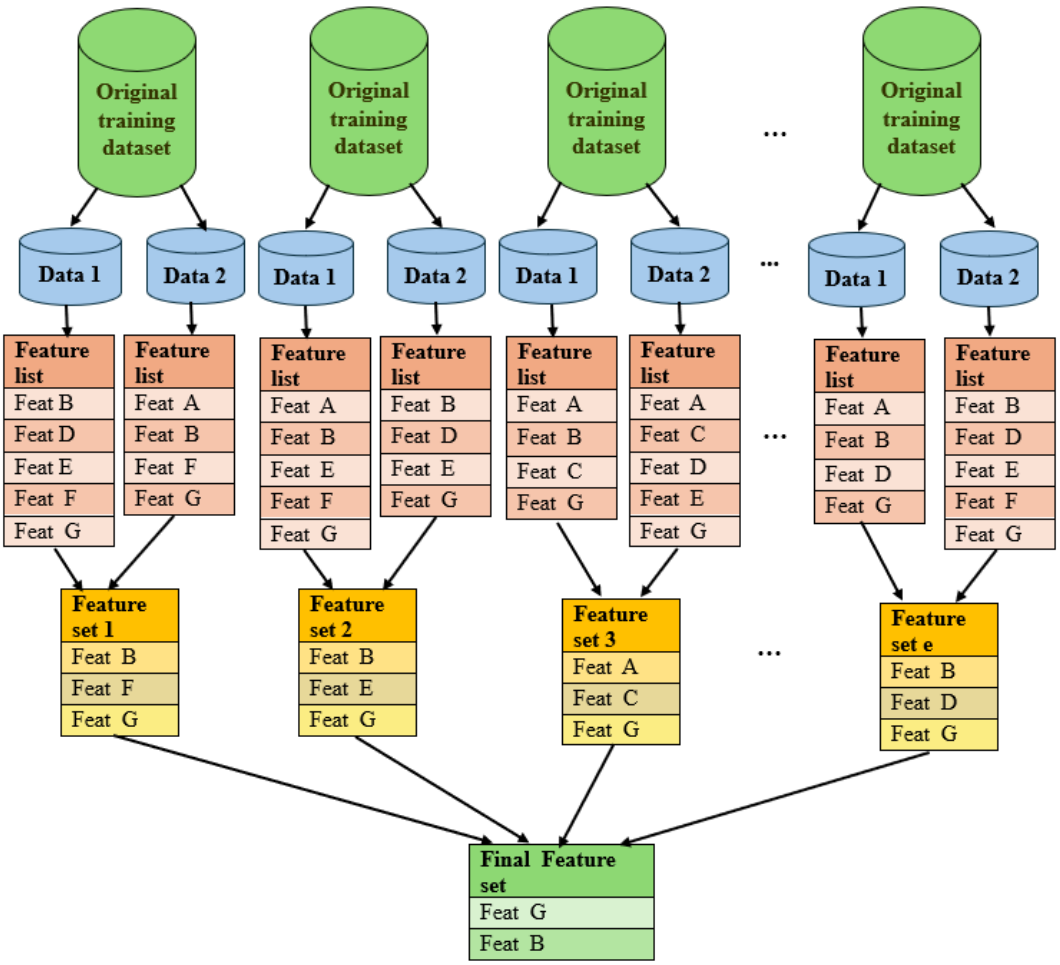


Figure 2 - Visual representation of the feature selection process using the Cross-Validated Feature Evaluation (CVFE) method.

The hypergraph-based feature evaluation (HFE) technique involves modeling feature interactions through a hypergraph structure. Unlike traditional graphs, represented as G(V, E), where vertices (V) are connected by pairwise edges (E), a hypergraph generalizes this concept by permitting each edge—termed a hyperedge—to link multiple vertices simultaneously. Formally, a hypergraph is defined as G(V, E), with V as the set of nodes and E comprising hyperedges, each of which is a subset of V. Figure 3 contrasts a standard graph with a hypergraph, while a hypergraph visualization of the complete training dataset is provided in Supplementary Figure S2 (**Supplementary Material**). In applying HFE, each continuous feature was discretized into a fixed number of intervals (bins), and the feature relevance scores were computed based on the complex, multi-dimensional relationships captured within the hypergraph structure. These scores were derived by assigning weights to the hyperedges that represent the corresponding feature bins. The ranking of features was guided by the hypergraph cut conductance minimization framework [27], ultimately producing a rating vector that reflects the relative importance of each feature. To determine the final feature subset, the top z features were selected based on their sorted importance values, where z

is computed as $\beta \times m$ (m being the total number of features and β representing the selection ratio). In our experiments, we used discretization settings of bin = 5 and bin = 10. The resulting feature counts for varying β values and bin sizes are reported in Table 2. The specific features retained through HFE under each bin configuration are detailed in Supplementary Tables S6 and S7 (**Supplementary Material**).

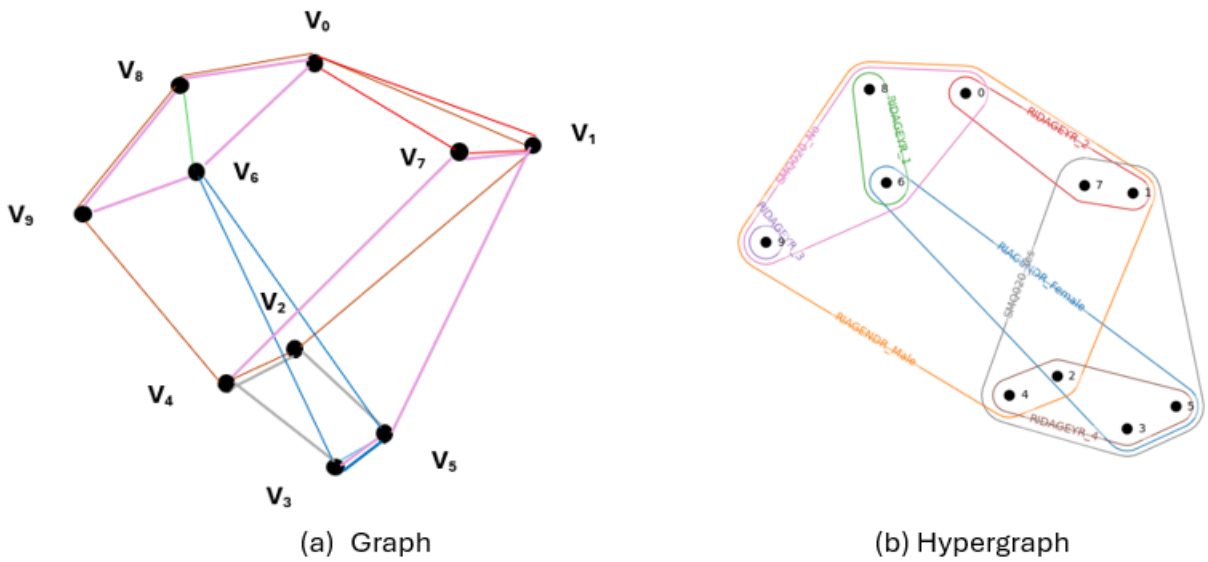


Figure 3 – Comparison between a standard graph and a hypergraph.

Code and data availability

The full set of experimental data along with all relevant scripts can be accessed at https://github.com/suraiya14/CVDRP, as referenced in the **Supplementary Material**.

Results

Following the reduction of the original feature set using three independent feature selection strategies, we developed individual predictive models based on the selected features using the XGBoost algorithm [29]. To interpret and analyze the contribution of each feature within these models, we applied SHapley Additive exPlanations (SHAP) [30]. SHAP provides a unified framework for attributing the impact of each input variable by calculating how much each feature contributes to the model's output, effectively measuring its marginal effect on prediction outcomes.

Performance assessment

An XGBoost-based predictive model was constructed by utilizing different sets of features selected through ADT, CVFE, and HFE methodologies in conjunction with the training data. The model's performance was assessed on the test data using the metrics defined in Equations 1 - 4, where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative counts, respectively. Among these evaluation metrics, accuracy was employed to determine the ratio of correctly predicted outcomes relative to the overall number of samples in the dataset.

$$Test_{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Test_{recall} = \frac{TP}{TP+FN} \tag{2}$$

$$Test_{precision} = \frac{TP}{TP+FP} \tag{3}$$

$$Test_{F-1} = 2 \times \frac{(Test_{precision} \times Test_{recall})}{(Test_{precision} + Test_{recall})} \tag{4}$$

In addition, we computed recall and precision to further evaluate model performance. Recall quantifies the ability of the model to correctly detect actual positive cases, whereas precision reflects the proportion of positive predictions that are genuinely correct. To provide a balanced evaluation that incorporates both precision and recall, we utilized the F1 score, which is defined as their harmonic mean. Moreover, the Area Under the Curve (AUC) metric was used to assess the effectiveness of the binary classifier. Higher AUC values correspond to stronger model performance, with a score of 1 indicating flawless classification and 0.5 representing performance equivalent to random guessing.

Table 2 - Summary of the number of selected features and corresponding evaluation metrics—accuracy, precision, recall, F1 score, and AUC—on the testing dataset for different feature subsets.

Feature evaluation algorithm	Configuration	Number of features	$Test_{Acc}$	$Test_{Precision}$	$Test_{Recall}$	$Test_{F1}$	$Test_{AUC}$
ADT	$B = 50$	10	0.6985	0.7143	0.6618	0.6870	0.7744

CVFE	(c = 2, e = 10, p = 0.2)	22	0.7353	0.7581	0.6912	0.7231	0.7829
	(c = 2, e = 10, p = 0.6)	21	0.7353	0.7581	0.6912	0.7231	0.7850
	(c = 2, e = 5, p = 0.8)	20	0.7353	0.7581	0.6912	0.7231	0.7841
	(c = 3, e = 5, p = 0.6)	20	0.7353	0.7581	0.6912	0.7231	0.7902
HFE Bin = 5	β = 25	6	0.7279	0.7460	0.6912	0.7176	0.7805
	β = 50	11	0.7279	0.7460	0.6912	0.7176	0.7956
	β = 75	16	0.7353	0.7424	0.7206	0.7313	0.7822
HFE Bin = 10	β = 25	6	0.75	0.7833	0.6912	0.7344	0.7857
	β = 50	11	0.7353	0.7424	0.7206	0.7313	0.7898
	β = 75	16	0.6912	0.7167	0.6324	0.6719	0.7803

Legend:

$Test_{acc}$: Accuracy on the testing dataset

$Test_{precision}$: Precision on the testing dataset

$Test_{recall}$: Recall on the testing dataset

$Test_{F1}$: F1 score on the testing dataset

$Test_{AUC}$: AUC on the testing dataset

B: Number of boosting cycles

c: Number of disjoint sub-parts

e: Number of repeated runs

p: Proportions of repeated runs for extracting common features

β: percentage of features selected

Table 2 presents a detailed comparison of XGBoost model performance using feature subsets selected through ADT, CVFE, and HFE techniques. Corresponding confusion matrices for each reduced feature set are visualized in Supplementary Figure S3 (**Supplementary Material**). Overall, the HFE feature set (bin=10, *β*=25) yielded superior predictive performance compared to the ADT and CVFE sets, with the best model successfully identifying 47 out of 60 patients at risk for cardiovascular disease.

Features identified using the HFE Approach

The highest predictive performance was attained using the XGBoost algorithm paired with the HFE feature selection method configured with bin=10 and *β*=25. Figure 4 presents a SHAP summary bar plot illustrating the importance rankings of the features identified through SHAP value analysis applied to that model. From an initial set of 22 features, HFE selected six key predictors: RIDAGEYR (age at screening), LBXTC (total cholesterol in mg/dL), LBXGH (glycohemoglobin), BPXOSY1 (systolic blood pressure via oscillometric measurement), SMQ020 (history of smoking at least 100 cigarettes), and DIQ010 (physician-diagnosed diabetes), ordered according to their relative influence on the model's predictions.

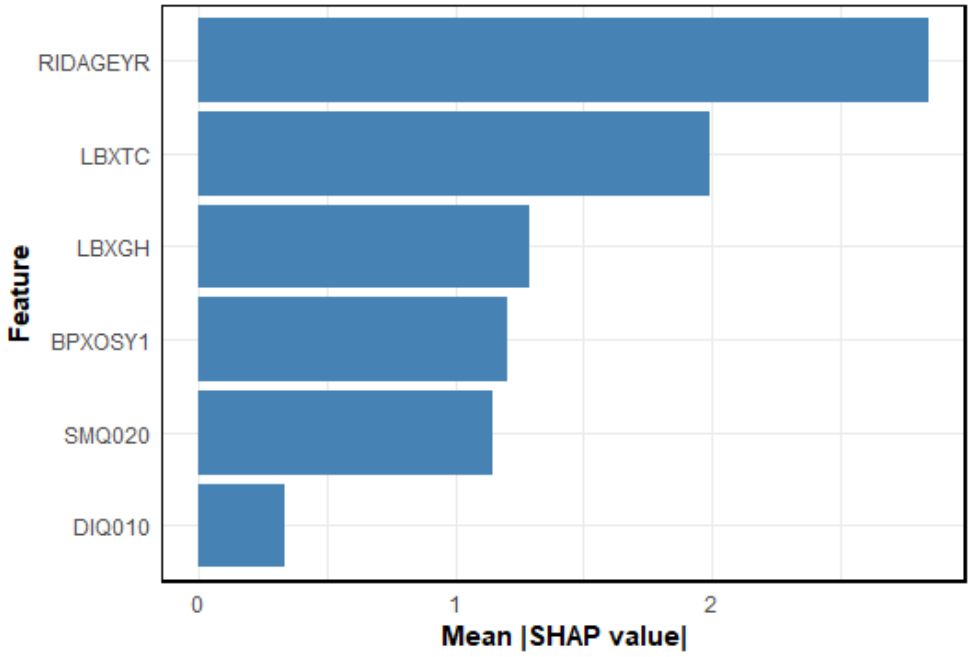


Figure 4 – Importance of features selected by the XGBoost model shown as a bar chart. The *x*-axis represents the mean absolute SHAP value for each feature, reflecting its average contribution to the model's predictions. Features are arranged from most to least influential.

Feature impact assessment

Figure 5 presents the importance of the key continuous features identified in Figure 4 through SHAP value analysis. Each point corresponds to an individual patient, with overlapping points slightly offset (jittered) to reflect the density of data. The *x*-axis displays how much each feature influences the model's prediction, where higher SHAP values correspond to an elevated risk of CVD. The color gradient—from yellow to purple—represents low to high feature values, respectively. Notably, increased SHAP values for age, glycohemoglobin levels, and systolic blood pressure (oscillometric measurement) are linked to a greater likelihood of CVD. While higher total cholesterol often aligns with increased risk, this pattern does not hold consistently across all patients.

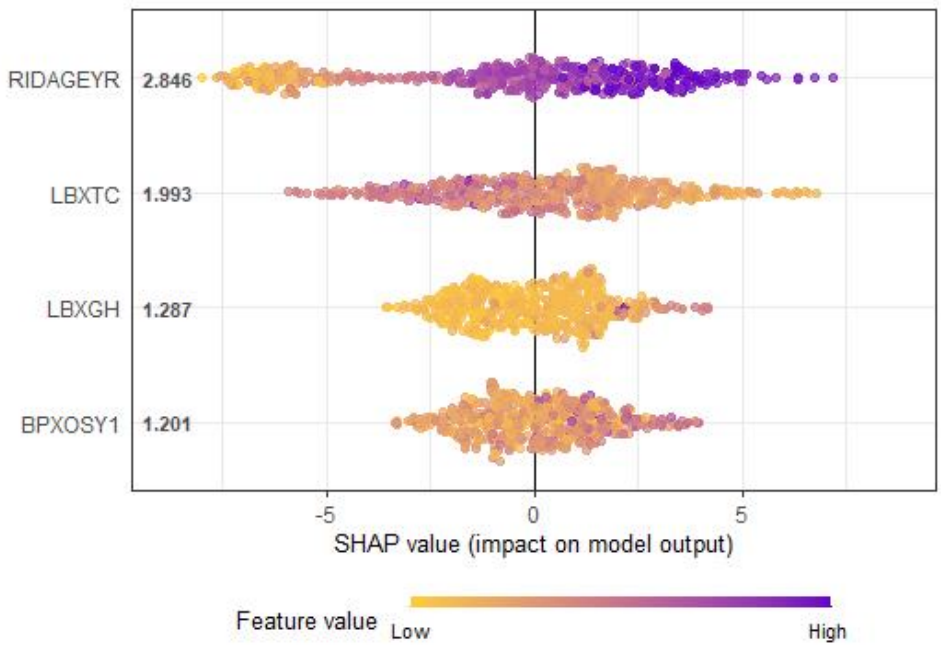
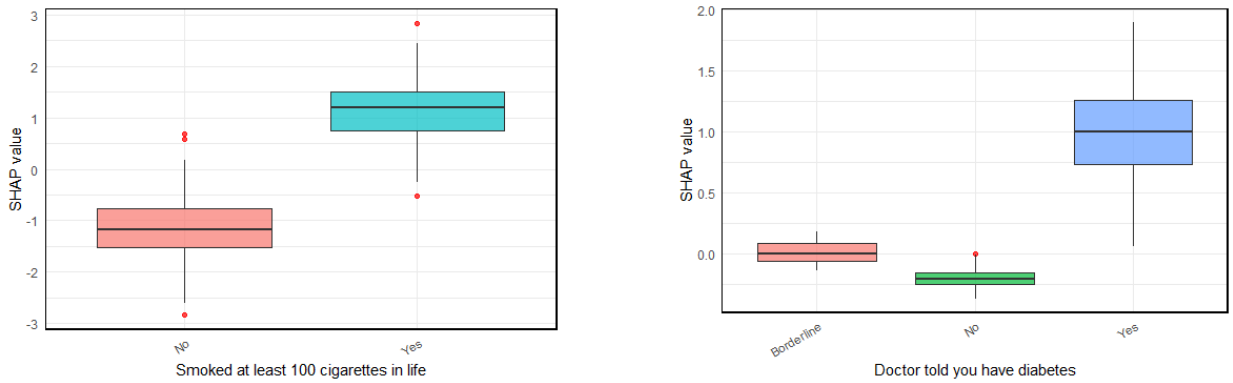


Figure 5 – SHAP analysis illustrating the impact of the top continuous features on model predictions.



(a) SMQ020 (b) DIQ010

Figure 6 – SHAP analysis illustrating the impact of the top categorical features on model predictions.

Figure 6 displays boxplots representing the SHAP values for the different categories of the categorical features identified in Figure 4. Since elevated SHAP values correspond to increased CVD risk, the boxplots allow for straightforward interpretation of risk levels across categories. Specifically, Figure 6(a) shows that individuals who have smoked at least 100

cigarettes in their lifetime exhibit a higher CVD risk. Likewise, Figure 6(b) indicates that patients diagnosed with diabetes tend to have an elevated risk of developing CVD.

Discussion

Accurate individualized prediction plays a vital role in enhancing patient outcomes in CVD risk evaluation. This work introduced a CVD risk prediction model built on the XGBoost machine learning framework. By systematically removing unnecessary and non-contributory features, we assessed the model's effectiveness using a streamlined feature subset. The developed prediction pipeline successfully pinpointed critical variables affecting CVD risk, offering valuable insights for tailored risk assessment, guiding therapeutic approaches, and aiding clinical decision processes. Our results are consistent with existing literature. For example, Peng et al. [31] demonstrated that the XGBoost and XGBH algorithms achieved superior predictive accuracy, with systolic blood pressure identified as the most influential variable. Likewise, Khaw and Wareham [32] emphasized the significant link between glycated hemoglobin levels and cardiovascular risk, noting that even slight increases below the diabetic threshold correspond to a 10–20% elevation in CVD risk. The widely implemented QRISK3 model in the UK incorporates factors such as age, systolic blood pressure, cholesterol ratios, smoking habits, and diabetes status as primary indicators of cardiovascular risk [33]. Additionally, Shishehbori and Awan [34] surveyed recent machine learning methods in this field, underscoring that conventional clinical markers—including age, blood pressure, cholesterol, smoking history, and diabetes—continue to be the most influential predictors across diverse models.

Our model highlighted age as the most impactful predictor, a conclusion corroborated by multiple prior investigations [35-38] that emphasize age as a key determinant of cardiovascular disease risk. Furthermore, the identification of other important variables—including total cholesterol, glycated hemoglobin, systolic blood pressure, smoking history, and diabetes status—carries substantial clinical relevance. Prioritizing these features allows healthcare providers to enhance the accuracy of risk stratification, tailor prevention efforts, and effectively allocate resources for managing cardiovascular conditions. While our methodology offers notable advantages, several limitations must be acknowledged. The accuracy of feature contribution assessments may be compromised in subpopulations where data are limited. Moreover, the present model is not designed to handle time-to-event analyses or longitudinal forecasting, restricting its usefulness for tracking disease progression over time. Future research should aim to expand the model's capabilities to address these outcomes.

In conclusion, we constructed and validated a machine learning framework to estimate cardiovascular disease risk by leveraging a carefully selected set of features from NHANES datasets. Although the model achieved satisfactory predictive performance, there remains room for improvement in both accuracy and AUC metrics. The inclusion of more detailed, high-dimensional data sources—such as multi-omics profiles or longitudinal clinical measurements—has the potential to boost predictive precision and clinical relevance. Additionally, efforts will be directed toward creating an intuitive, accessible interface to facilitate the model's adoption and practical use by healthcare professionals.

Supporting information

The supplementary material for this article is provided herewith the manuscript.

(DOCX)

Authors' contributions

SA: Conceptualization, data collection, formal analysis, software implementation, validation, visualization, and writing manuscript. JHM: Conceptualization, supervision, reviewing analyses, and editing the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethical approval statement

The study was conducted in accordance with the Declaration of Helsinki.

References

1. Organization WH. Health Stats 2017: World Health Organization; 2017 [June 2025]. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

2. Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics. 2024;14(2):144.

3. Mousa D, Zayed N, Yassine IA, editors. Automatic cardiac MRI localization method. 2014 Cairo International Biomedical Engineering Conference (CIBEC); 2014: IEEE.

4. Sen SK. Predicting and diagnosing of heart disease using machine learning algorithms. International Journal of Engineering and Computer Science. 2017;6(6):21623-31.

5. Yazdani A, Varathan KD, Chiam YK, Malik AW, Wan Ahmad WA. A novel approach for heart disease prediction using strength scores with significant predictors. BMC medical informatics and decision making. 2021;21(1):194.

6. Guarneros-Nolasco LR, Cruz-Ramos NA, Alor-Hernández G, Rodríguez-Mazahua L, Sánchez-Cervantes JL. Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms. Mathematics. 2021;9(20):2537.

7. Mandava M. MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning. Biomedical Signal Processing and Control. 2024;93:106147.

8. Pu L-N, Zhao Z, Zhang Y-T. Investigation on cardiovascular risk prediction using genetic information. IEEE Transactions on Information Technology in Biomedicine. 2012;16(5):795-808.

9. Hossen MA, Tazin T, Khan S, Alam E, Sojib HA, Monirujjaman Khan M, et al. Supervised machine learning-based cardiovascular disease analysis and prediction. Mathematical Problems in Engineering. 2021;2021(1):1792201.

10. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J, editors. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. 2017 IEEE symposium on computers and communications (ISCC); 2017: IEEE.

11. Ambekar S, Phalnikar R, editors. Disease risk prediction by using convolutional neural network. 2018 Fourth international conference on computing communication control and automation (ICCUBEA); 2018: IEEE.

12. Gil-Guillen V, Orozco-Beltran D, Maiques-Galan A, Aznar-Vicente J, Navarro J, Cea-Calvo L, et al. Agreement between REGICOR and SCORE scales in identifying high cardiovascular risk in the Spanish population. Revista Espanola De Cardiologia. 2007;60(10):1042.

13. Khandoker AH, Al Zaabi Y, Jelinek HF, editors. What can tone and entropy tell us about risk of cardiovascular diseases? 2019 Computing in Cardiology (CinC); 2019: IEEE.

14. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation. 2002;105(3):310-5.

15. Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015;7(1):129-37.

16. Solanki A, Barot MP. Study of heart disease diagnosis by comparing various classification algorithms. International Journal of Engineering and Advanced Technology. 2019;8(2S2):40-2.

17. Kiran P, Swathi A, Sindhu M, Manikanta Y. Effective heart disease prediction using hybrid machine learning technique. South Asian Journal of Engineering and Technology. 2022;12(3):123-30.

18. Holmes J, Sacchi L, Bellazzi R. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86:334-8.

19. Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Medical engineering & physics. 2022;105:103825.

20. Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW. An integrated machine learning framework for effective prediction of cardiovascular diseases. IEEE access. 2021;9:106575-88.

21. Rubini P, Subasini C, Katharine AV, Kumaresan V, Kumar SG, Nithya T. A cardiovascular disease prediction using machine learning algorithms. Annals of the Romanian Society for Cell Biology. 2021;25(2):904-12.

22. Elsayed HAG, Syed L, editors. An automatic early risk classification of hard coronary heart diseases using framingham scoring model. Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing; 2017.

23. DeGroat W, Abdelhalim H, Patel K, Mendhe D, Zeeshan S, Ahmed Z. Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. Scientific reports. 2024;14(1):1.

24. Akhter S, Miller JH. BPAGS: a web application for bacteriocin prediction via feature evaluation using alternating decision tree, genetic algorithm, and linear support vector classifier. Frontiers in Bioinformatics. 2024;3:1284705.

25. Freund Y, Mason L, editors. The alternating decision tree learning algorithm. icml; 1999.

26. Yang M-R, Wu Y-W. A Cross-Validated Feature Selection (CVFS) approach for extracting the most parsimonious feature sets and discovering potential antimicrobial resistance (AMR) biomarkers. Computational and Structural Biotechnology Journal. 2023;21:769-79.

27. Misiorek P, Janowski S. Hypergraph-based importance assessment for binary classification data. Knowledge and Information Systems. 2023;65(4):1657-83.

28. Terry A, Chiappa M, McAllister J, Woodwell D, Graber J. Plan and operations of the National Health and Nutrition Examination Survey, August 2021–August 2023. 2024.

29. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.

30. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020;2(1):56-67.

31. Peng M, Hou F, Cheng Z, Shen T, Liu K, Zhao C, et al. Prediction of cardiovascular disease risk based on major contributing features. Scientific Reports. 2023;13(1):4778.

32. Khaw K-T, Wareham N. Glycated hemoglobin as a marker of cardiovascular risk. Current opinion in lipidology. 2006;17(6):637-43.

33. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart. 2008;94(1):34-9.

34. Shishehbori F, Awan Z. Enhancing cardiovascular disease risk prediction with machine learning models. arXiv preprint arXiv:240117328. 2024.

35. Mansoori A, Allahyari M, Mirvahabi MS, Tanbakuchi D, Ghoflchi S, Derakhshan-Nezhad E, et al. Predictive properties of novel anthropometric and biochemical indexes for prediction of cardiovascular risk. Diabetology & Metabolic Syndrome. 2024;16(1):304.

36. van Os HJ, Kanning JP, Bonten TN, Rakers MM, Putter H, Numans ME, et al. Cardiovascular risk prediction in men and women aged under 50 years using routine care data. Journal of the American Heart Association. 2023;12(7):e027011.

37. Neumann JT, Thao LT, Callander E, Chowdhury E, Williamson JD, Nelson MR, et al. Cardiovascular risk prediction in healthy older people. Geroscience. 2022;44(1):403-13.

38. Wallisch C, Agibetov A, Dunkler D, Haller M, Samwald M, Dorffner G, et al. The roles of predictors in cardiovascular risk models-a question of modeling culture? BMC Medical Research Methodology. 2021;21:1-12.