

Machine Learning-Based Analysis of Lifestyle Risk Factors in Atherosclerotic Cardiovascular Disease Risk++

Hye-Jin Kim, Heeji Choi, Hyo-Jung Ahn, Seung-Ho Shin, Chulho Kim, Sang-Hwa Lee, Jong-Hee Sohn, Jae-Jun Lee

Submitted to: JMIR Medical Informatics
on: March 24, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 16

 Figures 17

 Figure 1..... 18

 Figure 2..... 19

 Figure 3..... 20

 Multimedia Appendixes 21

 Multimedia Appendix 1..... 22

Machine Learning-Based Analysis of Lifestyle Risk Factors in Atherosclerotic Cardiovascular Disease Risk++

Hye-Jin Kim¹ PhD; Heeji Choi² MSc; Hyo-Jung Ahn³ MSc; Seung-Ho Shin⁴ PhD; Chulho Kim⁵ MD, PhD; Sang-Hwa Lee⁵ MD, PhD; Jong-Hee Sohn⁵ MD, PhD; Jae-Jun Lee⁶ MD, PhD

¹Artificial Intelligence Research Center College of Medicine Hallym University Chuncheon-si KR

²Artificial Intelligence Research Center College of Medicine Hallym University Chuncheon-si KR

³Health Insurance Review and Assessment Research Institute Health Insurance Review and Assessment Service Wonju-si KR

⁴Artificial Intelligence Research Center College of Medicine Hallym University Chuncheon-si KR

⁵Department of Neurology College of Medicine Hallym University Chuncheon-si KR

⁶Department of Anesthesiology College of Medicine Hallym University Chuncheon-si KR

Corresponding Author:

Chulho Kim MD, PhD
Department of Neurology
College of Medicine
Hallym University
1, Hallymdaehak-gil
Chuncheon-si
KR

Abstract

Background: The risk of developing atherosclerotic cardiovascular disease (ASCVD) varies among individuals and is related to a variety of lifestyle factors in addition to the presence of chronic diseases.

Objective: We aimed to assess the predictive accuracy of machine learning (ML) models incorporating lifestyle risk behaviors for ASCVD risk using Korean nationwide database.

Methods: Utilizing data from the Korea National Health and Nutrition Examination Survey, five ML algorithms were employed for the prediction of high ASCVD risk: logistic regression, support vector machine, random forest, extreme gradient boosting (XGB), and light gradient boosting (LGB) models. ASCVD risk was assessed using the Pooled Cohort Equations, with a high-risk threshold of ≥7.5% over 10 years. Among the 8,573 participants aged 40–79 years, propensity score matching (PSM) was used to adjust for demographic confounders. We divided the dataset into a training and a test dataset in an 8:2 ratio. We also used bootstrapping to train the ML model with the area under the receiver operating characteristics curve (AUROC) score. Shapley additive explanations was used to identify the models' important variables in assessing high ASCVD risks. In sensitivity analysis, we additionally performed binary logistic regression analysis, in which the ML model's results were consistent with conventional statistical model.

Results: Of 8,573 participants, 41.7% had high ASCVD risk. Before PSM, age and sex differed significantly between groups. PSM (1:1) yielded 1,976 patients with balanced demographics. After PSM, the high ASCVD risk group had higher alcohol/tobacco use, lower omega-3 intake, higher BMI, less physical activity, and spent less time sitting. In 5 ML models, XGB model showed high AUROC values, with LGB model outperforming in accuracy, recall, and F1 score. Variable importance analysis using Shapley additive explanations identified smoking and age as the strongest predictors, while BMI, sodium or omega-3 intake, and LDL cholesterol also had significant variables. Sensitivity analysis using multivariable LR analysis also confirmed these findings, showing that smoking, BMI, and LDL cholesterol increased ASCVD risk, whereas omega-3 intake and physical activity were associated with lower risk.

Conclusions: Analyzing lifestyle behavioral factors in ASCVD risk with ML improves predictive performance compared to traditional models. Personalized prevention strategies tailored to an individual's lifestyle can effectively reduce ASCVD risk.

(JMIR Preprints 24/03/2025:74415)

DOI: <https://doi.org/10.2196/preprints.74415>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [http://www.jmir.org/](#)

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Original Paper

Machine Learning-Based Analysis of Lifestyle Risk Factors in Atherosclerotic Cardiovascular Disease Risk

Abstract

Background: The risk of developing atherosclerotic cardiovascular disease (ASCVD) varies among individuals and is related to a variety of lifestyle factors in addition to the presence of chronic diseases.

Objective: We aimed to assess the predictive accuracy of machine learning (ML) models incorporating lifestyle risk behaviors for ASCVD risk using Korean nationwide database.

Methods: Utilizing data from the Korea National Health and Nutrition Examination Survey, five ML algorithms were employed for the prediction of high ASCVD risk: logistic regression, support vector machine, random forest, extreme gradient boosting (XGB), and light gradient boosting (LGB) models. ASCVD risk was assessed using the Pooled Cohort Equations, with a high-risk threshold of $\geq 7.5\%$ over 10 years. Among the 8,573 participants aged 40–79 years, propensity score matching (PSM) was used to adjust for demographic confounders. We divided the dataset into a training and a test dataset in an 8:2 ratio. We also used bootstrapping to train the ML model with the area under the receiver operating characteristics curve (AUROC) score. Shapley additive explanations was used to identify the models' important variables in assessing high ASCVD risks. In sensitivity analysis, we additionally performed binary logistic regression analysis, in which the ML model's results were consistent with conventional statistical model.

Results: Of 8,573 participants, 41.7% had high ASCVD risk. Before PSM, age and sex differed significantly between groups. PSM (1:1) yielded 1,976 patients with balanced demographics. After PSM, the high ASCVD risk group had higher alcohol/tobacco use, lower omega-3 intake, higher BMI, less physical activity, and spent less time sitting. In 5 ML models, XGB model showed high AUROC values, with LGB model outperforming in accuracy, recall, and F1 score. Variable importance analysis using Shapley additive explanations identified smoking and age as the strongest predictors, while BMI, sodium or omega-3 intake, and LDL cholesterol also had significant variables. Sensitivity analysis using multivariable LR analysis also confirmed these findings, showing that smoking, BMI, and LDL cholesterol increased ASCVD risk, whereas omega-3 intake and physical activity were associated with lower risk.

Conclusions: Analyzing lifestyle behavioral factors in ASCVD risk with ML improves predictive performance compared to traditional models. Personalized prevention strategies tailored to an individual's lifestyle can effectively reduce ASCVD risk.

Keywords: Atherosclerotic Cardiovascular Disease; Machine Learning; Lifestyle Risk Behaviors; Predictive Modeling.

Introduction

Atherosclerotic cardiovascular disease (ASCVD) is the leading cause of morbidity and mortality worldwide [1]. In 2019, cardiovascular diseases (CVDs) were responsible for an estimated 32% of all global deaths [2]. Additionally, out of the 17 million premature deaths caused by non-communicable diseases in 2019, 38% were due to CVDs [2]. The prevalence of CVDs in South Korea is estimated to be increasing recently [3]. While the CVD mortality rate declined until 2010, it has been steadily rising since then [4]. Considering these trends, addressing CVDs is crucial for improving public health outcomes.

The main risk factors for developing ASCVD include hypertension, smoking, and dyslipidemia [5].

Additionally, ASCVD outcomes vary across different ethnic groups based on socioeconomic status, which also influences health-related lifestyle behaviors. [6-8]. As the prevalence of ASCVD is increasing, the importance of primary prevention has been emphasized from a public health perspective [9]. ASCVDs are closely related with modifiable risk factors that can be controlled through lifestyle modification, which is the cornerstone of ASCVD prevention [9]. The most effective prevention strategy is a comprehensive approach that promotes a healthy lifestyle and addresses all major risk factors [10].

Machine learning (ML) has been introduced to the field of medical data analysis, offering numerous benefits that enhance patient care and improve clinical outcomes. ML-based prediction using patients' data such as demographics, clinical and laboratory data, and image data can identify patient future outcomes that might be missed by conventional prediction model [11]. Additionally, ML enables to plan for the precision medicine, early detection of high-risk patients for the disease progression with a high accuracy [11]. Previous studies have constructed ML model for predicting ASCVD, but few have modeled the impact of lifestyle behavior on ASCVD risk [11-13]. ASCVD risk is associated with a variety of demographic factors, including age, race, and gender, as well as a patient's lifestyle behaviors [14, 15]. Therefore, we aimed to predict how patient lifestyle behaviors factors are associated with ASCVD risk.

Methods

Participants

This study a retrospective case control study using nationally representative data from the Korea National Health and Nutrition Examination Survey (KNHANES). The KNHANES is a nationwide cross-sectional survey initiated by the Korea Disease Control and Prevention Agency in 1998 to assess the health and nutritional status of the Korean population. Among the 22,889 participants in the 2019–2021 KNHANES, 10,481 participants 40–79 years of age with no cardiovascular event were initially screened. Subsequently, 1,908 participants with one or more missing data on sodium intake, low-density lipoprotein (LDL)-cholesterol, time spent sitting, occupation, marital status, smoking, income, body mass index (BMI), and weight change over one year were excluded. A total of 8,573 participants were finally selected for analysis (Figure 1).

Measurement of ASCVD risk score

The 10-year ASCVD risk was assessed using the American College of Cardiology/American Heart Association Pooled Cohort Equations (PCEs) risk score [15]. They recommended that PCEs for non-Hispanic whites may be considered for risk estimation in populations other than non-Hispanic African Americans and non-Hispanic whites [15, 16]. Therefore, the estimation function of the ASCVD risk score for the white population was used in this study. This score algorithm is designed to predict both the 10-year (40–79 years of age) and lifetime (20–59 years of age) risk of ASCVD. This includes the assessment of both fatal and non-fatal coronary heart disease and stroke [17, 18]. We defined cardiovascular disease high risk as having a 10-year ASCVD risk of $\geq 7.5\%$.

Measurement of lifestyle risk behaviors

The lifestyle risk behaviors analyzed in this study included sex, age, occupation, marital status, residential area, household income, alcohol consumption, smoking, omega-3 intake, sodium intake, BMI, weight change over 1 year, LDL cholesterol, physical activity, and time spent sitting. The occupation was classified into three categories of non-manual workers, manual workers, and other workers. Marital status was separated, widowed, and divorced participants were assigned the no spouse status. Residential area was grouped as urban or rural areas. Alcohol consumption was divided into two categories, where participants drank at least 2 days per week and fewer than 2 days

per week [19], regardless of the amount of alcohol consumed. Individuals who had smoked at least 100 cigarettes in their lifetime and currently smoked were classified as smokers, whereas those who had smoked fewer than 100 cigarettes in their lifetime but did not currently smoke and never smoked in their lifetime were classified as non-smokers. BMI was calculated as weight in kilograms divided by height in meters squared. Weight change over 1 year was classified into three categories of no weight change, weight gain, and weight loss. Using the criteria specified in the questionnaire, weight gains and losses of 0 to less than 3 kg over 1 year were classified as no change, gains of more than 3 kg as weight gain, and losses of more than 3 kg as weight loss. Physical activity was calculated using the sum of the minutes of exercise per week and considering the strength of the exercise intensity (vigorous and moderate). The combination of vigorous intensity and moderate intensity exercise was considered by calculating the minutes exercised per week as follows: $2 \times \text{moderate activity (min/week)} + \text{vigorous activity (min/week)}$ [20]. "Physically active" is defined as engaging in at least 150 minutes of moderate activity per week. We used omega-3 or sodium intake in grams per day, and time spent sitting in minutes per day, and these variables were entered into the ML model as a continuous variable.

Statistical Analysis

We categorized ASCVD risk into two groups: high ($\geq 7.5\%$) and low risk. ASCVD risk directly incorporates demographic factors such as age and sex in its score calculation. Therefore, we performed 1:1 propensity score matching based on sociodemographic characteristics to assess the impact of lifestyle behaviors on ASCVD risk. The propensity score was calculated with the following variables: sex, age, job or marital status, residential area and household income. In descriptive analyses of two group comparison, continuous variables were presented as median and interquartile range, and categorical variables were presented as proportions (%). Differences between high and low ASCVD groups were compared using the Mann-Whitney *U* Test for continuous variables after normality test and the chi-square test for categorical variables.

We used five ML algorithms for the prediction of high ASCVD risk: logistic regression (LR), support vector machine (SVM), random forest (RF), extreme gradient boosting (XGB), and light gradient boosting (LGB) model. We divided the dataset into a training and a test dataset in an 8:2 ratio. Then, the training dataset was divided 5-fold, and cross-validation was performed by composing the same ratio of classes in the divided dataset. In the ML model, categorical variables were one-hot encoded, and continuous variables were scaled using min-max normalization. We also used bootstrapping to train the ML model with the area under the receiver operating characteristics curve (AUROC) score and resampled 1000 times to calculate the model's performance metrics for each iteration. [21]. The ML model was optimized using Optuna, a hyperparameter optimization framework based on the random search [21]. Understanding the performance of ML models is becoming increasingly important, which highlights the growing significance of explainable artificial intelligence in interpreting model results. We utilized the SHapley Additive exPlanations (SHAP) value to visualize the importance and relationships among the input features of the model. In addition to the AUROC score, accuracy, precision, recall, F1 score, specificity, area under the precision-recall curve (AUPRC) were calculated to assess the performance of the developed ML models.

We conducted a multivariable LR analysis as a sensitivity test to verify whether the ML model's findings on high ASCVD risk were consistent with conventional statistical methods. Univariable LR analysis was first performed on 1:1 matched patients (high vs. low ASCVD risk) to identify lifestyle and sociodemographic factors associated with high ASCVD risk. Variables with a *p*-value < 0.05 in univariable model were entered into the multivariable model, and their significance was reported as adjusted odds ratios (ORs) with 95% confidence intervals (CIs).

Results

Baseline Characteristics

Of the 8,573 participants, proportion of high ASCVD risk group was 41.7% (3578/8573) and 4995 (59.0%) participants were women (Table 1). Before PSM, demographic variables such as age and sex showed significant differences between the high and low ASCVD risk groups. The propensity score distribution and standardized differences before and after PSM for the variables used in propensity score calculation are presented in Figure S1 and S2 (Multimedia Appendix I). Using the 1:1 PSM method, 1976 patients were included in the final population for analysis, ensuring no significant differences in sociodemographic variables between the two groups, except for lifestyle behaviors. After PSM, the high ASCVD risk group had higher alcohol and tobacco consumption and lower omega-3 intake than the low ASCVD risk group. Additionally, they had a higher BMI, lower physical activity, and spent less time sitting compared to lower ASCVD risk group.

Table 1. Comparison of baseline characteristics of high and low ASCVD risk group before and after propensity score matching.

Variable	Before PSM ^a			After PSM		
	ASCVD ^b <7.5% (n=4995)	ASCVD ≥7.5% (n=3578)	<i>p</i>	ASCVD <7.5% (n=988)	ASCVD ≥7.5% (n=988)	<i>p</i>
Sex, women	3695 (74.0%)	1364 (38.1%)	<0.001	493(49.9%)	482(48.8%)	0.653
Age, median (IQR ^c)	51.0 (45.0–58.0)	68.0 (62.0–73.0)	<0.001	59.0 (54.0–65.0)	61.0 (52.0–67.0)	0.201
Job			<0.001			0.620
Non-manual workers	1886 (37.8%)	1543(43.1%)		435 (44.0%)	454 (46.0%)	
Other workers	1632 (32.7%)	1641(45.9%)		355 (35.9%)	350 (35.4%)	
Manual workers	1477 (29.6%)	394(11.0%)		198 (20.0)	184(18.6%)	
Marital status, without spouse	867 (17.4%)	922(25.8%)	<0.001	229 (23.2%)	235 (23.8%)	0.362
Residential area, urban	881 (17.6%)	1001(28.0%)	<0.001	899 (20.5%)	315 (46.2%)	
Income, 10,000 won/month	492 (292–700)	230(109–450)	<0.001	360 (200–614)	355 (193–600)	0.984
Alcohol drinking, ≥2/week	839 (16.8%)	873(24.4%)	<0.001	181 (18.3%)	250 (25.3%)	<0.001
Smoking	442 (8.8%)	915(25.6%)	<0.001	62 (6.3%)	417 (42.2%)	<0.001
Omega-3 intake, g/day	1.4 (0.8–2.3)	1.3 (0.7–2.3)	<0.001	1.5 (0.9–2.6)	1.3 (0.7–2.3)	<0.001
Sodium intake, g/day	2.8 (1.9–4.0)	2.9 (1.9–4.2)	0.020	3.1 (2.0–4.2)	2.9 (2.0–4.2)	0.155
Body mass index, kg/m ²	23.5 (21.4–25.9)	24.4 (22.5–26.5)	<0.001	23.9 (21.9–25.8)	25.1 (23.2–27.3)	<0.001
Weight change over 1 year			<0.001			0.144
Weight loss	528 (10.6%)	506 (14.1%)		116 (11.7%)	121 (12.2%)	
No weight change	3171 (63.5%)	2584(72.2%)		703 (71.2%)	666 (67.4%)	
Weight gain	1296 (25.9%)	488(13.6%)		169 (17.1%)	201 (20.3%)	
LDL ^d , mg/dL	120 (97–143)	110 (85–135)	<0.001	114 (91–139)	116 (92–143)	0.122

physical activity	1074 (21.5%)	534 (14.9%)	<0.001	217 (22.0%)	164 (16.6%)	0.003
Time spent sitting, hour/day	8.0 (5.0–10.0)	8.0 (5.8–11.0)	<0.001	8.0 (5.0–10.0)	8.0 (5.7–10.9)	0.003

Data are represented as median (interquartile change) or number (percent) as appropriate.

^aPSM: propensity score matching.

^bASCVD: atherosclerotic cardiovascular disease.

^cIQR: Interquartile range.

^dLDL: low-density lipoprotein.

Evaluation Outcomes Performance of ML Model for ASCVD Risk

The 1,972 matched patients were allocated to the training and test datasets in an 8:2 ratio while maintaining the same ASCVD risk proportion. A comparison of variables between these two groups is presented in Table S1 (Multimedia Appendix I), showing no significant differences in variable distribution. Figure 2 shows the result of ML performance for predicting high ASCVD risk ($\geq 7.5\%$) in the test dataset, with XGB showing the highest AUROC (0.811; 95% CI, 0.748–0.867). Table 2 presents the other performances of the five ML models, the LGB model demonstrated the best performance including accuracy, recall, and F1-score.

Table 2. Overall performances of 5 machine learning models for high ASCVD risk prediction on test dataset with propensity score matching.

Model	LR ^a		SVM ^b		RF ^c		XGB ^d		LGB ^e	
Accuracy	0.730	(0.672–0.788)	0.730	(0.672–0.788)	0.732	(0.672–0.793)	0.722	(0.662–0.783)	0.750	(0.692–0.808)
Precision	0.776	(0.701–0.853)	0.827	(0.750–0.901)	0.791	(0.714–0.870)	0.768	(0.693–0.846)	0.783	(0.711–0.855)
Recall	0.646	(0.556–0.737)	0.581	(0.485–0.677)	0.631	(0.535–0.727)	0.636	(0.535–0.727)	0.692	(0.596–0.778)
F1 score	0.705	(0.630–0.776)	0.682	(0.600–0.757)	0.702	(0.624–0.774)	0.696	(0.618–0.767)	0.735	(0.663–0.802)
AUROC ^f	0.809	(0.748–0.867)	0.800	(0.737–0.859)	0.786	(0.720–0.847)	0.811	(0.748–0.867)	0.810	(0.749–0.869)
Specificity	0.813	(0.737–0.889)	0.879	(0.808–0.939)	0.833	(0.758–0.909)	0.808	(0.727–0.879)	0.808	(0.727–0.879)
AUPRC ^g	0.811	(0.741–0.881)	0.803	(0.730–0.873)	0.805	(0.737–0.868)	0.828	(0.766–0.884)	0.829	(0.771–0.886)

Data are represented as averaged value with interquartile change.

^aLR, logistic regression.

^bSVM: support vector machine.

^cRF: random forest.

^dXGB: extreme gradient boosting.

^eLGB: light gradient boosting.

^fAUROC: area under the receiver operating characteristics curve.

^gAUPRC: area under the precision-recall curve.

Variable importance in ML model

To visualize the importance of variables included in the ML model, we presented SHAP values for the XGB model used to predict high ASCVD risk (Figure 3). Current smoking and age were the strongest predictors of high ASCVD risk over 10 years. Among lifestyle factors, BMI, sodium or omega-3 intake, and LDL cholesterol were also associated with increased risk. In contrast,

factors such as “time spent sitting,” physical activity, sex, and marital status had a relatively smaller impact. The SHAP values for the LGB model, presented in Figure S3 (Multimedia Appendix I), showed a similar pattern of variable importance as the XGB model.

Associations of lifestyle risk factors with 10-year ASCVD risk

For sensitivity analysis, we examined whether the ML model's results were consistent with binary LR (Table 3) using high ASCVD risk as the outcome. Multivariable binary LR showed that smoking, BMI, and LDL cholesterol were positively associated with high ASCVD risk, while omega-3 intake and physical activity were significantly negatively associated. Compared to the SHAP values of the XGB and LGB models, factors such as smoking status, BMI, and omega-3 intake showed similar associations in multivariable binary LR model.

Table 3. Result of binary logistic regression analysis of lifestyle variables for predicting 10-year high ASCVD risk in propensity score matched cohort.

Variable	Univariable analysis		Multivariable analysis	
	OR ^a (95% CI ^b)	<i>p</i>	Adjusted OR (95% CI)	<i>p</i>
Sex, women	0.96 (0.80–1.14)	0.621	-	
Age, years	1.01 (1.00–1.02)	0.201	-	
Job			-	
Manual workers	1.00 (reference)	-		
Non-manual workers	1.12 (0.88–1.43)	0.343		
Other workers	1.06 (0.83–1.36)	0.642		
Marital status, without spouse	1.03 (0.84–1.27)	0.751	-	
Residential area, urban	1.11 (0.90–1.37)	0.334	-	
Income, 10,000 won/month	1.00 (1.00–1.00)	0.984	-	
Alcohol drinking, ≥2/week	1.51 (1.22–1.87)	<0.001	0.89 (0.68–1.15)	0.369
Smoking	10.91 (8.19–14.53)	<0.001	11.52 (8.20–15.60)	<0.001
Omega-3 intake, g/day	0.93 (0.89–0.97)	0.003	0.93 (0.88–0.98)	0.011
Sodium intake, g/day	1.00 (1.00–1.00)	0.973	-	
Body mass index, kg/m ²	1.13 (1.10–1.17)	<0.001	1.13 (1.09–1.16)	<0.001
Weight change over 1 year			-	
Weight loss	1.00 (reference)	-		
No weight change	0.91 (0.69–1.20)	0.494		
Weight gain	1.14 (0.82–1.58)	0.431		
LDL ^c , mg/dL	1.00 (1.00–1.01)	0.039	1.00 (1.00–1.00)	0.675
physical activity	0.71 (0.56–0.89)	0.003	0.61 (0.47–0.79)	<0.001
Time spent sitting, hour/day	1.00 (1.00–1.00)	0.003	1.00 (1.00–1.00)	0.423

^aOR: odds ratio.

^bCI: confidence interval.

^cLDL: low-density lipoprotein.

Discussion

This study assessed the performance of ML models including XGB and LGB in predicting 10-year

high ASCVD risk in adults and identifying key risk factors. After adjusting for confounders using PSM, both models demonstrated high AUROC values, with LGB outperforming in accuracy, recall, and F1 score. Variable importance analysis identified smoking and age as the strongest predictors, while BMI, sodium or omega-3 intake, and LDL cholesterol also had significant variables. Multivariable LR also confirmed these findings, showing that smoking, BMI, and LDL cholesterol increased ASCVD risk, whereas omega-3 intake and physical activity were associated with lower risk.

Age is associated with increased oxidative stress, which leads to an increased susceptibility of CVD onset [22]. Excessive generation of reactive oxygen species leads to a state of oxidative stress which is a major risk factor for the development and progression of atherosclerosis [23]. Aging is an unmodifiable ASCVD risk factor, but it can be prevented if other factors could be corrected. Cigarette smoking is widely accepted as a major risk factor for the development of clinical CVD resulting from direct effects on atherosclerosis. Epidemiologic studies strongly support that cigarette smoking in both men and women increases the incidence of myocardial infarction and fatal coronary artery disease [24, 25].

Cigarette smoking is widely accepted as a major risk factor for the development of clinical CVD resulting from direct effects on atherosclerosis. Epidemiologic studies strongly support that cigarette smoking in both men and women increases the incidence of myocardial infarction and fatal coronary artery disease [26]. Because the chemical constituents of smoke have high oxidant and inflammatory capacities, they can directly induce endothelial damage and potentiate an inflammatory response [27]. Smoke is able to increase LDL levels through metabolic alterations and the induction of LDL oxidation due to the direct oxidant capacity of smoke components [27, 28]. Obesity has been linked to persistent inflammation and oxidative stress. Oxidative stress plays a crucial role in disorders related to obesity, such as dyslipidemia and hypertension, causing cardiovascular diseases [29]. The prevalence of overweight and obesity has been strongly increasing over the last few decades, and it is considered to be one of the largest challenges for public health work worldwide [30].

The American Heart Association has recently outlined a new framework that focuses on defining and optimizing cardiovascular health through the adoption of 8 simple health components: healthy diet, engaging in regular physical activity, avoidance of nicotine, healthy sleep, and healthy levels of blood lipids and glucose, and blood pressure [31]. Most of clinical guidelines for the primary or secondary prevention of ASCVD emphasize the importance of lifestyle modification. These lifestyle changes are also an important part of public health policy for CVD prevention. Therefore, a lifestyle approach to ASCVD risk may be important for accurate prediction and prioritization of treatment strategies for prevention.

This study has several limitations. First, it is a cross-sectional study, which does not allow for establishing causal relationships. Second, some variables were measured by questionnaire, which may lead to underestimation or overestimation for the prediction of ASCVD risk. Especially in the case of nutrition, it is particularly difficult to accurately reflect daily intake. Third, the study involved a secondary analysis of data from the KNHANES, so it was not possible to include all lifestyle and socioeconomic status variables in the analysis. Nevertheless, the strength of this study is that it used representative and reliable data from KNHANES to predict modifiable lifestyle risk factors predicted using an ML approach.

Conclusions

This study demonstrates that ML models are effective tools for assessing ASCVD risk and highlights the significant impact of lifestyle factors such as smoking, BMI, and omega-3 intake. These findings offer valuable insights for public health strategies aimed at preventing cardiovascular disease and reinforce the importance of lifestyle modifications in reducing ASCVD risk.

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HR21C0198), and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI); funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C1498) ; supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2022RIS-005) and supported by Hallym University Research Fund.

Authors' contributions

HJK designed the study, and prepared the manuscript. HC, HJA and SHS collected data, and literature search. CK drafted the manuscript. SHL, JHS and JJL data analysis. All authors carried out critical review. And approved the final manuscript.

Conflicts of Interest

None declared.

Abbreviations

ASCVD: atherosclerotic cardiovascular disease
ML: machine learning
KNHANES: Korea National Health and Nutrition Examination Survey
XGB: extreme gradient boosting
LGB: light gradient boosting
LR: logistic regression
SVM: support vector machine
RF: random forest
PSM: propensity score matching
AUROC: area under the receiver operating characteristic curve
LDL: low-density lipoprotein
BMI: body mass index
PCE: Pooled Cohort Equation
SHAP: SHapley Additive exPlanations
AUPRC: area under the precision-recall curve
RCT: randomized controlled trial

References

1. Weir HK, Anderson RN, King SMC, Soman A, Thompson TD, Hong Y, et al. Peer reviewed: heart disease and cancer deaths—trends and projections in the United States, 1969–2020.

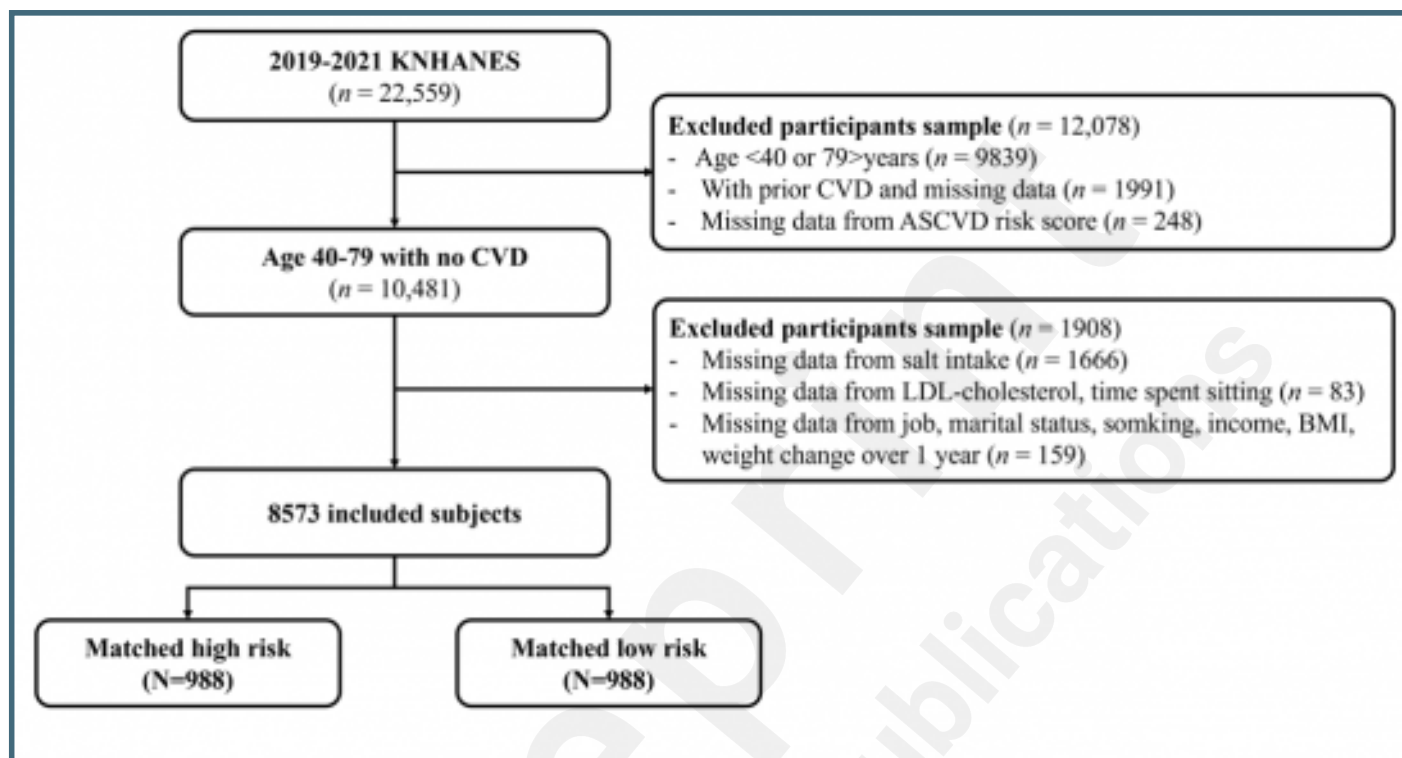
- Prev Chronic Dis. Nov, 2016;13:E157. doi: 10.5888/pcd13.160211.
2. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol*. 2020 Dec 22;76(25):2982-3021. doi: 10.1016/j.jacc.2020.11.010.
 3. Kim HC. Epidemiology of cardiovascular disease and its risk factors in Korea. *Glob Health Med*. 2021 Jun 30;3(3):134-141. doi: 10.35772/ghm.2021.01008.
 4. Lee HH, Cho SMJ, Lee H, Baek J, Bae JH, Chung WJ, Kim HC. Korea heart disease fact sheet 2020: analysis of nationwide data. *Korean Circ J*. 2021 Jun;51(6):495-503. doi: 10.4070/kcj.2021.0097.
 5. Poznyak AV, Sadykhov NK, Kartuesov AG, Borisov EE, Melnichenko AA, Grechko AV, Orekhov AN. Hypertension as a risk factor for atherosclerosis: Cardiovascular risk assessment. *Front Cardiovasc Med*. 2022 Aug 22;9:959285. doi: 10.3389/fcvm.2022.959285.
 6. Devareddy A, Sarraju A, Rodriguez F. Health disparities across the continuum of ASCVD risk. *Curr Cardiol Rep*. 2022 Sep;24(9):1129-1137. doi: 10.1007/s11886-022-01736-y.
 7. Mollborn S, Lawrence EM, Hummer RA. A gender framework for understanding health lifestyles. *Soc Sci Med*. 2020 Nov;265:113182. doi: 10.1016/j.socscimed.2020.113182. Epub 2020 Jul 29.
 8. Rajendran A, Minhas AS, Kazzi B, Varma B, Choi E, Thakkar A, Michos ED. Sex-specific differences in cardiovascular risk factors and implications for cardiovascular disease prevention in women. *Atherosclerosis*. 2023 Nov;384:117269. doi: 10.1016/j.atherosclerosis.2023.117269.
 9. Yusuf S, Hawken S, Ôunpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet*. 2004 Sep;364(9438):937-52. doi: 10.1016/S0140-6736(04)17018-9.
 10. Arnett DK, Khera A, Blumenthal RS. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: part 1, lifestyle and behavioral factors. *JAMA Cardiol*. 2019 Oct 1;4(10):1043-1044. doi: 10.1001/jamacardio.2019.2604.
 11. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS one*. Apr 4;12(4):e0174944. doi: 10.1371/journal.pone.0174944. eCollection 2017.
 12. Nakanishi R, Slomka PJ, Rios R, Betancur J, Blaha MJ, Nasir K, et al. Machine learning adds to clinical and CAC assessments in predicting 10-year CHD and CVD deaths. *JACC Cardiovasc Imaging*. 2021 Mar;14(3):615-625. doi: 10.1016/j.jcmg.2020.08.024. Epub 2020 Oct 28.
 13. Cho SY, Kim SH, Kang SH, Lee KJ, Choi D, Kang S, et al. Pre-existing and machine learning-based models for cardiovascular risk prediction. *Sci Rep*. 2021 Apr 26;11(1):8886. doi: 10.1038/s41598-021-88257-w.
 14. Connelly PJ, Azizi Z, Alipour P, Delles C, Pilote L, Raparelli V. The importance of gender to understand sex differences in cardiovascular disease. *Can J Cardiol*. 2021 May;37(5):699-710. doi: 10.1016/j.cjca.2021.02.005.
 15. Goff Jr DC, Lloyd-Jones DM, Bennett G, Coady S, D'agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2014 Jun 24;129(25 Suppl 2):S49-73. doi: 10.1161/01.cir.0000437741.48606.98.
 16. Wong ND, Budoff MJ, Ferdinand K, Graham IM, Michos ED, Reddy T, et al. Atherosclerotic cardiovascular disease risk assessment: An American Society for Preventive Cardiology clinical practice statement. *Am J Prev Cardiol*. 2022 Mar 15;10:100335. doi: 10.1016/j.ajpc.2022.100335. eCollection 2022 Jun.

17. Lloyd-Jones DM, Braun LT, Ndumele CE, Smith Jr SC, Sperling LS, Virani SS, Blumenthal RS. Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the American Heart Association and American College of Cardiology. *Circulation*. Jun 18;139(25):e1162-e1177. doi: 10.1161/CIR.0000000000000638. Epub 2018 Nov 10.
18. Mortensen MB, Fuster V, Muntendam P, Mehran R, Baber U, Sartori S, Falk E. A simple disease-guided approach to personalize ACC/AHA-recommended statin allocation in elderly people: the BioImage Study. *J Am Coll Cardiol*. 2016 Aug 30;68(9):881-91. doi: 10.1016/j.jacc.2016.05.084..
19. Lee HK, Lee BH. The epidemiology of alcohol use disorders. *J Korean Diab*. 2012 June 30;13(2): 69-75. doi: 10.4093/jkd.2012.13.2.69
20. Lopez JPR, Sabag A, Juan MM, Rezende LF, Pastor-Valero M. Do vigorous-intensity and moderate-intensity physical activities reduce mortality to the same extent? A systematic review and meta-analysis. *BMJ Open Sport Exerc Med*. 2020 Oct 5;6(1):e000775. doi: 10.1136/bmjsem-2020-000775.
21. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*; 2019. p. 2623-31.
22. Jebari-Benslaïman S, Galicia-García U, Larrea-Sebal A, Olaetxea JR, Alloza I, Vandenbroeck K, et al. Pathophysiology of atherosclerosis. *Int J Mol Sci*. 2022 Mar 20;23(6):3346. doi: 10.3390/ijms23063346.
23. Man JJ, Beckman JA, Jaffe IZ. Sex as a biological variable in atherosclerosis. *Circ. Res*. 2020 Apr 24;126(9):1297-1319. doi: 10.1161/CIRCRESAHA.120.315930.
24. Rodgers JL, Jones J, Bolleddu SI, Vanthenapalli S, Rodgers LE, Shah K, et al. Cardiovascular risks associated with gender and aging. *J Cardiovasc Dev Dis*. 2019 Apr 27;6(2):19. doi: 10.3390/jcdd6020019.
25. Batty M, Bennett MR, Yu E. The role of oxidative stress in atherosclerosis. *Cells*. Nov 30;11(23):3843. doi: 10.3390/cells11233843.
26. Howard G, Wagenknecht LE, Burke GL, Diez-Roux A, Evans GW, McGovern P, et al. Cigarette smoking and progression of atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) Study. *JAMA*. Jan 14;279(2):119-24. doi: 10.1001/jama.279.2.119
27. Gambardella J, Sardu C, Sacra C, Del Giudice C, Santulli G. Quit smoking to outsmart atherogenesis: Molecular mechanisms underlying clinical evidence. *Atherosclerosis*. 2017 Feb;257:242-245. doi: 10.1016/j.atherosclerosis.2016.12.010.
28. Haslam D. Obesity: a medical history. *Obesity reviews*. 2007 Mar;8 Suppl 1:31-6. doi: 10.1111/j.1467-789X.2007.00314.x.
29. Kelly T, Yang W, Chen C-S, Reynolds K, He J. Global burden of obesity in 2005 and projections to 2030. *Int J Obes*. 2008 Sep;32(9):1431-7. doi: 10.1038/ijo.2008.102.
30. Williams EP, Mesidor M, Winters K, Dubbert PM, Wyatt SB. Overweight and obesity: prevalence, consequences, and causes of a growing public health problem. *Curr Obes Rep*. 2015 Sep;4(3):363-70. doi: 10.1007/s13679-015-0169-4.
31. Kris-Etherton PM, Petersen KS, Després JP, Anderson CA, Deedwania P, Furie KL, et al. Strategies for promotion of a healthy lifestyle in clinical settings: pillars of ideal cardiovascular health: a science advisory from the American Heart Association. *Circulation*. 2021 Dec 14;144(24):e495-e514. doi: 10.1161/CIR.0000000000001018.

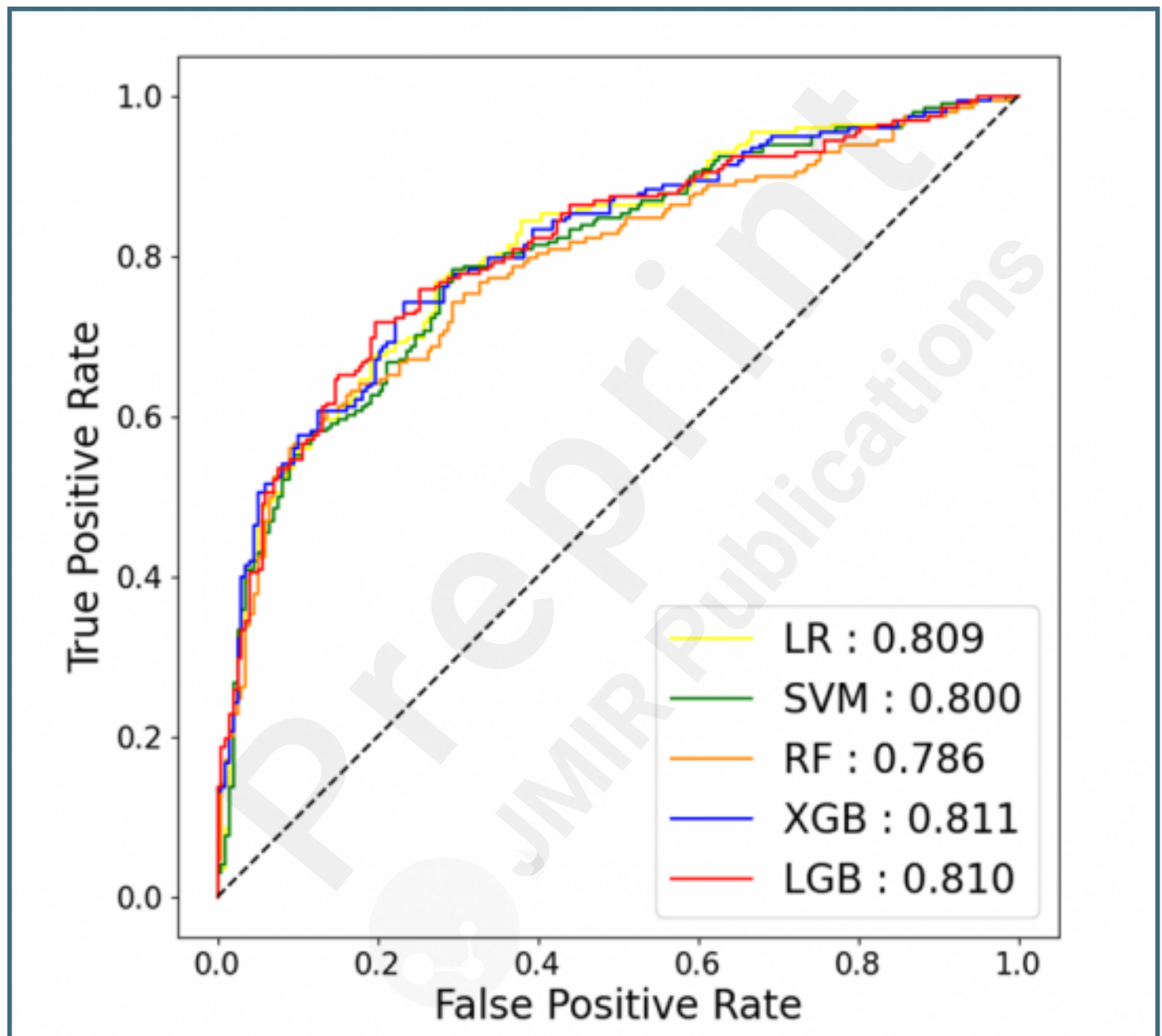
Supplementary Files

Figures

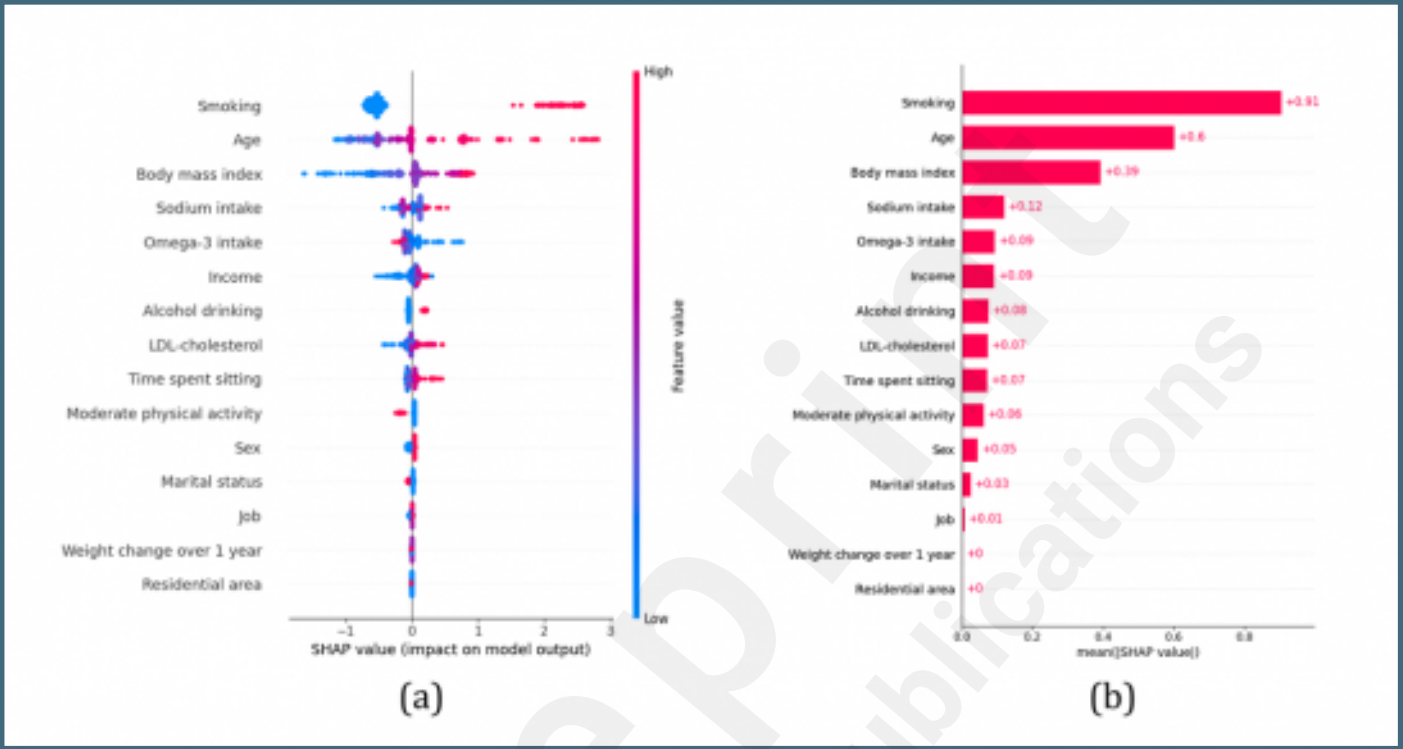
Flow chart of the participants. KNHANES: Korea National Health and Nutrition Examination Survey. CVD: cardiovascular disease. ASCVD: atherosclerotic cardiovascular disease. LDL: low-density lipoprotein. BMI: body mass index.



The area under the receiver operating characteristics curve values off 5 machine learning model for predicting high atherosclerotic cardiovascular disease risk (>7.5% of 10-year risk) in the test dataset. LR: logistic regression. SVM: support vector machine. RF: random forest. XGB: extreme gradient boosting. LGB: light gradient boosting.



Variable importances of extreme gradient boosting models for high atherosclerotic cardiovascular disease risk in the test dataset. (a) Importance matrix plot showing the direction of the relationship between a input variable and high atherosclerotic cardiovascular disease risk (b) Shapley Additive exPlanations (SHAP) summary plot of lifestyle variables predictive features of the ML model. As demonstrated by the colorbar, higher values are shown in red, while lower values are shown in blue.



Multimedia Appendixes

Untitled.

URL: <http://asset.jmir.pub/assets/cf6d272e905a62f3c67b4132b35fb416.pdf>

