Machine Learning-Based Prediction of Coronary Heart Disease: Comprehensive Insights from the Suita Study

Thien Vu, Yoshihiro Kokubo, Mai Inoue, Masaki Yamamoto, Attayeb Mohsen, Agustin Martin-Morales, Research Dawadi, Takao Inoue, Tay Jie Ting, Mari Yoshizaki, Naoki Watanabe, Yuki Kuriya, Chisa Matsumoto, Ahmed Arafa, Yoko M Nakao, Yuka Kato, Masayuki Teramoto, Michihiro Araki

Submitted to: JMIR Cardio
on: October 29, 2024

Table of Contents

Machine Learning-Based Prediction of Coronary Heart Disease: Comprehensive Insights from the Suita Study

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

[1]Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, Japan Osaka JP
[2]NCD Epidemiology Research Center, Shiga University of Medical Science, Otsu, Shiga, Japan. Otsu JP
[3]Department of Cardiac Surgery, Cardiovascular Center, Cho Ray hospital, Vietnam. Hochiminh VN
[4]Department of Preventive Cardiology, National Cerebral and Cardiovascular Center, Suita, Osaka, 564-8565, Japan. Osaka JP
[5]Department of Cardiology, Center for Health Surveillance and Preventive Medicine, Tokyo Medical University Hospital, Shinjuku, Japan Tokyo JP
[6]Department of Public Health, Faculty of Medicine, Beni-Suef University. Beni-Suef EG
[7]Division of Health Sciences, Osaka University Graduate School of Medicine, Suita, Japan Suita JP
[8]- Graduate School of Medicine, Kyoto University, 54 Shogoin-Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan. Kyoto JP
[9]- Graduate School of Science, Technology and Innovation, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe, Hyogo 657-8501, Japan. Kobe JP

Corresponding Author:
Thien Vu
Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, Japan
3-17 Senrioka-shinmachi, Osaka, Settsu 566-0002, Japan.
Osaka
JP

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. Machine learning (ML) offers advanced methods for analyzing complex datasets, revealing novel predictors of CHD beyond traditional models.

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

Methods: The study recruited 7,672 participants aged 30 to 84 years from Suita City, Japan, between 1989 and 1999. Over an average of 15 years, participants were monitored for cardiovascular events. Five ML models—Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression (LR), and LightGBM—were used. The optimal model was identified based on accuracy, sensitivity, specificity, and AUC. SHapley Additive exPlanations (SHAP) were then employed to explore the contribution of various risk factors to CHD.

Results: RF achieved the highest AUC (95% CI) of 0.94 (0.93-0.96), outperforming LR, SVM, XGBoost, and LightGBM. SHAP on the best model identified the top CHD predictors. Intima-media thickness of common carotid artery (IMT_cMax) was identified as the strongest predictor of CHD, highlighting the importance of arterial health. Systolic and diastolic blood pressure, along with lipid profiles (non-HDL cholesterol, HDL cholesterol, and triglycerides), were closely associated with CHD incidence. eGFR underscored the link between renal function and CHD. Novel insights included the impact of lower calcium levels, systemic inflammation (elevated WBC counts), fructosamine levels, and obesity-related factor (body fat percentage). A protective effect in females indicated the need for sex-specific CHD management strategies.

Conclusions: ML, particularly the RF model combined with SHAP, effectively identified key risk factors for CHD, including arterial health, blood pressure, lipid profiles, renal function, and novel markers. These findings support a multifactorial approach to CHD risk assessment.

(JMIR Preprints 29/10/2024:68066)

DOI: https://doi.org/10.2196/preprints.68066

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 v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: Machine Learning-Based Prediction of Coronary Heart Disease: Comprehensive Insights from the Suita Study

Running Head: Coronary Heart Disease and Machine Learning

Authors, Affiliations, Contact Information

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

1. Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, 3-17 Senrioka-shinmachi, Osaka, Settsu 566-0002, Japan.
2. NCD Epidemiology Research Center, Shiga University of Medical Science, Otsu, Shiga, Japan.
3. Department of Cardiac Surgery, Cardiovascular Center, Cho Ray hospital, Vietnam.
4. Department of Preventive Cardiology, National Cerebral and Cardiovascular Center, Suita, Osaka, 564-8565, Japan.
5. Department of Cardiology, Center for Health Surveillance and Preventive Medicine, Tokyo Medical University Hospital, Shinjuku, Japan.
6. Department of Public Health, Faculty of Medicine, Beni-Suef University.
7. Division of Health Sciences, Osaka University Graduate School of Medicine, Suita, Japan
8. Graduate School of Medicine, Kyoto University, 54 Shogoin-Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan.
9. Graduate School of Science, Technology and Innovation, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe, Hyogo 657-8501, Japan.

*** Corresponding authors:**

1. Thien VU, MD, PhD

- Institute: National Institutes of Biomedical Innovation, Health and Nutrition, Japan.
- Email: thien-vu@nibiohn.go.jp, thienvuyd01@gmail.com

2. Michihiro ARAKI, PhD

- Institute: National Institutes of Biomedical Innovation, Health and Nutrition, Japan.
- Email: araki@nibiohn.go.jp

-

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. Machine learning (ML) offers advanced methods for analyzing complex datasets, revealing novel predictors of CHD beyond traditional models.

Objective:

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

Methods:

The study recruited 7,672 participants aged 30 to 84 years from Suita City, Japan, between 1989 and 1999. Over an average of 15 years, participants were monitored for cardiovascular events. Five ML models—Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression (LR), and LightGBM—were used. The optimal model was identified based on accuracy, sensitivity, specificity, and AUC. SHapley Additive exPlanations (SHAP) were then employed to explore the contribution of various risk factors to CHD.

Results:

RF achieved the highest AUC (95% CI) of 0.94 (0.93-0.96), outperforming LR, SVM, XGBoost, and LightGBM. SHAP on the best model identified the top CHD predictors. Intima-media thickness of common carotid artery (IMT_cMax) was identified as the strongest predictor of CHD, highlighting the importance of arterial health. Systolic and diastolic blood pressure, along with lipid profiles (non-HDL cholesterol, HDL cholesterol, and triglycerides), were closely associated with CHD incidence. eGFR underscored the link between renal function and CHD. Novel insights included the impact of lower calcium levels, systemic inflammation (elevated WBC counts), fructosamine levels, and obesity-related factor (body fat percentage). A protective effect in females indicated the need for sex-

specific CHD management strategies.

Conclusion:

ML, particularly the RF model combined with SHAP, effectively identified key risk factors for CHD, including arterial health, blood pressure, lipid profiles, renal function, and novel markers. These findings support a multifactorial approach to CHD risk assessment.

Keywords: Coronary Heart Disease (CHD), Machine Learning (ML), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boost (XGBoost), Light Gradient Boosted Machine (LightGBM), Shapley Addictive ExPlanations (SHAP).

1. Introduction

Coronary heart disease (CHD) is a prevalent and life-threatening cardiovascular condition characterized by narrowing or blockage of the coronary arteries [1], [2]. Accurate prediction and timely risk assessment are crucial for effective preventive measures and personalized interventions [3]. While conventional risk assessment models have been used, there is growing recognition of the potential of machine learning in enhancing CHD prediction [4], [5].

Machine learning 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], [7], [8], [9]. 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 [10]. Several studies have demonstrated the effectiveness of machine learning 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 machine learning model using electronic health records and assessed its probabilities as in silico scores for coronary artery disease in participants in two longitudinal biobank cohorts [11]. Another study applied an ensemble ML model for coronary disease prediction, using ML classifiers to predict heart disease [12]. These findings highlight the potential of machine learning in driving innovation and improving the accuracy of CHD diagnosis and prediction [13].

However, challenges exist in utilizing machine learning 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.

This study aimed to address the role of machine learning 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.

2. Methods

2.1. Study Participants

The Suita Study, a prospective population-based cohort study, was conducted in Suita City, Osaka, Japan. Previous publications have provided extensive information regarding the methodology and selection criteria of the study [14], [15]. Briefly, from 1989 to 1999, a total of 7672 men and women aged 30 to 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 two years for an average of 15 years until their first occurrence of stroke, myocardial infarction (MI), death, or relocation. Opt-out procedures were implemented for those who preferred not to participate in this study. Informed consent was obtained from all participants. The study was conducted in accordance with the guidelines and regulations outlined in the Declaration of Helsinki, with approval from the Institutional Review Board at the National Cerebral and Cardiovascular Center. (No. R21024-2).

2.2. Data collection

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. Detailed information regarding the data collection procedure can be found elsewhere [14], [15]. These evaluations served as the baseline examination for the current investigation.

2.3. Risk Factors and Anthropomorphic Measurements

Each participant's blood pressure (BP) was measured using a mercury column sphygmomanometer, an appropriately sized cuff, and a standardized protocol to ensure accuracy and precision [15]. 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 one minute to allow for adequate observation and recording. Hypertension was defined as a systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, or the use of antihypertensive medications. Body mass index (BMI) was calculated as weight (kg) divided by the square of height (m^2). At baseline, routine blood tests were performed, including serum total and high-density lipoprotein cholesterol and glucose levels. Non-high-density lipoprotein cholesterol levels were calculated by subtracting high-density lipoprotein cholesterol (HDL-c) from total cholesterol. Diabetes mellitus was defined as a fasting plasma glucose (FPG) ≥126 mg/dL, a non-FPG ≥200 mg/dL, or the use of diabetes mellitus medication. Metabolic syndrome was defined as a combination of abdominal obesity, impaired fasting glucose, atherogenic dyslipidemia, and elevated blood pressure according to the original Japanese criteria for metabolic syndrome, which were a waist circumference ≥85 cm in men and ≥90 cm in women and/or a BMI ≥25.0 kg/m^2, an essential component plus ≥2 (definite MetS) of the following [16], [17]: (1) systolic blood pressure ≥130 mm Hg and/or diastolic blood pressure ≥85 mm Hg or medication use; (2) triglyceride level ≥150 mg/dL and/or HDL cholesterol level <40 mg/dL; and (3) fasting glucose level ≥100 mg/dL and/or medication use. The estimated glomerular filtration rate (eGFR) (mL/min/1.73 m^2) was calculated according to the original Modification of Diet in Renal Disease (MDRD) equation modified by the Japanese coefficient (0.881) as follows:

$0.881 \times 186 \times \text{serum creatinine}^{-1.154} \times \text{age}^{-0.203} \times (0.742 \text{ if female})$ [18]. Carotid artery was measured by high-resolution ultrasound machine with atherosclerotic indexes of intima-media thickness (IMT) on both sides of common carotid artery (CCA), carotid artery bulb, internal carotid artery (ICA) and external carotid artery (ECA). The maximum IMT in the CCA (IMT_cMax) was defined as the maximum measurable IMT in the scanned CCA areas, and the maximum IMT (IMT_MAX) in the entire area was defined as the maximum measurable IMT in the entire scanned CCA, bulb, ICA and ECA areas for both sides [19]. Atrial Fibrillation was checked by standard 12-lead ECGs from all participants and was determined by well-trained physicians [14].

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.

2.4. Confirmation of CHD

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. Myocardial infarctions were classified as definite or probable according to the criteria established by the MONICA project [20].

Every two 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 [14], [15]. CHD was diagnosed based on specific criteria such as primary heart attack, sudden death within 24 hours of acute illness onset, and coronary artery disease with bypass surgery or intervention.

2.5 Analysis methods

The goal of this analysis was to predict the incidence of Coronary Heart Disease (CHD) using machine learning models and examine the contribution of each risk factor on the CHD incidence. A

comprehensive process was followed, which included descriptive analysis, feature selection, model training, hyperparameter optimization, and interpretability through SHAP (SHapley Additive exPlanations) values.

Data Preprocessing

Initially, we imported two datasets: one containing baseline clinical and demographic information, and the other containing ultrasound data. To ensure data quality, we conducted extensive missing data analysis. Variables with over 30% missing data were excluded from the analysis to improve model robustness. Refer to **Supplementary Figure S1** for details. Various visual tools were employed to inspect missingness patterns, including naniar, VIM and dlookr packages. Then missing data was addressed using Multivariate Imputation by Chained Equations (MICE), ensuring that all variables had sufficient data for downstream analysis.

Descriptive Analysis

Before model building, we conducted a detailed descriptive analysis to understand patient characteristics. This step involved generating summary statistics for continuous and categorical variables, stratified by CHD incidence. Continuous variables were summarized using means and standard deviations for normally distributed data, or medians and interquartile ranges for non-normally distributed data. Categorical variables were reported as frequencies and percentages. To compare differences in patient characteristics based on CHD incidence (yes/no), we employed various statistical tests including Student t-tests, Mann–Whitney U tests, or Chi–squared 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 employed to identify and remove variables with high multicollinearity, defined as having a correlation

coefficient greater than 0.8. Refer to Correlation Coefficients heat map in the **Supplementary Figure S2** for details. The next step involved applying Least Absolute Shrinkage and Selection Operator (LASSO) 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 cholesterol, and blood pressure were retained, given their established association with coronary heart disease (CHD). The variables used for model development were described in **Supplementary Table S1**, and their contributions to the incidence of CHD were illustrated in **Supplementary Figure S3**.

Handling Class Imbalance

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.

Machine learning model development

The objective was to predict CHD incidence, and various machine learning models were trained for this purpose. The dataset was split into training (80%) and testing (20%) sets while maintaining balanced target variable distributions across both. Several machine learning algorithms were implemented to compare their predictive power. LR was used as a baseline model, offering simplicity and interpretability [21]. RF, an ensemble learning method, was employed due to its strength in handling high-dimensional data and offering feature importance insights [8], [22]. SVM with radial basis kernels were used for their effectiveness in non-linear classification tasks [23], [24]. In addition, two gradient-boosting models, XGBoost and LightGBM, were implemented because of their superior performance in handling complex datasets and large-scale data [9], [25].

Hyperparameter optimization was carried out to fine-tune the performance of each model. For LR, RF and SVM models, the caret package was used to perform grid search-based hyperparameter tuning and cross-validation. For XGBoost and LightGBM, internal tuning methods were used. Key hyperparameters were optimized for each model. In the RF, the number of trees and maximum depth were optimized, while XGBoost's parameters, including learning rate, maximum depth, subsample ratio, and boosting rounds, were fine-tuned. LightGBM was optimized by adjusting the number of leaves, learning rate, and feature fraction. The hyperparameter tuning process aimed to maximize key performance metrics such as accuracy, area under the curve (AUC), and F1 score, ensuring the models achieved the best possible predictive performance.

Model evaluation was conducted using several performance metrics. Accuracy was used to assess the overall proportion of correct predictions. Sensitivity measured the models' ability to correctly identify CHD cases, while specificity assessed their ability to correctly identify non-CHD cases. Precision reflected the proportion of true positive CHD cases among all predicted positives, and AUC provided an overall measure of the models' discriminatory power between CHD and non-CHD cases. The F1 score, balancing precision and recall, was used to evaluate model performance in the context of imbalanced data. Additionally, bootstrapping was performed to estimate 95% confidence intervals for each performance metric, further enhancing the robustness of the model evaluation. The performance metrics were presented in **Supplementary Table S2**.

SHAP Analysis for Interpretability

To interpret the models' decisions, we used SHAP values, particularly for the XGBoost and RF models. SHAP values provided insight into the contribution of each feature to the prediction of CHD, allowing for model transparency [8], [9], [26]. SHAP summary plots visualized the importance of key features, while SHAP dependence plots highlighted the non-linear relationships between features and incidence of CHD.

3. Results

In this study, 7,260 participants were analyzed, of which 305 (4.2%) were diagnosed with coronary heart disease. The participants with CHD had a significantly older median age of 63 years compared to 55 years in those without. CHD was more prevalent in men (66.2%) compared to women (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 estimated glomerular filtration rate (eGFR) was lower in participants with CHD compared to those without. The intima-media thickness of common carotid arteries, IMT_cMax, were also significantly higher in CHD patients (1.10 mm vs. 1.00 mm, $p < 0.001$).

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 cholesterol levels and higher non-HDL cholesterol and triglyceride levels in CHD patients.

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 two groups.

Regarding lifestyle factors, current smoking was more prevalent among participants with CHD (36.1% vs. 29.0%, $p < 0.001$), 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**.

The performance metrics of the five machine learning models employed in our CHD prediction study provide valuable insights into their effectiveness, as shown in **Table 2.** The RF model performed the best, achieving an accuracy of 0.97, a high sensitivity of 0.89, and perfect specificity of 1.0, indicating it is both highly reliable in identifying CHD cases and in correctly classifying non-CHD

participants. XGBoost showed similarly strong performance, with an accuracy of 0.95 and an AUC of 0.91, providing high sensitivity and specificity, which makes it another robust model for this application. SVM model also delivered solid results, with an accuracy of 0.94, a sensitivity of 0.82, and an AUC of 0.90. While its sensitivity is slightly lower than RF and XGBoost, it maintains a good balance between identifying CHD and non-CHD cases. In contrast, LR served as a baseline model with lower overall performance. Its accuracy of 0.77 and a sensitivity of 0.53 suggest that while it correctly identifies non-CHD cases relatively well (specificity of 0.87), it struggles to detect CHD cases. The LightGBM model, however, was the weakest performer in this analysis, with an accuracy of only 0.30 and an AUC of 0.50. While LightGBM achieved a sensitivity of 1.0, indicating that it correctly identified all CHD cases, its inability to classify non-CHD participants accurately (specificity of 0.0) makes it unreliable for this task. Overall, RF and XGBoost emerged as the most effective models for predicting CHD, balancing both sensitivity and specificity while maintaining high predictive power across all metrics.

The results presented in the figures provide a comprehensive overview of the most important variables in predicting coronary heart disease (CHD) using SHAP values.

In the F**igure 1**, the bar plot on the left ranks the top features contributing to CHD prediction, with maximum carotid intima-media thickness (IMT_cMax) identified as the most influential variable, followed by hypertension (htn), estimated glomerular filtration rate (eGFR), non-HDL cholesterol (non_HDLc), and systolic blood pressure (SBP). 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 cholesterol, and blood pressure are positively associated with an increased likelihood of CHD, whereas lower levels of protective factors like HDL cholesterol and eGFR are associated with a higher risk of CHD. Other

important variables, such as triglycerides (TG), age, and glucose levels, 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 (WBC) and other factors such as calcium levels, fructosamine level, sex, and body fat (bf) play roles in determining CHD risk.

The **Figure 2** 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 cholesterol shows a generally positive association with CHD, where higher levels correspond to a higher risk. For systolic blood pressure (SBP), the risk of CHD increases sharply with rising SBP values. HDL cholesterol 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, where 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.

Overall, these findings highlight the importance of a combination of traditional cardiovascular risk factors, metabolic markers, arterial health (IMT_cMax), and kidney function (eGFR), along with novel predictors such as WBC, calcium level, fructosamine level and body fat percentage, in predicting CHD. The detailed SHAP visualizations provide insights into the complex interactions between these variables and their influence on CHD risk, reinforcing the need for comprehensive risk management strategies targeting these modifiable factors.

4. Discussion

The present study provides a comprehensive evaluation of the role of machine learning in predicting coronary heart disease, underscoring the importance of key cardiovascular and metabolic predictors while identifying potential novel risk factors. Machine learning's ability to process and analyze large, multidimensional data has proven to be a powerful tool in identifying predictors of CHD that go beyond conventional risk models. Especially, by utilizing Random Forest (RF) models combined with SHapley Additive exPlanations (SHAP), this analysis enhances our understanding of the relative contributions of various risk factors and demonstrates the effectiveness of ML techniques in cardiovascular risk prediction.

Role of Machine Learning in CHD Prediction

Machine learning offers a distinct advantage over traditional statistical models by uncovering complex patterns in high-dimensional data. In this study, the RF model, in conjunction with SHAP values, successfully identified critical predictors of CHD, including IMT_cMax, hypertension, estimated glomerular filtration rate (eGFR), non-HDL cholesterol, and triglycerides (TG). These ranking underscores the importance of arterial health, blood pressure regulation, lipid levels, and kidney function in assessing CHD risk. Other significant variables, such as age and glucose levels, were also linked to CHD, with older age and impaired glucose regulation associated with higher risk. Additionally, markers of inflammation like white blood cell count (WBC) and other factors such as calcium levels, sex, fructosamine levels, and body fat percentage (bf) play crucial roles in determining CHD risk. These findings reinforce the growing recognition of ML as a valuable tool for enhancing risk assessment and preventive strategies in clinical settings.

Importance of Key Predictors

Arterial health

Carotid intima-media thickness (IMT) emerged as the strongest predictor of CHD in our study.

IMT_cMax, which measures the thickness of the common carotid arteries, is a well-established indicator of atherosclerosis and future cardiovascular events, including myocardial infarction and stroke [27], [28]. Multiple studies support this, showing that even a small increase in IMT correlates with a significantly elevated risk of acute myocardial infarction and stroke. For instance, in the Atherosclerosis Risk in Communities (ARIC) study, a 0.1 mm increase in IMT corresponded to a 50% increase in CHD risk [27], [29]. Therefore, measuring IMT through non-invasive 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

Systolic blood pressure (SBP) and hypertension were among the most critical predictors of CHD, aligning with the well-established association between elevated blood pressure and cardiovascular risk [30], [31]. Both SBP and diastolic blood pressure (DBP) were prominent, emphasizing the need for effective blood pressure management in reducing CHD risk [30], [32].

The study also confirmed the importance of lipid management, as non-HDL cholesterol and triglycerides were strongly related to CHD incidence [33], [34], [35], [36]. Additionally, glucose levels were shown to be significant, with higher levels associated with an increased risk of CHD, pointing to the importance of monitoring glucose metabolism in cardiovascular health [37], [38], [39]. These results underline the role of dyslipidemia and impaired glucose metabolism in the pathogenesis of CHD.

Renal Function and Metabolic Factors

The role of eGFR as a key predictor highlights the connection between renal function and CHD [40]. Impaired kidney function has been increasingly recognized as a cardiovascular risk factor, particularly due to its association with hypertension and dyslipidemia [41], [42]. The results support incorporating kidney function markers in future CHD risk assessments. In addition, metabolic

marker, body fat percentage (bf) was identified as important predictors, signaling the impact of obesity-related factors on cardiovascular health. These findings suggest that obesity-related measures beyond body mass index (BMI) should be considered in CHD risk assessments.

Gender

Gender-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 [43], [44]. These findings suggest the need for sex-specific strategies in managing CHD risk.

Novel Insights and Potential Risk Factors

One of the notable strengths of this study is its ability to uncover novel risk factors, such as WBC count, which serves as a marker of systemic inflammation. Recent evidence suggests that inflammation plays a pivotal role in the development of atherosclerosis and cardiovascular events, and our study's findings align with this growing body of research. Additionally, lower calcium levels were associated with a higher risk of CHD, highlighting the importance of mineral balance in cardiovascular health. Fructosamine level, a marker of total glycated serum proteins, was positively associated to CHD and offers a valuable alternative to HbA1c for monitoring glycemic status, particularly over shorter periods, making it useful for tracking rapid metabolic changes. Furthermore, body fat percentage (bf) and waist circumference 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.

Strenth and Limitation

While machine learning models like RF demonstrated superior predictive power in this study, several challenges remain. Data quality, particularly in terms of missing values and feature selection, plays a crucial role in building robust models. Although systematic feature selection methods, including

LASSO regression and SHAP analysis, were employed, ensuring model interpretability is a critical challenge that must be addressed for the successful integration of ML algorithms into clinical practice. SHAP values, which provide insights into how individual features contribute to model predictions, help mitigate this challenge by enhancing model transparency.

However, a significant limitation of this study is the generalizability of the findings. The dataset used is specific to a particular population, and further validation is required across more diverse populations to assess the external validity of the models. 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.

Conclusion

This study underscores the potential of machine learning in predicting coronary heart disease by identifying key risk factors like IMT_cMax, hypertension, lipid profiles, and renal function. The identification of novel predictors such as WBC count, Calcium level, frutosamine level and body fat percentage highlights the importance of a multifactorial approach to CHD risk assessment. Although the integration of ML models into clinical practice presents challenges related to data quality, interpretability, and generalizability, this study paves the way for more personalized and effective cardiovascular risk management strategies.

Author contributions

Study concept and design: TV, YK, MA; data analysis and interpretation: TV, MI, MY; drafting of the manuscript: TV; resources: YK; data curation: YK, MA; supervision: YK, MA; reviewing and editing: TV, RD, AMM, JTT, AA, MT, YMN, TI. All authors critically revised and approved the final version of the manuscript.

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.

Conflicts of interest: YMN reports a study grant from Bayer outside the submitted work. All authors declare no conflicts of interest.

Acknowledgments: This article was supported by the Japan Science and Technology Agency (JST) COI-NEXT Grant number JPMJPF2018 to M.A.

References

[1] T. Vos *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,” *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/S0140-6736(20)30925-9.

[2] G. A. Roth *et al.*, “Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019,” *J Am Coll Cardiol*, vol. 76, no. 25, pp. 2982–3021, Dec. 2020, doi: 10.1016/j.jacc.2020.11.010.

[3] H. Y. Lim, L. M. Burrell, R. Brook, H. H. Nandurkar, G. Donnan, and P. Ho, “The Need for Individualized Risk Assessment in Cardiovascular Disease,” *J Pers Med*, vol. 12, no. 7, p. 1140, Jul. 2022, doi: 10.3390/jpm12071140.

[4] M. B. Matheson, Y. Kato, S. Baba, C. Cox, J. A. C. Lima, and B. Ambale-Venkatesh, “Cardiovascular Risk Prediction Using Machine Learning in a Large Japanese Cohort,” *Circ Rep*, vol. 4, no. 12, p. CR-22-0101, Dec. 2022, doi: 10.1253/circrep.CR-22-0101.

[5] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?,” *PLoS One*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.

[6] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised Machine Learning: A Brief Primer,” *Behav Ther*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/j.beth.2020.05.002.

[7] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[8] T. Vu *et al.*, “Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study,” *J Cardiovasc Dev Dis*, vol. 11, no. 7, p. 207, Jul. 2024, doi: 10.3390/jcdd11070207.

[9] A. Martin-Morales, M. Yamamoto, M. Inoue, T. Vu, R. Dawadi, and M. Araki, “Predicting Cardiovascular Disease Mortality: Leveraging Machine Learning for Comprehensive Assessment of Health and Nutrition Variables,” *Nutrients*, vol. 15, no. 18, p. 3937, Sep. 2023, doi:

10.3390/nu15183937.

[10] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, “Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda,” *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, pp. 8459–8486, Jul. 2023, doi: 10.1007/s12652-021-03612-z.

[11] I. S. Forrest *et al.*, “Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts,” *The Lancet*, vol. 401, no. 10372, pp. 215–225, Jan. 2023, doi: 10.1016/S0140-6736(22)02079-7.

[12] S. H. Bani Hani and M. M. Ahmad, “Machine-learning Algorithms for Ischemic Heart Disease Prediction: A Systematic Review,” *Curr Cardiol Rev*, vol. 19, no. 1, Jan. 2023, doi: 10.2174/1573403X18666220609123053.

[13] R. Alizadehsani *et al.*, “Machine learning-based coronary artery disease diagnosis: A comprehensive review,” *Comput Biol Med*, vol. 111, p. 103346, Aug. 2019, doi: 10.1016/j.compbiomed.2019.103346.

[14] Y. Kokubo *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*, vol. 28, no. 11, pp. 1355–1361, Nov. 2015, doi: 10.1093/ajh/hpv038.

[15] Y. Kokubo *et al.*, “Impact of High-Normal Blood Pressure on the Risk of Cardiovascular Disease in a Japanese Urban Cohort,” *Hypertension*, vol. 52, no. 4, pp. 652–659, Oct. 2008, doi: 10.1161/HYPERTENSIONAHA.108.118273.

[16] Y. M. Nakao *et al.*, “Effectiveness of nationwide screening and lifestyle intervention for abdominal obesity and cardiometabolic risks in Japan: The metabolic syndrome and comprehensive lifestyle intervention study on nationwide database in Japan (MetS ACTION-J study),” *PLoS One*, vol. 13, no. 1, p. e0190862, Jan. 2018, doi: 10.1371/journal.pone.0190862.

[17] H. Iso *et al.*, “Risk Classification for Metabolic Syndrome and the Incidence of Cardiovascular Disease in Japan With Low Prevalence of Obesity: A Pooled Analysis of 10 Prospective Cohort

Studies," *J Am Heart Assoc*, vol. 10, no. 23, Dec. 2021, doi: 10.1161/JAHA.121.020760.

[18] E. Imai *et al.*, "Estimation of glomerular filtration rate by the MDRD study equation modified for Japanese patients with chronic kidney disease," *Clin Exp Nephrol*, vol. 11, no. 1, pp. 41–50, Mar. 2007, doi: 10.1007/s10157-006-0453-4.

[19] Y. Kokubo, M. Watanabe, A. Higashiyama, Y. M. Nakao, F. Nakamura, and Y. Miyamoto, "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*, vol. 7, no. 11, Jun. 2018, doi: 10.1161/JAHA.117.007720.

[20] H. Tunstall-Pedoe, K. Kuulasmaa, P. Amouyel, D. Arveiler, A. M. Rajakangas, and A. Pajak, "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*, vol. 90, no. 1, pp. 583–612, Jul. 1994, doi: 10.1161/01.CIR.90.1.583.

[21] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Crit Care*, vol. 9, no. 1, p. 112, 2005, doi: 10.1186/cc3045.

[22] X. Su *et al.*, "Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model," *J Clin Lab Anal*, vol. 34, no. 9, Sep. 2020, doi: 10.1002/jcla.23421.

[23] P. Unnikrishnan, D. K. Kumar, S. Poosapadi Arjunan, H. Kumar, P. Mitchell, and R. Kawasaki, "Development of Health Parameter Model for Risk Prediction of CVD Using SVM," *Comput Math Methods Med*, vol. 2016, pp. 1–7, 2016, doi: 10.1155/2016/3016245.

[24] Y.-J. Son, H.-G. Kim, E.-H. Kim, S. Choi, and S.-K. Lee, "Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients," *Healthc Inform Res*, vol. 16, no. 4, p. 253, 2010, doi: 10.4258/hir.2010.16.4.253.

[25] H. Yang, Z. Chen, H. Yang, and M. Tian, "Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison," *IEEE Access*, vol. 11, pp. 23366–23380, 2023, doi: 10.1109/ACCESS.2023.3253885.

[26] L. Bloch and C. M. Friedrich, “Data analysis with Shapley values for automatic subject selection in Alzheimer’s disease data sets using interpretable machine learning,” *Alzheimers Res Ther*, vol. 13, no. 1, p. 155, Sep. 2021, doi: 10.1186/s13195-021-00879-4.

[27] T. Kawai *et al.*, “Carotid plaque score and intima media thickness as predictors of stroke and mortality in hypertensive patients,” *Hypertension Research*, vol. 36, no. 10, pp. 902–909, Oct. 2013, doi: 10.1038/hr.2013.61.

[28] V. Nambi *et al.*, “Carotid Intima-Media Thickness and Presence or Absence of Plaque Improves Prediction of Coronary Heart Disease Risk,” *J Am Coll Cardiol*, vol. 55, no. 15, pp. 1600–1607, Apr. 2010, doi: 10.1016/j.jacc.2009.11.075.

[29] Y. Kokubo, M. Watanabe, A. Higashiyama, Y. M. Nakao, F. Nakamura, and Y. Miyamoto, “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*, vol. 7, no. 11, Jun. 2018, doi: 10.1161/JAHA.117.007720.

[30] C. Ji *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*, vol. 36, no. 10, pp. 933–939, Oct. 2022, doi: 10.1038/s41371-021-00598-1.

[31] S. P. Whelton *et al.*, “Association of Normal Systolic Blood Pressure Level With Cardiovascular Disease in the Absence of Risk Factors,” *JAMA Cardiol*, vol. 5, no. 9, p. 1011, Sep. 2020, doi: 10.1001/jamacardio.2020.1731.

[32] J. Li *et al.*, “Evaluation of Optimal Diastolic Blood Pressure Range Among Adults With Treated Systolic Blood Pressure Less Than 130 mm Hg,” *JAMA Netw Open*, vol. 4, no. 2, p. e2037554, Feb. 2021, doi: 10.1001/jamanetworkopen.2020.37554.

[33] L.-L. Guo *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*, vol. 8, Apr. 2021, doi: 10.3389/fcvm.2021.649181.

[34] I. Saito *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," *J Atheroscler Thromb*, vol. 27, no. 4, pp. 363–374, Apr. 2020, doi: 10.5551/jat.50385.

[35] J. Dong *et al.*, "The Associations of Lipid Profiles With Cardiovascular Diseases and Death in a 10-Year Prospective Cohort Study," *Front Cardiovasc Med*, vol. 8, Nov. 2021, doi: 10.3389/fcvm.2021.745539.

[36] X. Zhao, D. Wang, and L. Qin, "Lipid profile and prognosis in patients with coronary heart disease: a meta-analysis of prospective cohort studies," *BMC Cardiovasc Disord*, vol. 21, no. 1, p. 69, Dec. 2021, doi: 10.1186/s12872-020-01835-0.

[37] A. V. Poznyak, L. Litvinova, P. Poggio, V. N. Sukhorukov, and A. N. Orekhov, "Effect of Glucose Levels on Cardiovascular Risk," *Cells*, vol. 11, no. 19, p. 3034, Sep. 2022, doi: 10.3390/cells11193034.

[38] H. K. R. Riise *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*, vol. 9, no. 1, p. e001928, Feb. 2021, doi: 10.1136/bmjdrc-2020-001928.

[39] E. Selvin, J. Coresh, S. H. Golden, F. L. Brancati, A. R. Folsom, and M. W. Steffes, "Glycemic Control and Coronary Heart Disease Risk in Persons With and Without Diabetes," *Arch Intern Med*, vol. 165, no. 16, p. 1910, Sep. 2005, doi: 10.1001/archinte.165.16.1910.

[40] P. Charoen *et al.*, "Mendelian Randomisation study of the influence of eGFR on coronary heart disease," *Sci Rep*, vol. 6, no. 1, p. 28514, Jun. 2016, doi: 10.1038/srep28514.

[41] J. Jankowski, J. Floege, D. Fliser, M. Böhm, and N. Marx, "Cardiovascular Disease in Chronic Kidney Disease," *Circulation*, vol. 143, no. 11, pp. 1157–1172, Mar. 2021, doi: 10.1161/CIRCULATIONAHA.120.050686.

[42] J. J. Brugts, A. M. Knetsch, F. U. S. Mattace-Raso, A. Hofman, and J. C. M. Witteman, "Renal Function and Risk of Myocardial Infarction in an Elderly Population," *Arch Intern Med*, vol. 165, no.

22, p. 2659, Dec. 2005, doi: 10.1001/archinte.165.22.2659.

[43] A. H. E. M. Maas and Y. E. A. Appelman, “Gender differences in coronary heart disease,” *Netherlands Heart Journal*, vol. 18, no. 12, pp. 598–603, Nov. 2010, doi: 10.1007/s12471-010-0841-y.

[44] T. Shah, N. Palaskas, and A. Ahmed, “An Update on Gender Disparities in Coronary Heart Disease Care,” *Curr Atheroscler Rep*, vol. 18, no. 5, p. 28, May 2016, doi: 10.1007/s11883-016-0574-5.

X

Legends for Tables and Figures

Table 1: Characteristics of study participants with and without Coronary Heart Disease incidence (healthy Japanese, aged 30–84 years, Suita study at baseline).

Table 2: Performance metrics and their 95% confidence intervals of different machine learning approaches.

Figure 1: Top most important variables for incident Coronary Heart Disease (CHD) using SHapley Additive exPlanations (SHAP) values.

Figure 2: SHAP Dependency Plots Illustrate the Relationship Between Key Variables and Coronary Heart Disease Risk.

Supplementary files

Table S1: List of the variables included in the prediction model.

Table S2: Interpretation of the model performance metrics.

Figure S1: Percentage of missing data across all variables immediately prior to imputation.

Figure S2: Correlation Coefficients heat map.

Figure S3: The heat plot of all variables for incident Coronary Heart Disease based on SHAP values.

Tables

Table 1: Characteristics of study participants with and without coronary heart disease incidence (healthy Japanese, aged 30–84 years, Suita study at baseline).

	Coronary heart disease, n (%)		**p-value**
	No 6955 (95.8)	**Yes** 305 (4.2)	
Age, Years	55.0 [44.0;65.0]	63.0 [56.0;71.0]	<0.001
Sex, n (%)			<0.001
Male	3147 (45.2%)	202 (66.2%)	
Female	3808 (54.8%)	103 (33.8%)	
SBP, mmHg	123 [110;137]	138 [125;153]	<0.001
DBP, mmHg	77.0 [70.0;85.0]	83.0 [74.0;89.0]	<0.001
eGFR, mL/min/1.73 m2	104 (32.2)	95.3 (63.3)	0.014
IMT_cMax, mm	1.00 [0.80;1.10]	1.10 [1.00;1.30]	<0.001
BMI, kg/m2	22.5 (3.10)	23.3 (3.26)	<0.001
Body Fat, %	23.2 (6.32)	22.6 (7.06)	0.154
Waist Circumference, cm	80.0 [73.0;86.0]	83.0 [77.0;90.0]	<0.001
HDL-c, mg/dL	53.0 [44.0;63.0]	46.0 [38.0;56.0]	<0.001
non-HDL-c, mg/dL	152 (36.9)	172 (40.5)	<0.001
Triglycerides, mg/dL	98.0 [70.0;143]	121 [90.0;174]	<0.001
Calcium, mg/dL	9.35 (0.46)	9.34 (0.43)	0.612
Fructosamine, μmol/L	251 [237;266]	257 [242;276]	<0.001
Glucose, mg/dL	95.0 [89.0;101]	100 [93.0;109]	<0.001
White Blood Cell Count, /mm^3	5.33 [4.48;6.36]	5.65 [4.81;6.78]	<0.001
Red Blood Cell Count, $10^3/mm^3$	4.53 (0.44)	4.60 (0.46)	0.008

	Coronary heart disease, n (%)		p-value
	No 6955 (95.8)	**Yes** 305 (4.2)	
Hemoglobin, g/dL	13.9 (1.56)	14.3 (1.49)	<0.001
Smoking Status			<0.001
Current	2019 (29.0%)	110 (36.1%)	
Past	1091 (15.7%)	79 (25.9%)	
Never	3845 (55.3%)	116 (38.0%)	
Drinking Status			0.266
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%)	<0.001
Hypertension, n (%)	2056 (29.6%)	172 (56.4%)	<0.001
Diabetes Mellitus, n (%)	426 (6.13%)	49 (16.1%)	<0.001
Dyslipidemia, n (%)	5280 (75.9%)	265 (86.9%)	<0.001

Coronary heart disease (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 (standard deviation) for continuous variables with approximately normally distribution or by median [interquartile range] with skewed distribution and n (%) for categorical variables. Differences in characteristics were evaluated by using the unpaired Student's t-test, Wilcoxon rank sums test, or Chi squared test.

Abbreviations: BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; non-HDL-c, non-high-density lipoprotein cholesterol; eGFR, estimated glomerular filtration rate; IMT_cMax, the maximum intima-media thickness of common carotid arteries.

Table 2: Performance metrics and their 95% confidence intervals of different machine learning approaches.

Model	Accuracy	Sensitivity	Precision	AUC	F1 Score
LR	0.77 (0.75 - 0.79)	0.53 (0.49 - 0.57)	0.64 (0.60 - 0.69)	0.7 (0.68 - 0.72)	0.58 (0.55 - 0.62)
RF	0.97 (0.96 - 0.97)	0.89 (0.87 - 0.91)	1.0 (1 - 1)	0.94 (0.93 - 0.96)	0.94 (0.55 - 0.62)
SVM	0.94 (0.92 - 0.95)	0.82 (0.78 - 0.85)	0.96 (0.60 - 0.69)	0.9 (0.89 - 0.92)	0.88 (0.86 - 0.90)
XGBoost	0.95 (0.94 - 0.96)	0.82 (0.79 - 0.85)	1.0 (1 - 1)	0.91 (0.89 - 0.92)	0.9 (0.89 - 0.92)
LightGBM	0.3 (0.28 - 0.33)	1.0 (1 - 1)	0.3 (0.29 - 0.33)	0.5 (0.49 - 0.57)	0.47 (0.89 - 0.92)

Abbreviation: AUC, Area Under the Curve; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XG Boost, eXtreme Gradient Boost; Light GBM, Light Gradient Boosted Machine

Figures

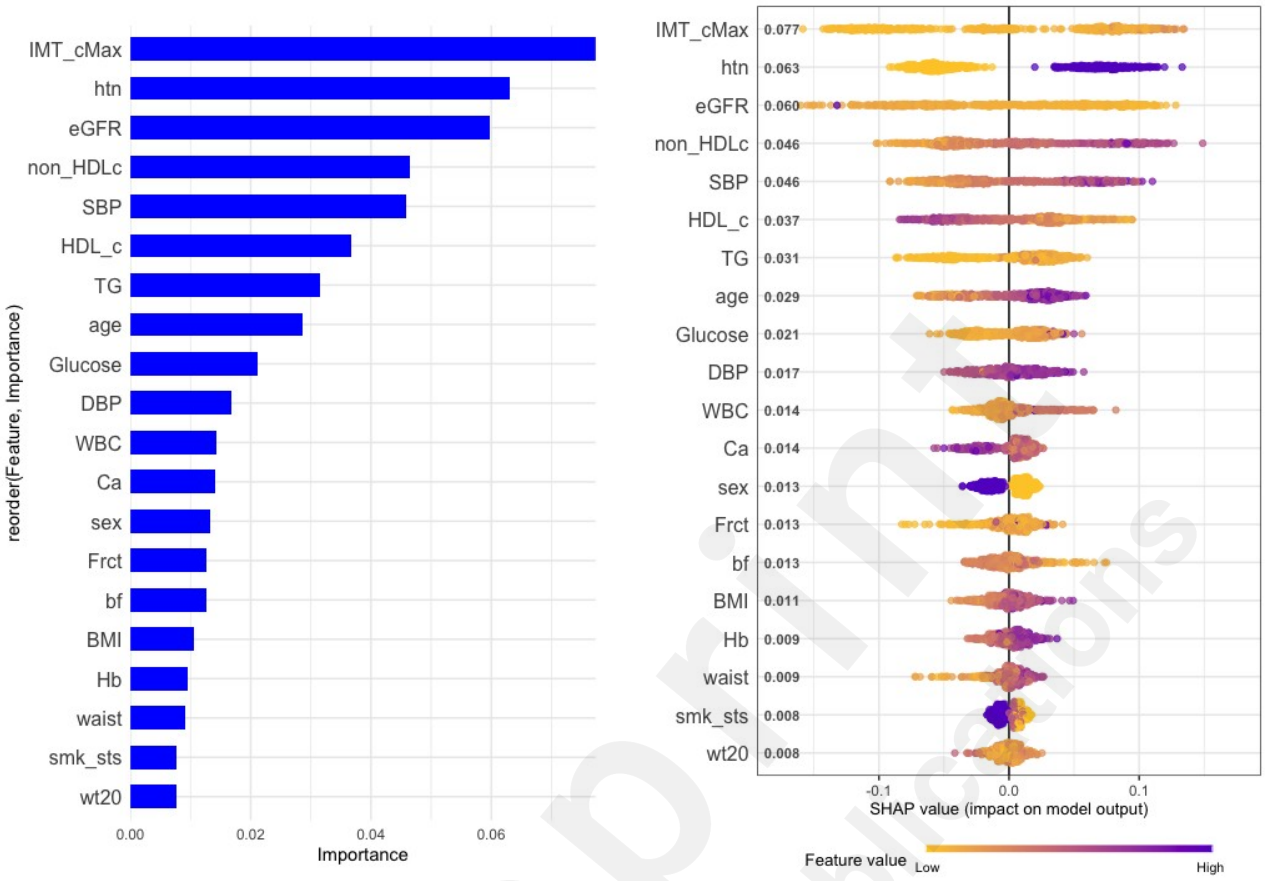

Figure 1: Top most important variables for incident Coronary Heart Disease (CHD) using SHapley Additive exPlanations (SHAP) values.

(**Left**) The contribution levels of all variables to CHD. A variable is represented by each bar in the plot, and the length of the bar indicates the extent of the variable's contribution to CHD. (**Right**) The relationships between the variables and CHD are revealed by the heat plot of SHAP values. Purple indicates a positive relationship, while yellow indicates a negative relationship. Through it, we can obtain a basic understanding of the relationship between the value of a particular variable and its influence on prediction. Each data point corresponds to a certain participant and their corresponding Shapley value for a particular variable. The position of a data point on this plot is defined by the variable's importance, which is represented on the y-axis, and the Shapley value, which is represented on the x-axis.

Abbreviations: IMT_cMax, the maximum intima-media thickness of common carotid arteries; htn, hypertension; eGFR, estimated glomerular filtration rate; non-HDL-c, non-high-density lipoprotein cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglycerides; WBC, white blood cell; Ca, Calcium level; Frct: Fructosamine; bf: body fat; BMI, body mass index; Hb, Hemoglobin; smk_sts, smoking status; wt20, weight at age of 20.

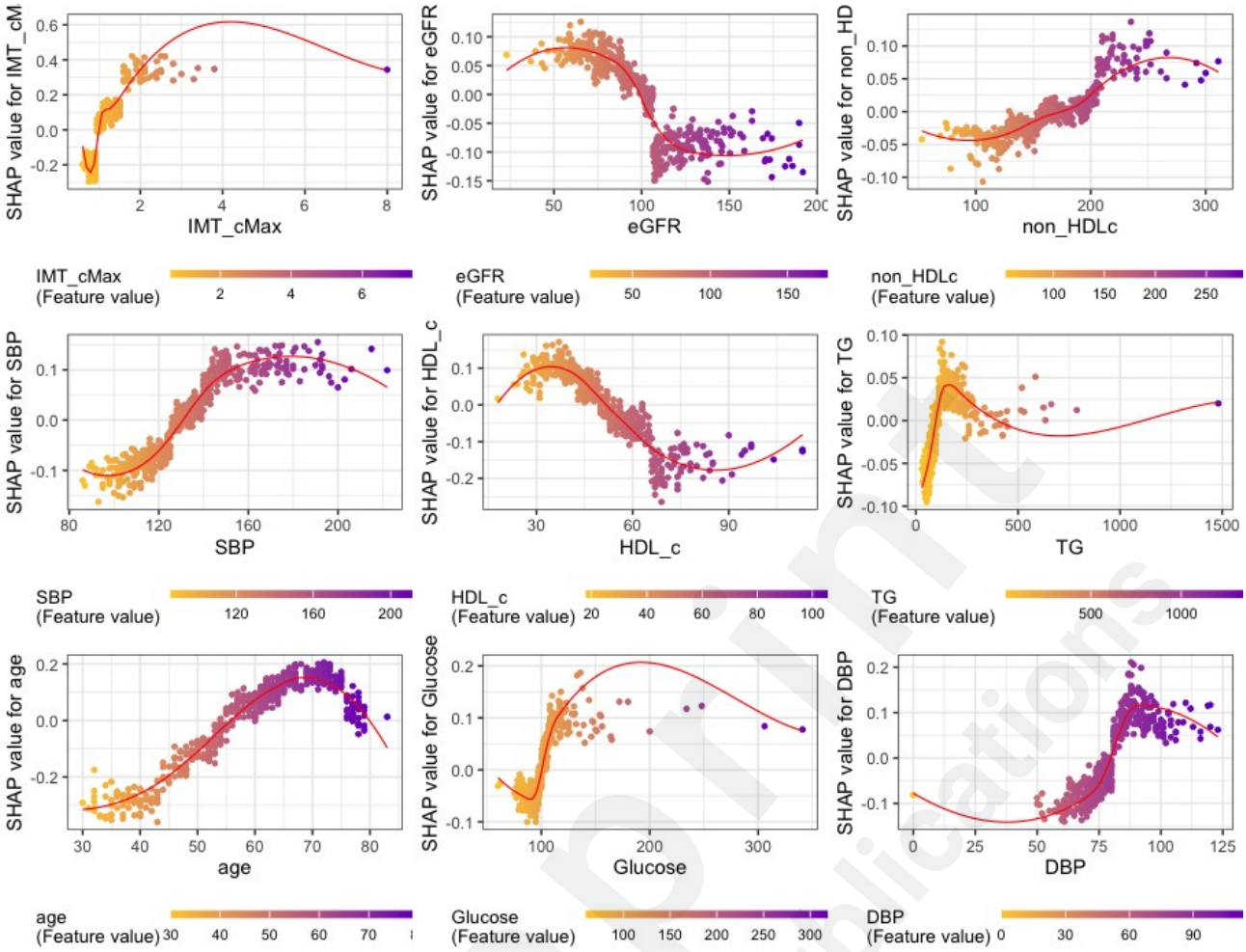

Figure 2: SHAP Dependency Plots Illustrate the Relationship Between Key Variables and Coronary Heart Disease Risk.

Abbreviations: IMT_cMax, the maximum intima-media thickness of common carotid arteries; eGFR, estimated glomerular filtration rate; non-HDL-c, non-high-density lipoprotein cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglycerides.

Supplementary Files