Original Paper

Machine Learning Model for Predicting Coronary Heart Disease Risk: Development and Validation Using Insights From a Japanese Population–Based Study

Thien Vu[1,2,3], MD, PhD; Yoshihiro Kokubo[4], MD, PhD; Mai Inoue[1], ME; Masaki Yamamoto[1], BE; Attayeb Mohsen[1,5], MD, PhD; Agustin Martin-Morales[1], PhD; Research Dawadi[1], PhD; Takao Inoue[1,6], PhD; Jie Ting Tay[1], MRES; Mari Yoshizaki[1], PhD; Naoki Watanabe[1], PhD; Yuki Kuriya[1], PhD; Chisa Matsumoto[4,7], MD, PhD; Ahmed Arafa[4,8], MD, PhD; Yoko M Nakao[4], MD, PhD; Yuka Kato[4,9], MD, PhD; Masayuki Teramoto[4], MD, MPH; Michihiro Araki[1,10,11], PhD

[1]Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan
[2]NCD Epidemiology Research Center, Shiga University of Medical Science, Shiga, Otsu, Japan
[3]Department of Cardiac Surgery, Cardiovascular Center, Cho Ray Hospital, Ho Chi Minh City, Vietnam
[4]Department of Preventive Cardiology, National Cerebral and Cardiovascular Center, Suita, Osaka, Japan
[5]Libyan Centre for Dental Research, Zliten, Libya
[6]Faculty of Informatics, Yamato University, Osaka, Japan
[7]Department of Cardiology, Center for Health Surveillance and Preventive Medicine, Tokyo Medical University Hospital, Tokyo, Japan
[8]Department of Public Health, Faculty of Medicine, Beni-Suef University, Beni-Suef, Egypt
[9]Division of Health Sciences, Osaka University Graduate School of Medicine, Suita, Osaka, Japan
[10]Graduate School of Medicine, Kyoto University, Kyoto, Japan
[11]Graduate School of Science, Technology and Innovation, Kobe University, Kobe, Japan

Corresponding Author:

Thien Vu, MD, PhD
Artificial Intelligence Center for Health and Biomedical Research
National Institutes of Biomedical Innovation, Health and Nutrition
3-17 Senrioka-shinmachi
Osaka, 566-0002
Japan
Phone: 81 8093069457
Email: thienvuyd01@gmail.com

Abstract

Background: Coronary heart disease (CHD) is a major cause of morbidity and mortality worldwide. Identifying key risk factors is essential for effective risk assessment and prevention. A data-driven approach using machine learning (ML) offers advanced techniques to analyze complex, nonlinear, and high-dimensional datasets, uncovering novel predictors of CHD that go beyond the limitations of traditional models, which rely on predefined variables.

Objective: This study aims to evaluate the contribution of various risk factors to CHD, focusing on both established and novel markers using ML techniques.

Methods: The study recruited 7672 participants aged 30-84 years from Suita City, Japan, between 1989 and 1999. Over an average of 15 years, participants were monitored for cardiovascular events. A total of 7260 participants and 28 variables were included in the analysis after excluding individuals with missing outcome data and eliminating unnecessary variables. Five ML models—logistic regression, random forest (RF), support vector machine, Extreme Gradient Boosting, and Light Gradient-Boosting Machine—were applied for predicting CHD incidence. Model performance was evaluated using accuracy, sensitivity, specificity, precision, area under the curve, F_1-score, calibration curves, observed-to-expected ratios, and decision curve analysis. Additionally, Shapley Additive Explanations (SHAPs) were used to interpret the prediction models and understand the contribution of various risk factors to CHD.

Results: Among 7260 participants, 305 (4.2%) were diagnosed with CHD. The RF model demonstrated the highest performance, with an accuracy of 0.73 (95% CI 0.64-0.80), sensitivity of 0.74 (95% CI 0.62-0.84), specificity of 0.72 (95% CI 0.61-0.83), and an area under the curve of 0.73 (95% CI 0.65-0.80). RF also showed excellent calibration, with predicted

probabilities closely aligning with observed outcomes, and provided substantial net benefit across a range of risk thresholds, as demonstrated by decision curve analysis. SHAP analysis elucidated key predictors of CHD, including the intima-media thickness (IMT_cMax) of the common carotid artery, blood pressure, lipid profiles (non–high-density lipoprotein cholesterol, high-density lipoprotein cholesterol, and triglycerides), and estimated glomerular filtration rate. Novel risk factors identified as significant contributors to CHD risk included lower calcium levels, elevated white blood cell counts, and body fat percentage. Furthermore, a protective effect was observed in women, suggesting the potential necessity for gender-specific risk assessment strategies in future cardiovascular health evaluations.

Conclusions: We developed a model to predict CHD using ML and applied SHAP methods for interpretation. This approach highlights the multifactor nature of CHD risk evaluation, aiming to support health care professionals in identifying risk factors and formulating effective prevention strategies.

JMIR Cardio 2025;9:e68066; doi: 10.2196/68066

Keywords: coronary heart disease; machine learning; logistic regression; random forest; support vector machine; Extreme Gradient Boosting; Light Gradient-Boosting Machine; Shapley Additive Explanations; CHD; SVM; XGBoost; LightGBM; SHAP

Introduction

Coronary heart disease (CHD) remains a leading cause of morbidity and mortality worldwide, responsible for approximately 9.14 million deaths in 2019 [1,2]. Early identification of individuals at high risk is crucial, as timely interventions can significantly reduce the likelihood of severe outcomes like heart attacks and strokes. Studies have shown that early prediction and intervention can lead to a notable reduction in CHD-related mortality through preventive treatments such as statins and lifestyle changes [1-3]. While conventional risk assessment models have been used, there is growing recognition of the potential of machine learning (ML) in enhancing CHD prediction [4,5].

ML algorithms have proven their ability to analyze complex data and identify intricate patterns and relationships that are not easily detected by traditional statistical methods [6-10]. By integrating diverse data sources, such as demographics, medical history, lifestyle habits, and diagnostic findings, these algorithms can predict the likelihood of developing CHD. This approach offers comprehensive risk evaluation, adaptability to new data, and the potential to uncover novel risk factors and disease mechanisms [11].

Several studies have demonstrated the effectiveness of ML models in deriving quantitative markers for coronary artery disease and predicting the presence of heart disease. For example, a study developed and validated a coronary artery disease–predictive ML model using electronic health records and assessed its probabilities as in silico scores for coronary artery disease in participants in 2 longitudinal biobank cohorts [12]. Another study applied an ensemble ML model for coronary disease prediction, using ML classifiers to predict heart disease [13]. These findings highlight the potential of ML in driving innovation and improving the accuracy of CHD diagnosis and prediction [14].

However, challenges exist in utilizing ML for CHD prediction, including data quality, feature selection, model interpretability, and generalizability. These issues must be carefully addressed to ensure the reliability and robustness of the predictive models. Rigorous validation, regulatory compliance, and effective communication strategies are essential for its successful integration into clinical practice.

While several established CHD prediction models rely on traditional statistical techniques with predefined risk factors, they are limited by linear assumptions and struggle with complex, high-dimensional datasets. This restricts their ability to uncover novel or subtle risk factors. In contrast, ML models can handle these complexities, offering more nuanced and accurate predictions by identifying nonlinear interactions and discovering previously overlooked factors. Therefore, ML may enhance the overall understanding of CHD and improve both risk assessment and prevention strategies.

This study aimed to address the role of ML techniques in predicting incident CHD and identifying novel risk factors. This study sought to deepen our understanding of the factors contributing to CHD development by analyzing a comprehensive dataset. These findings will enhance risk assessment, enabling the development of personalized interventions and preventive strategies.

Methods

Study Design and Participants

The Suita Study, a prospective population-based cohort study, was conducted in Suita City, Osaka, Japan. From 1989 to 1999, a total of 7672 men and women aged 30-84 years who did not have a previous history of cardiovascular disease were recruited for the study. Participants were selected from the population registry of the municipality and were followed up every 2 years for an average of 15 years until their first occurrence of stroke, myocardial infarction (MI), death, or relocation.

After excluding participants with missing outcome data and removing unnecessary variables, the analysis included 7260 participants and 28 variables. Opt-out procedures were implemented for those who preferred not to participate in this study. Informed consent was obtained from all participants at the time of enrollment. The study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis and Artificial Intelligence

(TRIPOD+AI) statement guidelines for reporting prediction models in medicine, and we have added the completed checklist in Checklist 1 [15].

Ethical Considerations

The study was conducted in compliance with the ethical standards outlined in the Declaration of Helsinki, and approval was granted by the Institutional Review Board at the National Cerebral and Cardiovascular Center (approval R21024-2). As this study involves secondary data analysis, it is important to note that the original informed consent, obtained during the primary data collection, permits the use of the data for secondary analyses without requiring additional consent from participants. Participants' privacy was protected by anonymizing or deidentifying the data to prevent identification.

Outcome

The primary outcome was CHD, including MI, sudden death within 24 hours of acute illness onset, and coronary artery disease requiring bypass surgery or intervention. Medical records were carefully reviewed by hospital doctors or researchers who were blinded to the baseline data to provide an unbiased approach to the analysis. MIs were classified as definite or probable according to the criteria established by the MONICA Project [16].

Every 2 years, each participant's health was evaluated at the National Cerebral and Cardiovascular Center in Osaka, Japan, to detect the occurrence of CHD. Yearly questionnaires were also completed by all participants by mail or telephone. CHD surveillance was completed by systematically searching for death certificates [17,18].

Predictors

Predictors were measured at baseline and processed according to a standardized protocol. A comprehensive and prospective data collection process was implemented, encompassing various aspects such as demographics, medical history, medical imaging, laboratory data, lifestyle habits, and outcomes.

Blood Pressure and Physical Measurements

Blood pressure was measured in each participant using a mercury column sphygmomanometer, an appropriately sized cuff, and a standardized protocol to ensure accuracy and precision [17]. Before the initial blood pressure reading, the participants were instructed to rest for at least 5 minutes to establish a stable baseline. Blood pressure readings were obtained by averaging the second and third measurements, which were performed at intervals of more than 1 minute to allow for adequate observation and recording. Hypertension was defined as systolic blood pressure ≥140 mm Hg, diastolic blood pressure ≥90 mm Hg, or the use of antihypertensive medications.

BMI was calculated as weight (kg) divided by the square of height (m^2).

Biochemical Measurements

At baseline, routine blood tests were conducted, including measurements of total cholesterol, high-density lipoprotein cholesterol (HDL-c), and fasting glucose levels. Non-HDL-c was calculated by subtracting HDL-c from total cholesterol. Diabetes mellitus was diagnosed if participants had fasting plasma glucose (FPG) ≥126 mg/dL, a non-FPG ≥200 mg/dL, or the use of diabetes mellitus medication.

The estimated glomerular filtration rate (eGFR; mL/min/1.73 m^2) was calculated according to the original Modification of Diet in Renal Disease equation modified by the Japanese coefficient (0.881) as follows: $0.881 \times 186 \times \text{serum creatinine}^{-1.154} \times \text{age}^{-0.203} \times (0.742 \text{ if female})$ [19].

Imaging Diagnostics

Carotid artery measurements were performed using a high-resolution ultrasound machine to assess atherosclerotic indices, specifically intima-media thickness (IMT), on both sides of the common carotid artery (CCA), carotid artery bulb, internal carotid artery, and external carotid artery. The maximum IMT in the CCA (IMT_cMax) was defined as the highest measurable IMT in the scanned CCA regions, while the maximum IMT (IMT_MAX) was the highest measurable IMT across the entire scanned area, including the CCA, bulb, internal carotid artery, and external carotid artery on both sides [20].

Atrial fibrillation was checked by standard 12-lead ECGs from all participants and was determined by well-trained physicians [18].

Lifestyle and Medical History

Smoking status and drinking statuses were categorized as current, quit, or never. A questionnaire was used to ask participants about their past and present history of CHD.

Sample Size

All available data were used, and no formal sample size calculation was performed. The dataset included 7260 participants, among whom 305 had CHD, with 28 predictors selected after the feature selection process used in the model. Based on the events per predictor ratio, which is approximately 10.89 (305/28), the sample size is sufficient to ensure model stability and reliability [21,22]. Therefore, this dataset is adequate to answer the research questions.

Missing Data

Missing data analysis was conducted, and variables with more than 30% missing values were excluded to enhance model robustness. Missing data were imputed using Multivariate Imputation by Chained Equations. See Multimedia Appendix 1 for details on the percentage of missing data for each variable before imputation.

Statistical Analysis Methods

Descriptive Analysis

Continuous variables were summarized using means and SDs for normally distributed data, or medians and IQR for nonnormally distributed data. Categorical variables were reported as frequencies and percentages. To compare differences in patient characteristics based on CHD incidence (yes or no), we used various statistical tests including 2-tailed Student *t* tests, Mann-Whitney *U* tests, or chi-square tests, as appropriate.

Feature Selection

Feature selection was conducted in a stepwise manner to ensure that only the most relevant variables were included in the predictive models. Initially, variables with more than 30% missing data were excluded to avoid potential bias from imputation. Following this, a correlation matrix was used to identify and remove variables with high multicollinearity, defined as having a correlation coefficient greater than 0.8. See correlation coefficients heat map in the Multimedia Appendix 2 for details. The next step involved applying the least absolute shrinkage and selection operator regression. This technique shrinks the coefficients of less significant predictors toward zero, effectively removing them from the model, and was performed using cross-validation to identify the most important features based on the data. Finally, after statistical feature selection, medical knowledge was applied to confirm the clinical relevance of the remaining variables. Important predictors such as age, glucose levels, HDL-c, and blood pressure were retained, given their established association with CHD. The list of variables (predictors) used for model development was described in Multimedia Appendix 3.

Development of ML Models

Overview

The goal of this analysis was to predict the incidence of CHD using ML models and examine the contribution of each risk factor to the CHD incidence. A comprehensive process was followed, which included descriptive analysis, feature selection, model training, hyperparameter optimization, and interpretability through Shapley Additive Explanation (SHAP) values.

To manage the imbalance between CHD and non-CHD cases, we used down sampling on the majority class (non-CHD) to create a balanced dataset. This approach helps to ensure that the models do not disproportionately favor the majority class during training, improving prediction performance on the minority class.

The dataset was split into training (80%) and testing (20%) sets while maintaining balanced target variable distributions across both. Next, one-hot encoding was applied to convert categorical variables into a binary format, and normalization was performed to scale numerical features.

Several ML algorithms were implemented to compare their predictive power. Logistic regression (LR) was used as a baseline model, offering simplicity and interpretability [23]. Random forest (RF), an ensemble learning method, was used due to its strength in handling high-dimensional data and offering feature-importance insights [8,24]. Support vector machines (SVMs) with radial basis kernels were used for their effectiveness in nonlinear classification tasks [25,26]. Extreme Gradient Boosting (XGBoost) is an ML algorithm that improves model performance by using a series of decision trees, where each tree corrects the mistakes of the previous one. This sequential approach helps make predictions more accurate. Light Gradient-Boosting Machine (LightGBM) is another efficient algorithm that works similarly to XGBoost but is designed to be faster and more scalable, especially when working with large datasets and many features. Both algorithms are known for their high performance in handling complex data and large-scale problems [9,27].

Model Evaluation

We used 5-fold cross-validation during model training to ensure robustness and mitigate overfitting. Hyperparameter optimization was conducted using a grid search approach. The model's performance on the testing set was evaluated using 5 metrics: accuracy, sensitivity, specificity, precision, area under the curve (AUC), and F_1-score [15].

Calibration plots are used to evaluate the predictive accuracy of ML models in estimating CHD incidence. Calibration measures how closely the predicted absolute risk corresponds to the observed (true) risk across groups of patients categorized into different risk levels. The overall observed-to-expected (OE) ratio is calculated by dividing the total observed events by the total expected events for the entire population. For each decile, the OE ratio is determined by dividing the observed events within that decile by the expected events for the same decile. An ideal model is represented by a straight line bisecting the calibration plot, with an OE ratio of 1, indicating perfect calibration. An OE ratio <1 suggests overprediction, while a ratio >1 indicates underprediction [15].

Decision curve analysis (DCA) assesses the clinical use of ML models for predicting CHD incidence. DCA uses net benefit as a metric, reflecting the tradeoff between true-positive and false-positive predictions for a specific strategy [15,28,29].

Model Interpretation

SHAP is a method used in ML to make the predictions of a model more understandable. It helps explain how each input feature (such as age, cholesterol levels, or blood pressure) affects the model's decision. Essentially, SHAP breaks down the prediction to show how much each feature contributes to the final result, allowing us to see which factors are most important for predicting a condition like CHD [8-10,30]. SHAP summary plots visualized the importance of key features, while SHAP dependence plots highlighted the non-linear relationships between features and CHD incidence.

Results

Study Participants' Characteristics

In this study, 7260 participants were analyzed, of which 305 (4.2%) were diagnosed with CHD. The median age of participants with CHD was 63 (IQR 56-71) years , which was significantly older than that of those without CHD, whose median age was 55 (IQR 44-65) years. CHD was more prevalent in men (n=202, 66.2%) compared to women (n=103, 33.8%), and this gender difference was statistically significant.

Several cardiovascular risk factors were also associated with CHD. Participants with CHD had higher systolic and diastolic blood pressures. The eGFR was lower in participants with CHD compared to those without. The IMT of CCAs, IMT_cMax, was also significantly higher in patients with CHD (1.10 mm vs 1.00 mm; *P*<.001).

BMI and waist circumference were also higher in participants with CHD, indicating a greater degree of obesity. Additionally, lipid profiles showed significant differences, with lower HDL-c levels and higher non-HDL-c and triglyceride levels in patients with CHD.

Higher glucose levels and white blood cell counts were observed in participants with CHD, along with elevated hemoglobin levels. Regarding lifestyle factors, smoking was more common in those with CHD, while drinking status did not differ significantly between the 2 groups.

Regarding lifestyle factors, current smoking was more prevalent among participants with CHD (36.1% vs 29.0%; *P*<.001), while drinking status did not significantly differ between the groups.

In terms of comorbidities, atrial fibrillation, hypertension, diabetes mellitus, and dyslipidemia were all significantly more common in participants with CHD, as outlined in Table 1.

Table 1. Characteristics of study participants with and without CHD[a] incidence (Japanese participants aged 30-84 years, Suita Study). CHD was diagnosed by a first-ever acute myocardial infarction, sudden cardiac death within 24 hours of illness, or coronary artery disease followed by bypass or angioplasty. Values are presented as mean (SD) for continuous variables with approximately normally distribution or by median (IQR) with skewed distribution and n (%) for categorical variables. Differences in characteristics were evaluated by using the unpaired 2-tailed Student *t* test, Wilcoxon rank sums test, or chi-square test.

	CHD		*P* value
	No (n=6955)	Yes (n=305)	
Age (years), median (IQR)	55.0 (44.0-65.0)	63.0 (56.0-71.0)	<.001
Sex, n (%)			<.001
Male	3147 (45.2)	202 (66.2)	
Female	3808 (54.8)	103 (33.8)	
SBP[b] (mm Hg), median (IQR)	123 (110-137)	138 (125-153)	<.001
DBP[c] (mm Hg), median (IQR)	77.0 (70.0-85.0)	83.0 (74.0-89.0)	<.001
IMT_cMax[d] (mm), median (IQR)	1.00 (0.80-1.10)	1.10 (1.00-1.30)	<.001
eGFR[e] (mL/min/1.73 m^2), mean (SD)	104 (32.2)	95.3 (63.3)	.014
BMI (kg/m^2), mean (SD)	22.5 (3.10)	23.3 (3.26)	<.001
Body fat (%), mean (SD)	23.2 (6.32)	22.6 (7.06)	.15
Waist circumference (cm), median (IQR)	80.0 (73.0-86.0)	83.0 (77.0-90.0)	<.001
HDL-c[f] (mg/dL), median (IQR)	53.0 (44.0-63.0)	46.0 (38.0-56.0)	<.001
non-HDL-c (mg/dL), mean (SD)	152 (36.9)	172 (40.5)	<.001
Triglycerides (mg/dL), median (IQR)	98.0 (70.0-143)	121 (90.0-174)	<.001
Calcium (mg/dL), mean (SD)	9.35 (0.46)	9.34 (0.43)	.61
Fructosamine (μmol/L), median (IQR)	251 (237-266)	257 (242-276)	<.001
Glucose (mg/dL), median (IQR)	95.0 (89.0-101)	100 (93.0-109)	<.001
WBC[g] count (/mm^3), median (IQR)	5.33 (4.48-6.36)	5.65 (4.81-6.78)	<.001
RBC[h] count (10^3/mm^3), mean (SD)	4.53 (0.44)	4.60 (0.46)	.008
Smoking status, n (%)			<.001
Current	2019 (29)	110 (36.1)	
Past	1091 (15.7)	79 (25.9)	
Never	3845 (55.3)	116 (38)	
Drinking status, n (%)			.27

	CHD		*P* value
	No (n=6955)	Yes (n=305)	
Current	3613 (51.9)	152 (49.8)	
Past	156 (2.24)	11 (3.61)	
Never	3186 (45.8)	142 (46.6)	
Atrial fibrillation, n (%)	123 (1.77)	20 (6.56)	<.001
Hypertension, n (%)	2056 (29.6)	172 (56.4)	<.001
Diabetes mellitus, n (%)	426 (6.13)	49 (16.1)	<.001
Dyslipidemia, n (%)	5280 (75.9)	265 (86.9)	<.001

[a]CHD: coronary heart disease.
[b]SBP: systolic blood pressure.
[c]DBP: diastolic blood pressure.
[d]IMT_cMax: maximum intima-media thickness of common carotid arteries.
[e]eGFR: estimated glomerular filtration rate.
[f]HDL-c: high-density lipoprotein cholesterol.
[g]WBC: white blood cell.
[h]RBC: red blood cell.

Model Performance

The performance metrics of the 5 ML models used in our CHD prediction study provide valuable insights into their effectiveness, as shown in Table 2.

Table 2. Performance metrics and 95% CIs for machine learning models predicting CHD[a] incidence (Japanese participants, aged 30-84 years, Suita Study).

Model	Accuracy	Sensitivity	Specificity	Precision	AUC[b]	F_1-score
LR[c]	0.66 (0.58-0.75)	0.59 (0.46-0.71)	0.74 (0.62-0.84)	0.69 (0.55-0.81)	0.66 (0.57-0.75)	0.64 (0.52-0.73)
RF[d]	0.73 (0.64-0.80)	0.74 (0.62-0.84)	0.72 (0.61-0.83)	0.73 (0.61-0.84)	0.73 (0.65-0.80)	0.73 (0.64-0.82)
SVM[e]	0.71 (0.62-0.80)	0.70 (0.59-0.81)	0.72 (0.62-0.83)	0.72 (0.60-0.84)	0.71 (0.63-0.79)	0.71 (0.61-0.80)
XGBoost[f]	0.72 (0.64-0.80)	0.74 (0.63-0.84)	0.70 (0.58-0.82)	0.71 (0.60-0.82)	0.72 (0.64-0.80)	0.73 (0.63-0.81)
LightGBM[g]	0.50 (0.43-0.58)	1.00 (1.00-1.00)	0.00 (0.00-0.00)	0.50 (0.41-0.59)	0.5 (0.49-0.57)	0.67 (0.58-0.74)

[a]CHD: coronary heart disease.
[b]AUC: area under the curve.
[c]LR: logistic regression.
[d]RF: random forest.
[e]SVM: support vector machine.
[f]XGBoost: Extreme Gradient Boosting.
[g]LightGBM: Light Gradient-Boosting Machine.

RF emerged as the strongest model for CHD prediction in this study, achieving the highest overall performance with an accuracy of 0.73 (95% CI 0.64-0.80), sensitivity of 0.74 (95% CI 0.62-0.84), specificity of 0.72 (95% CI 0.61-0.83), and an AUC of 0.73 (95% CI 0.65-0.80). These results highlight its balanced ability to identify both CHD and non-CHD cases effectively. In comparison, XGBoost delivered robust, yet slightly inferior, results with an accuracy of 0.72 (95% CI 0.64-0.80), sensitivity of 0.74 (95% CI 0.63-0.84), specificity of 0.70 (95% CI 0.58-0.82), an AUC of 0.72 (95% CI 0.64-0.80), and an F_1-score of 0.73 (95% CI 0.63-0.81). SVM demonstrated competitive performance, achieving an AUC of 0.71 (95% CI 0.63-0.79), but ranked slightly behind RF and XGBoost. In contrast, LightGBM, despite its perfect sensitivity of 1.00 (95% CI 1.00-1.00), showed a specificity of 0.00 (95% CI 0.00-0.00) and an AUC of 0.50 (95% CI 0.49-0.57), rendering it unsuitable for this task. LR, while serving as a baseline model, exhibited moderate performance with an accuracy of 0.66 (95% CI 0.58-0.75), sensitivity of 0.59 (95% CI 0.46-0.71), specificity of 0.74 (95% CI 0.62-0.84), and an AUC of 0.66 (95% CI 0.57-0.75), but lacked the sensitivity required for effective CHD prediction.

The calibration curves for the 5 models (Figure 1) and the OE ratios by decile (Figure 2) provide critical insights into their predictive reliability. Among the models, RF demonstrated excellent calibration, with predicted probabilities closely aligning with observed outcomes across all deciles. This strong calibration is complemented by its performance in DCA (Figure 3).

Figure 1. Calibration plots for machine learning models predicting CHD incidence (Japanese participants, aged 30-84 years, Suita Study). CHD: coronary heart disease; LightGBM: Light Gradient-Boosting Machine; XGBoost: Extreme Gradient Boosting.

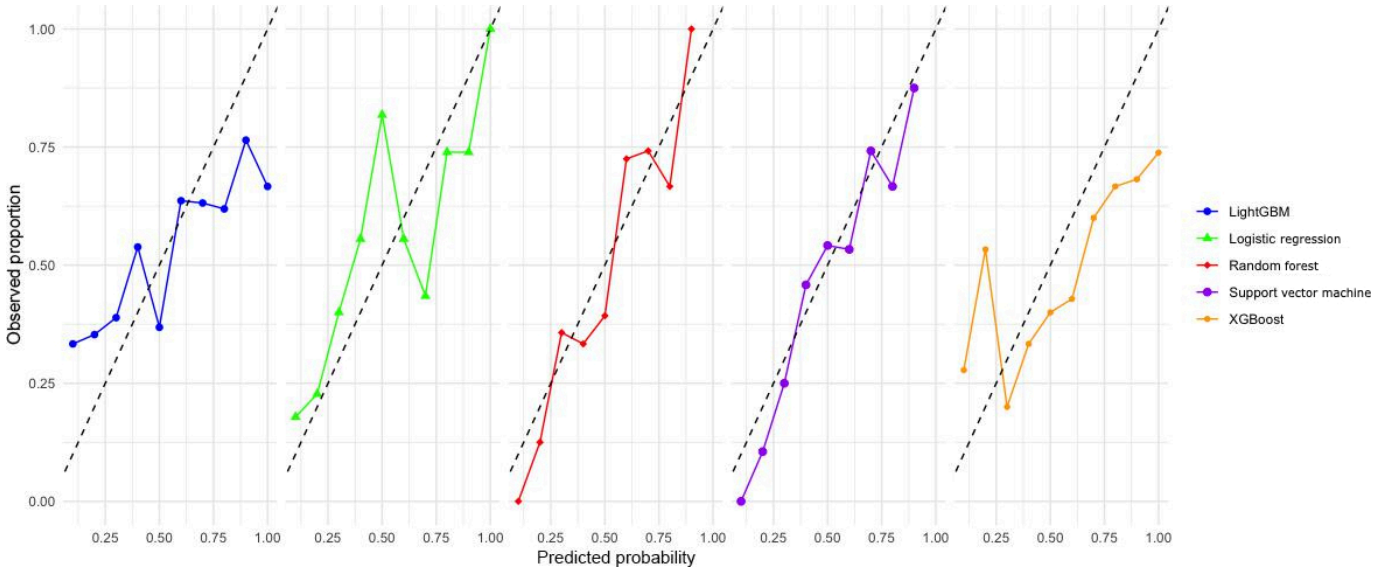


Figure 2. Calibration plots displaying observed-to-expected ratios for each decile of predicted CHD incidence risk (Japanese participants, aged 30-84 years, Suita Study). CHD: coronary heart disease; LightGBM: Light Gradient-Boosting Machine; SVM: support vector machine; XGBoost: Extreme Gradient Boosting.

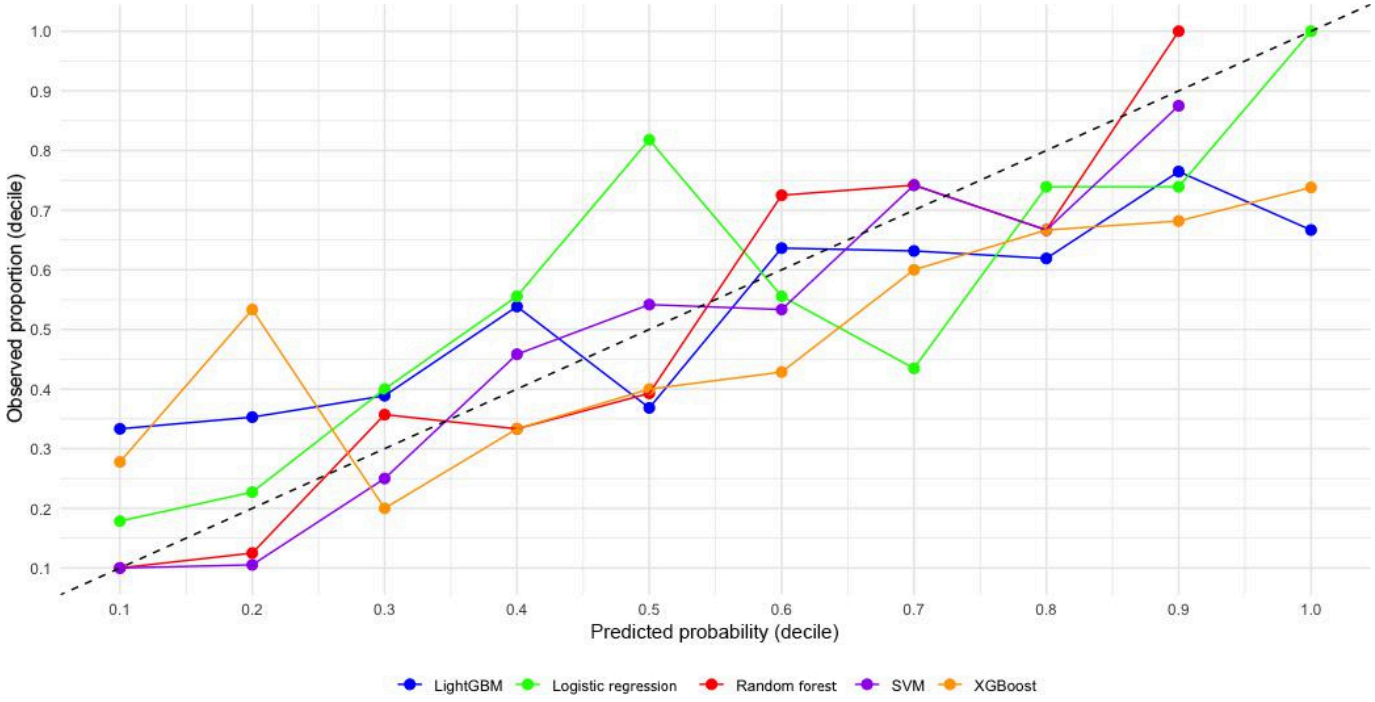
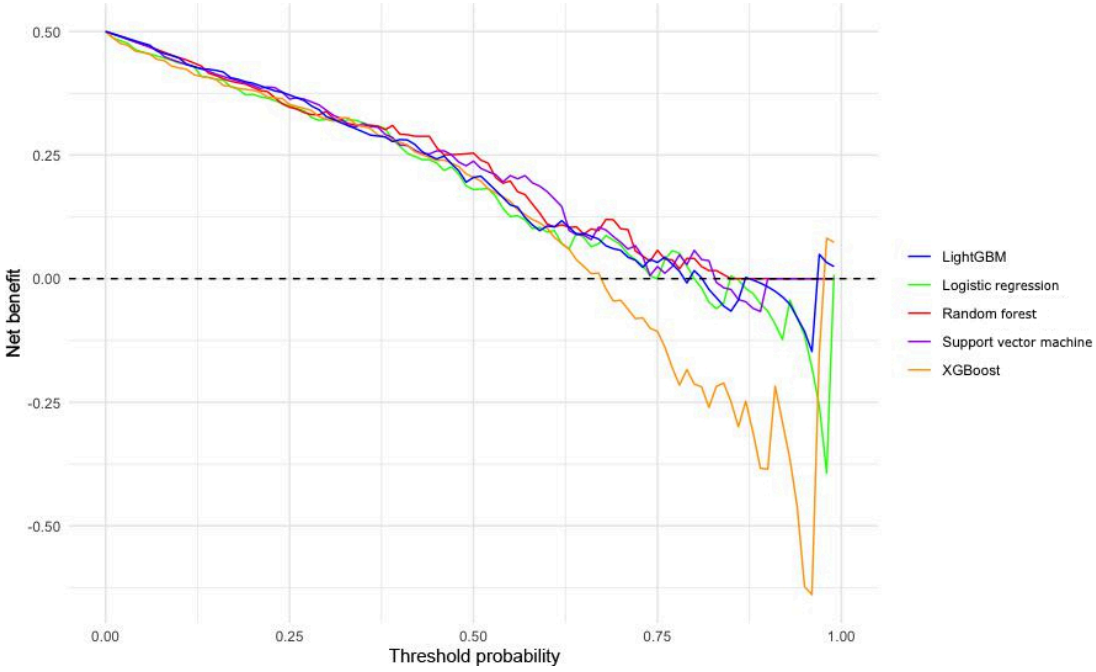


Figure 3. Decision curve analysis comparing machine learning models for predicting CHD incidence (Japanese participants, aged 30-84 years, Suita Study). CHD: coronary heart disease; LightGBM: Light Gradient-Boosting Machine; XGBoost: Extreme Gradient Boosting.



In terms of clinical use, as illustrated in Figure 3, all models exhibit a similar positive net benefit when the threshold is below 0.5, meaning that using the predictive models is better than not using any model (treat none). However, when the threshold exceeds 0.5, the models tend to decline rapidly, with LR and XGBoost showing the most pronounced decrease, declining earlier compared to the other models.

Based on the performance metrics, RF emerges as the best model for CHD prediction in this study due to its highest overall accuracy, balanced sensitivity and specificity, strong AUC, excellent calibration, and robust clinical use across various threshold probabilities.

Model Interpretation

In Figure 4, the bar plot on the left ranks the top features contributing to CHD prediction, with IMT_cMax identified as the most influential variable, followed by systolic blood pressure (SBP), HDL-c, non-HDL-c, and eGFR. This ranking emphasizes the significance of arterial health, blood pressure regulation, lipid levels, and kidney function in assessing CHD risk. The SHAP summary heat plot on the right provides a detailed visualization of how each feature influences individual model predictions. It shows that higher values of IMT_cMax, non-HDL-c, and blood pressure are positively associated with an increased likelihood of CHD, whereas lower levels of protective factors like HDL-c and eGFR are associated with a higher risk of CHD. Other important variables, such as age, glucose levels, and triglycerides, also contribute significantly, with older age and impaired glucose regulation being linked to a higher CHD risk. Additionally, markers of inflammation like white blood cell count and other factors such as calcium levels, sex, body fat, and BMI play roles in determining CHD risk.

Figure 5 consists of several SHAP dependency plots that illustrate the relationship between each key variable and CHD risk in more detail. For IMT_cMax, there is a positive association with CHD risk, showing that as the thickness of the carotid artery increases, so does the risk of CHD. The eGFR plot shows that lower eGFR values are associated with a higher risk of CHD, while higher eGFR values are associated with a lower risk, indicating the crucial role of kidney function in cardiovascular health. Non-HDL-c shows a generally positive association with CHD, where higher levels correspond to a higher risk. For SBP, the risk of CHD increases sharply with rising SBP values. HDL-c is inversely related to CHD risk, indicating its protective role, while higher triglycerides (TG) are linked to increased risk, especially at moderate levels. Age and glucose levels show a direct relationship with CHD risk, whereas older age and higher glucose levels are associated with increased risk. The SHAP value for diastolic blood pressure (DBP) also shows a positive relationship, suggesting that higher DBP levels contribute to the increased risk of CHD.

Figure 4. Contribution of variables to CHD incidence prediction using SHAP values (Japanese participants, aged 30-84 years, Suita Study). (A) The bar plot shows each variable's contribution to CHD, with bar length indicating the contribution extent. (B) The heat plot of SHAP values illustrates the relationships between variables and CHD. Purple signifies a positive relationship and yellow a negative one. Each point represents a participant, with the x-axis showing SHAP values and the y-axis indicating variable importance. bf: body fat; Ca: calcium; CHD: coronary heart disease; DBP: diastolic blood pressure; eGFR: estimated glomerular filtration rate; Frct: Fructosamine; Hb: hemoglobin; htn: hypertension; IMT_cMax: maximum intima-media thickness of common carotid arteries; HDL-c: high-density lipoprotein cholesterol; SBP: systolic blood pressure; smk_sts: smoking status; TG: triglycerides; WBC: white blood cell; wt20: weight at age of 20 years.

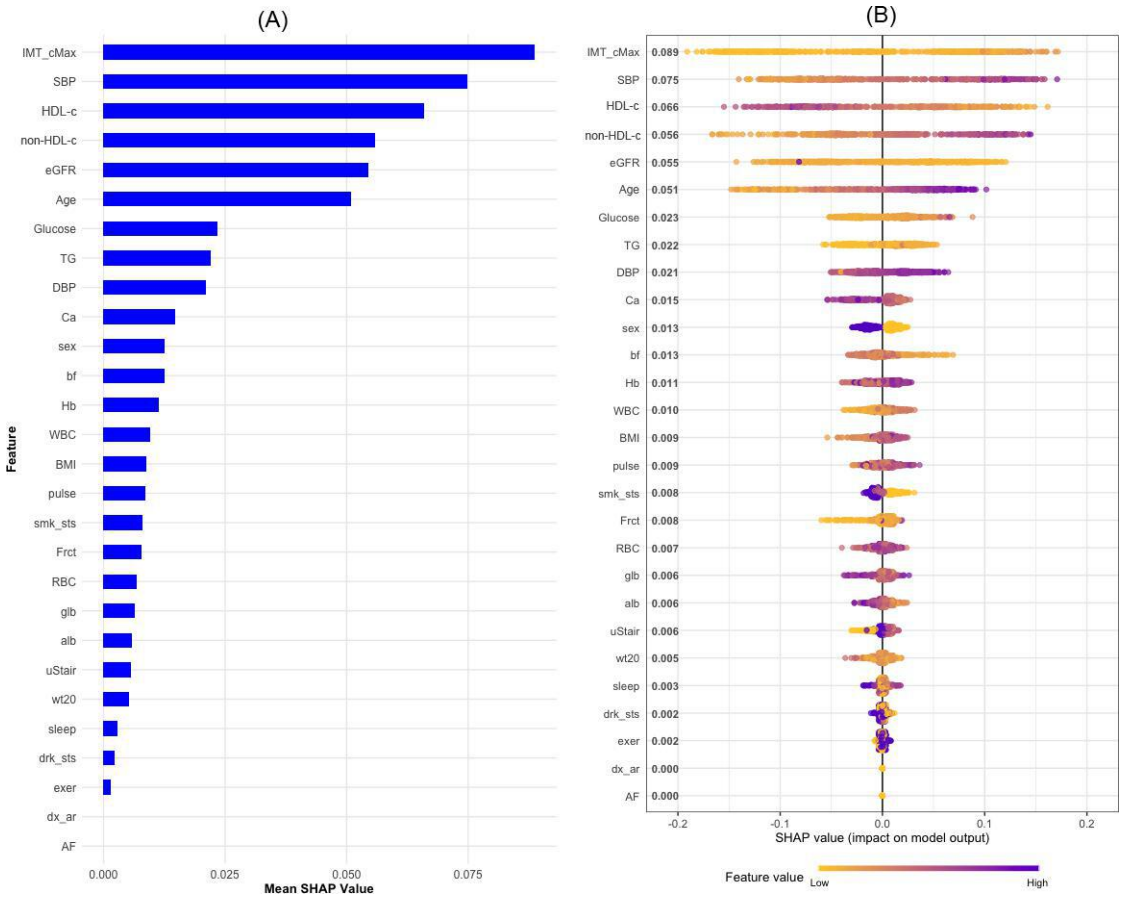
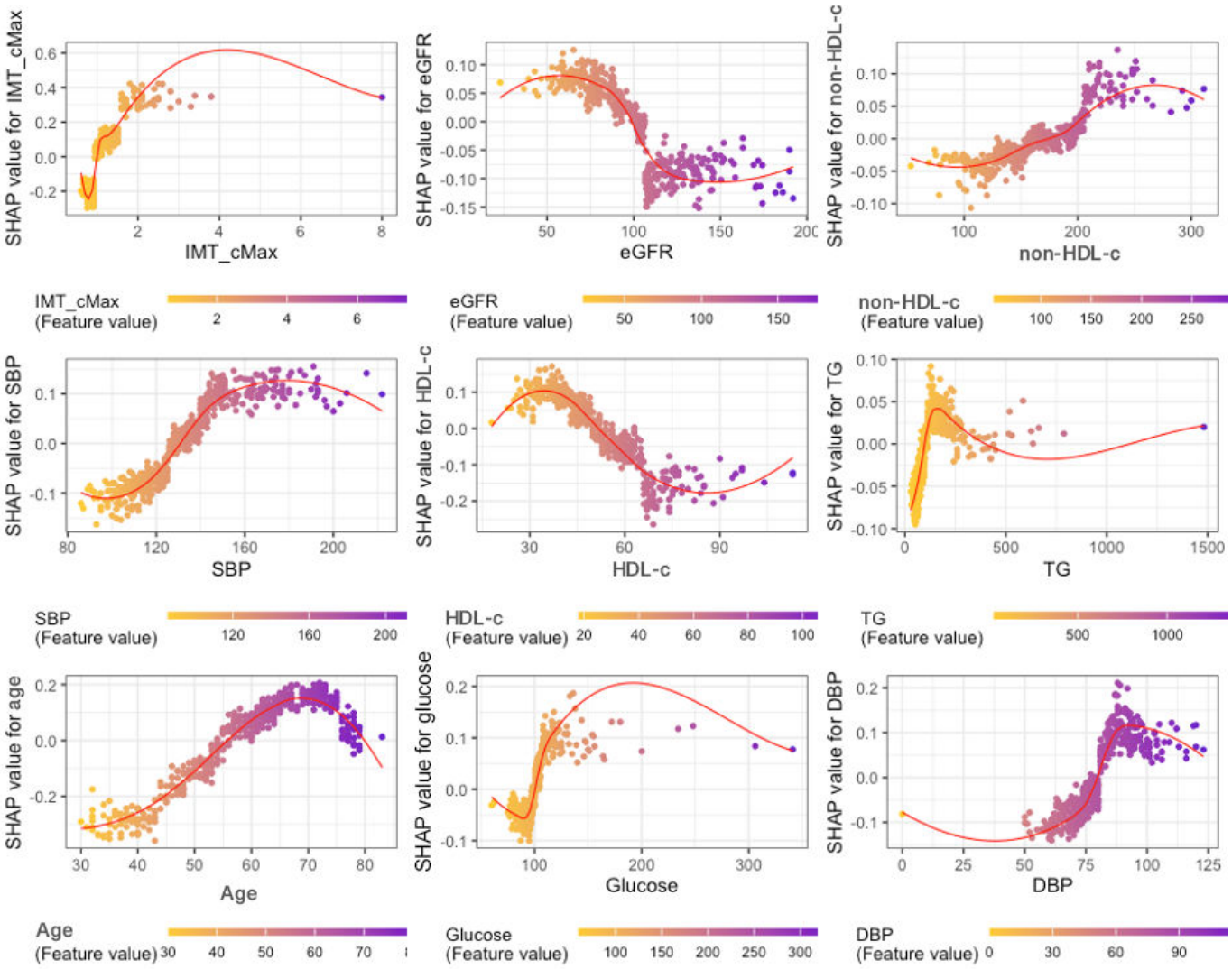


Figure 5. SHAP dependency plots showing the relationship between key variables and CHD risk (Japanese participants, aged 30-84 years, Suita Study). CHD: coronary heart disease; DBP: diastolic blood pressure; eGFR: estimated glomerular filtration rate; IMT_cMax: maximum intima-media thickness of common carotid arteries; HDL-c: high-density lipoprotein cholesterol; SBP: systolic blood pressure; SHAP: Shapley Additive Explanation; TG: triglycerides.



Discussion

Principal Findings

This study provides a comprehensive evaluation of the role of ML in predicting CHD. Among a cohort of 7260 participants, 305 were diagnosed with CHD. The analysis not only validated several well-established cardiovascular and metabolic risk factors but also identified novel predictors of CHD. Importantly, the findings underscore the use of ML models and the SHAP method in elucidating key contributors to CHD risk, with RF demonstrating superior performance, excelling in both discrimination and calibration for CHD prediction.

Comparison With Prior Work

Arterial Health

Carotid IMT emerged as the strongest predictor of CHD in our study. IMT_cMax, which measures the thickness of the CCAs, is a well-established indicator of atherosclerosis and future cardiovascular events, including MI and stroke [31,32]. Multiple studies support this, showing that even a small increase in IMT correlates with a significantly elevated risk of acute MI and stroke. For instance, in the Atherosclerosis Risk in Communities study, a 0.1 mm increase in IMT corresponded to a 50% increase in CHD risk [20,31]. Therefore, measuring IMT through noninvasive techniques like ultrasound has important clinical applications in evaluating subclinical atherosclerosis and assessing CHD risk. Given that many coronary artery assessments are invasive, the use of ultrasound to measure carotid artery IMT offers a valuable alternative for early detection and risk stratification.

Blood Pressure, Lipid Profiles, and Glucose

SBP and hypertension were among the most critical predictors of CHD, aligning with the well-established association between elevated blood pressure and cardiovascular risk [33,34]. Both SBP and diastolic blood pressure were prominent, emphasizing the need for effective blood pressure management in reducing CHD risk [33,35].

Furthermore, non-HDL-c and triglycerides were strongly associated with CHD, confirming the importance of lipid management in cardiovascular health [36-39]. Glucose levels were also significant, suggesting that monitoring glucose metabolism is essential in cardiovascular risk management [40-42].

Renal Function and Metabolic Factors

The role of eGFR as a key predictor highlights the connection between renal function and CHD [43]. Impaired kidney function has been increasingly recognized as a cardiovascular risk factor, particularly due to its association with hypertension and dyslipidemia [44,45]. The results support incorporating kidney function markers in future CHD risk assessments. In addition, metabolic markers and body fat percentage were identified as important predictors, signaling the impact of obesity-related factors on cardiovascular health. These findings suggest that obesity-related measures beyond BMI should be considered in CHD risk assessments.

Sex

The sex-specific analysis highlighted the protective effect of being female, consistent with existing research showing that premenopausal women are generally at a lower risk of developing CHD due to protective hormonal factors [46,47]. These findings suggest the need for sex-specific strategies in managing CHD risk.

Potential Risk Factors

One of the strengths of this study is its ability to uncover novel predictors, such as white blood cell count, which serves as a marker of systemic inflammation. Inflammation is increasingly recognized as a key player in the development of atherosclerosis and cardiovascular events. Additionally, lower calcium levels were associated with a higher risk of CHD, highlighting the importance of mineral balance in cardiovascular health. Furthermore, body fat percentage and BMI were highlighted as significant predictors of CHD, further emphasizing the need for a comprehensive evaluation of obesity-related metrics in cardiovascular risk assessments. These novel insights could lead to more personalized prevention strategies for individuals who may not exhibit classic cardiovascular risk profiles.

Limitations

Despite the promising results, several limitations of the study need to be considered. First, the quality of the data, particularly with respect to missing values, poses a challenge. Although feature selection techniques, such as least absolute shrinkage and selection operator regression and SHAP analysis, were used to mitigate this, the impact of missing data remains a potential limitation. Second, the generalizability of the findings is limited because the study relies on a specific population. The results may not fully apply to populations with different demographic and clinical characteristics. To address this, future research should focus on evaluating these ML models in real-world clinical settings, where variability in clinical practice, missing data, and other factors may affect model performance.

Conclusions

This study demonstrates the potential of ML in predicting CHD. The SHAP method enhances the interpretability of the prediction model, aiding health care professionals in clinical practice by supporting effective risk management and intervention strategies.

Acknowledgments

This article was supported by the Japan Science and Technology Agency COI-NEXT (grant JPMJPF2018) to MA.

Data Availability

The dataset examined in this study is not available to the public due to the inclusion of individuals' personal information but is available from the corresponding author at a reasonable request.

Authors' Contributions

TV conceptualized and designed the study, conducted the data analysis and interpretation, and drafted the manuscript. Y Kokubo contributed to the study concept and design, curated the data, provided resources, and supervised the project. MA assisted in study design, data curation, and supervision. MI and MY contributed to data analysis and interpretation. RD, AMM, JTT, AA, MT, YMN, and TI provided critical feedback and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

YMN is employed by the Department of Digital Health and Epidemiology, Graduate School of Medicine and Public Health, Kyoto University, an Industry-Academia Collaboration Course supported by Eisai Co., Ltd. and Kyowa Kirin Co., Ltd. Additionally, YMN reports a study grant from Bayer outside the submitted work.

Multimedia Appendix 1

Percentage of missing data across all variables prior to imputation (Japanese participants, aged 30-84 years, Suita Study).
[PNG File (Portable Network Graphics File), 94 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Correlation Coefficients for variables used in CHD Incidence prediction (Japanese participants, aged 30-84 years, Suita Study).

[PNG File (Portable Network Graphics File), 260 KB-Multimedia Appendix 2]

Multimedia Appendix 3

List of variables included in the CHD Incidence prediction model (Japanese participants, aged 30-84 years, Suita Study).
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 3]

Checklist 1

TRIPOD + AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis and Artificial Intelligence) checklist
[PDF File (Adobe File), 1664 KB-Checklist 1]

References

1. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. Oct 2020;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9]
2. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019. J Am Coll Cardiol. Dec 2020;76(25):2982-3021. [doi: 10.1016/j.jacc.2020.11.010]
3. Lim HY, Burrell LM, Brook R, Nandurkar HH, Donnan G, Ho P. The need for individualized risk assessment in cardiovascular disease. J Pers Med. Jul 14, 2022;12(7):1140. [doi: 10.3390/jpm12071140] [Medline: 35887637]
4. Matheson MB, Kato Y, Baba S, Cox C, Lima JAC, Ambale-Venkatesh B. Cardiovascular risk prediction using machine learning in a large Japanese cohort. Circ Rep. Dec 9, 2022;4(12):595-603. [doi: 10.1253/circrep.CR-22-0101] [Medline: 36530840]
5. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944. [doi: 10.1371/journal.pone.0174944]
6. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. Behav Ther. Sep 2020;51(5):675-687. [doi: 10.1016/j.beth.2020.05.002] [Medline: 32800297]
7. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel). Dec 25, 2020;23(1):18. [doi: 10.3390/e23010018] [Medline: 33375658]
8. Vu T, Kokubo Y, Inoue M, et al. Machine learning approaches for stroke risk prediction: findings from the Suita Study. J Cardiovasc Dev Dis. Jul 1, 2024;11(7):207. [doi: 10.3390/jcdd11070207] [Medline: 39057627]
9. Martin-Morales A, Yamamoto M, Inoue M, Vu T, Dawadi R, Araki M. Predicting cardiovascular disease mortality: leveraging machine learning for comprehensive assessment of health and nutrition variables. Nutrients. Sep 11, 2023;15(18):3937. [doi: 10.3390/nu15183937] [Medline: 37764721]
10. Thanh NT, Luan VT, Viet DC, Tung TH, Thien V. A machine learning-based risk score for prediction of mechanical ventilation in children with dengue shock syndrome: a retrospective cohort study. PLoS ONE. 2024;19(12):e0315281. [doi: 10.1371/journal.pone.0315281] [Medline: 39642139]
11. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2023;14(7):8459-8486. [doi: 10.1007/s12652-021-03612-z] [Medline: 35039756]
12. Forrest IS, Petrazzini BO, Duffy Á, et al. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. The Lancet. Jan 2023;401(10372):215-225. [doi: 10.1016/S0140-6736(22)02079-7]
13. Bani Hani SH, Ahmad MM. Machine-learning algorithms for ischemic heart disease prediction: a systematic review. Curr Cardiol Rev. 2023;19(1):e090622205797. [doi: 10.2174/1573403X18666220609123053] [Medline: 35692135]
14. Alizadehsani R, Abdar M, Roshanzamir M, et al. Machine learning-based coronary artery disease diagnosis: a comprehensive review. Comput Biol Med. Aug 2019;111:103346. [doi: 10.1016/j.compbiomed.2019.103346] [Medline: 31288140]
15. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
16. Tunstall-Pedoe H, Kuulasmaa K, Amouyel P, Arveiler D, Rajakangas AM, Pajak A. Myocardial infarction and coronary deaths in the World Health Organization MONICA Project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. Circulation. Jul 1994;90(1):583-612. [doi: 10.1161/01.cir.90.1.583] [Medline: 8026046]
17. Kokubo Y, Kamide K, Okamura T, et al. Impact of high-normal blood pressure on the risk of cardiovascular disease in a Japanese urban cohort: the Suita Study. Hypertension. Oct 2008;52(4):652-659. [doi: 10.1161/HYPERTENSIONAHA.108.118273] [Medline: 18725580]

18. Kokubo Y, Watanabe M, Higashiyama A, et al. Interaction of blood pressure and body mass index with risk of incident atrial fibrillation in a Japanese urban cohort: the Suita Study. Am J Hypertens. Nov 2015;28(11):1355-1361. [doi: 10.1093/ajh/hpv038] [Medline: 25845964]
19. Imai E, Horio M, Nitta K, et al. Estimation of glomerular filtration rate by the MDRD study equation modified for Japanese patients with chronic kidney disease. Clin Exp Nephrol. Mar 2007;11(1):41-50. [doi: 10.1007/s10157-006-0453-4] [Medline: 17384997]
20. Kokubo Y, Watanabe M, Higashiyama A, Nakao YM, Nakamura F, Miyamoto Y. Impact of intima-media thickness progression in the common carotid arteries on the risk of incident cardiovascular disease in the Suita Study. J Am Heart Assoc. Jun 1, 2018;7(11):e007720. [doi: 10.1161/JAHA.117.007720] [Medline: 29858361]
21. Christodoulou E, van Smeden M, Edlinger M, et al. Adaptive sample size determination for the development of clinical prediction models. Diagn Progn Res. Mar 22, 2021;5(1):6. [doi: 10.1186/s41512-021-00096-5] [Medline: 33745449]
22. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. Mar 18, 2020;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
23. Bewick V, Cheek L, Ball J. Statistics review 14: logistic regression. Crit Care. Feb 2005;9(1):112-118. [doi: 10.1186/cc3045] [Medline: 15693993]
24. Su X, Xu Y, Tan Z, et al. Prediction for cardiovascular diseases based on laboratory data: an analysis of random forest model. J Clin Lab Anal. Sep 2020;34(9):e23421. [doi: 10.1002/jcla.23421] [Medline: 32725839]
25. Unnikrishnan P, Kumar DK, Poosapadi Arjunan S, Kumar H, Mitchell P, Kawasaki R. Development of health parameter model for risk prediction of CVD using SVM. Comput Math Methods Med. 2016;2016:3016245. [doi: 10.1155/2016/3016245] [Medline: 27594895]
26. Son YJ, Kim HG, Kim EH, Choi S, Lee SK. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthc Inform Res. Dec 2010;16(4):253-259. [doi: 10.4258/hir.2010.16.4.253] [Medline: 21818444]
27. Vu T, Dawadi R, Yamamoto M, et al. Prediction of depressive disorder using machine learning approaches: findings from the NHANES. BMC Med Inform Decis Mak. Feb 17, 2025;25(1):83. [doi: 10.1186/s12911-025-02903-1] [Medline: 39962516]
28. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. [doi: 10.1186/s41512-019-0064-7] [Medline: 31592444]
29. Zhang Z, Rousson V, Lee WC, et al. Decision curve analysis: a technical note. Ann Transl Med. Aug 2018;6(15):308. [doi: 10.21037/atm.2018.07.02] [Medline: 30211196]
30. Bloch L, Friedrich CM, Alzheimer's Disease Neuroimaging Initiative. Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. Alzheimers Res Ther. Sep 15, 2021;13(1):155. [doi: 10.1186/s13195-021-00879-4] [Medline: 34526114]
31. Kawai T, Ohishi M, Takeya Y, et al. Carotid plaque score and intima media thickness as predictors of stroke and mortality in hypertensive patients. Hypertens Res. Oct 2013;36(10):902-909. [doi: 10.1038/hr.2013.61] [Medline: 23823172]
32. Nambi V, Chambless L, Folsom AR, et al. Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk: the ARIC (Atherosclerosis Risk In Communities) study. J Am Coll Cardiol. Apr 13, 2010;55(15):1600-1607. [doi: 10.1016/j.jacc.2009.11.075] [Medline: 20378078]
33. Ji C, Wang N, Shi J, et al. Level of systolic blood pressure within the normal range and risk of cardiovascular events in the absence of risk factors in Chinese. J Hum Hypertens. Oct 2022;36(10):933-939. [doi: 10.1038/s41371-021-00598-1] [Medline: 34480099]
34. Whelton SP, McEvoy JW, Shaw L, et al. Association of normal systolic blood pressure level with cardiovascular disease in the absence of risk factors. JAMA Cardiol. Sep 1, 2020;5(9):1011-1018. [doi: 10.1001/jamacardio.2020.1731] [Medline: 32936272]
35. Li J, Somers VK, Gao X, et al. Evaluation of optimal diastolic blood pressure range among adults with treated systolic blood pressure less than 130 mm Hg. JAMA Netw Open. Feb 1, 2021;4(2):e2037554. [doi: 10.1001/jamanetworkopen.2020.37554] [Medline: 33595663]
36. Guo LL, Chen YQ, Lin QZ, et al. Non-HDL-C Is more stable than LDL-C in assessing the percent attainment of non-fasting lipid for coronary heart disease patients. Front Cardiovasc Med. 2021;8:649181. [doi: 10.3389/fcvm.2021.649181] [Medline: 33869310]
37. Saito I, Yamagishi K, Kokubo Y, et al. Non-high-density lipoprotein cholesterol and risk of stroke subtypes and coronary heart disease: The Japan Public Health Center-Based Prospective (JPHC) Study. JAT. 2020;27(4):363-374. [doi: 10.5551/jat.50385]

38. Dong J, Yang S, Zhuang Q, et al. The associations of lipid profiles with cardiovascular diseases and death in a 10-year prospective cohort study. Front Cardiovasc Med. 2021;8:745539. [doi: 10.3389/fcvm.2021.745539] [Medline: 34901209]
39. Zhao X, Wang D, Qin L. Lipid profile and prognosis in patients with coronary heart disease: a meta-analysis of prospective cohort studies. BMC Cardiovasc Disord. Feb 3, 2021;21(1):69. [doi: 10.1186/s12872-020-01835-0] [Medline: 33535982]
40. Poznyak AV, Litvinova L, Poggio P, Sukhorukov VN, Orekhov AN. Effect of glucose levels on cardiovascular risk. Cells. Sep 28, 2022;11(19):3034. [doi: 10.3390/cells11193034] [Medline: 36230996]
41. Riise HKR, Igland J, Sulo G, et al. Casual blood glucose and subsequent cardiovascular disease and all-cause mortality among 159 731 participants in Cohort of Norway (CONOR). BMJ Open Diabetes Res Care. Feb 2021;9(1):e001928. [doi: 10.1136/bmjdrc-2020-001928] [Medline: 33622686]
42. Selvin E, Coresh J, Golden SH, Brancati FL, Folsom AR, Steffes MW. Glycemic control and coronary heart disease risk in persons with and without diabetes. Arch Intern Med. Sep 12, 2005;165(16):1910. [doi: 10.1001/archinte.165.16.1910]
43. Charoen P, Nitsch D, Engmann J, et al. Mendelian randomisation study of the influence of eGFR on coronary heart disease. Sci Rep. Jun 24, 2016;6:28514. [doi: 10.1038/srep28514] [Medline: 27338949]
44. Jankowski J, Floege J, Fliser D, Böhm M, Marx N. Cardiovascular disease in chronic kidney disease. Circulation. Mar 16, 2021;143(11):1157-1172. [doi: 10.1161/CIRCULATIONAHA.120.050686]
45. Brugts JJ, Knetsch AM, Mattace-Raso FUS, Hofman A, Witteman JCM. Renal function and risk of myocardial infarction in an elderly population: the Rotterdam Study. Arch Intern Med. 2005;165(22):2659-2665. [doi: 10.1001/archinte.165.22.2659] [Medline: 16344425]
46. Maas A, Appelman YEA. Gender differences in coronary heart disease. Neth Heart J. Dec 2010;18(12):598-602. [doi: 10.1007/s12471-010-0841-y] [Medline: 21301622]
47. Shah T, Palaskas N, Ahmed A. An update on gender disparities in coronary heart disease care. Curr Atheroscler Rep. May 2016;18(5):28. [doi: 10.1007/s11883-016-0574-5] [Medline: 27029220]

Abbreviations

AUC: area under the curve
CCA: common carotid artery
CHD: coronary heart disease
DCA: decision curve analysis
eGFR: estimated glomerular filtration rate
HDL-c: high-density lipoprotein cholesterol
IMT: intima-media thickness
IMT_cMax: maximum intima-media thickness of common carotid arteries
LightGBM: Light Gradient-Boosting Machine
LR: logistic regression
MI: myocardial infarction
ML: machine learning
OE: observed-to-expected
RF: random forest
SBP: systolic blood pressure
SHAP: Shapley Additive Explanation
SVM: support vector machine
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis and Artificial Intelligence
XGBoost: Extreme Gradient Boosting

Edited by Andrew Coristine; peer-reviewed by Maha Gasmi, Mahin Nomali, Neeladri Misra; submitted 29.10.2024; final revised version received 03.02.2025; accepted 24.02.2025; published 12.05.2025

Please cite as:
Vu T, Kokubo Y, Inoue M, Yamamoto M, Mohsen A, Martin-Morales A, Dawadi R, Inoue T, Tay JT, Yoshizaki M, Watanabe N, Kuriya Y, Matsumoto C, Arafa A, Nakao YM, Kato Y, Teramoto M, Araki M
Machine Learning Model for Predicting Coronary Heart Disease Risk: Development and Validation Using Insights From a Japanese Population–Based Study
JMIR Cardio 2025;9:e68066
URL: https://cardio.jmir.org/2025/1/e68066

doi: 10.2196/68066