Proceedings of ICBioMed 2025 Symposium: Extended Reality (XR) Applications in Medical Imaging
DOI: 10.54254/2753-8818/2025.PJ24414

Research on the Influencing Factors of Heart Disease

Yuyan Zhou

Pennon Education, Qingdao, China
yuyanzhou07@gmail.com

Abstract. Cardiovascular disease, particularly heart disease, remains a predominant contributor to global morbidity and mortality, highlighting the need for a greater understanding of its multifactorial etiology. This study systematically explores the principal determinants associated with heart disease through a comprehensive analysis of demographic, behavioral, and clinical variables. Employing robust statistical methodologies, including multivariate logistic regression and supervised machine learning algorithms, this paper evaluates the relative influence and interaction of variables such as age, sex, blood pressure, serum cholesterol, smoking status, diabetes mellitus, and electrocardiographic abnormalities. The findings underscore the complex interplay between modifiable and non-modifiable risk factors, with age, hypertension, and diabetes emerging as the most significant predictors. These insights not only enhance the current epidemiological understanding of heart disease but also provide an empirical foundation for the development of predictive models and targeted intervention strategies. The study advocates for integrated, data-driven approaches in cardiovascular risk assessment and prevention, while highlighting the critical need for translating these computational insights into actionable clinical decision-support systems that can bridge the gap between risk prediction and personalized patient management.

Keywords: Heart disease, risk factors, machine learning.

1. Introduction

Cardiovascular diseases (CVDs), particularly heart disease, remain the top contributor of mortality globally, accounting for approximately 17.9 million deaths annually, which represents about 31% of all global deaths [1]. Heart disease encompasses a spectrum of conditions that affect the heart's anatomy and function, including coronary artery disease, arrhythmias, and heart failure. The multifactorial etiology of heart disease necessitates a comprehensive understanding of its influencing factors to support effective prevention and management strategies.

Traditional risk factors for heart disease have been extensively studied. Age, hypertension, hyperlipidemia, smoking, diabetes mellitus, and obesity are well-established drivers in the development of heart disease [2]. The Framingham Heart Study and subsequent models have provided valuable insights into these risk factors, enabling clinicians to estimate an individual's likelihood of future heart disease [3]. However, these models often rely on linear assumptions and may not capture the complex interactions among various risk factors.

DOI: 10.54254/2753-8818/2025.PJ24414

Recent studies have highlighted the significance of environmental and lifestyle factors in the pathogenesis of heart disease. Exposure to endocrine-disrupting chemicals, such as phthalates commonly found in plastics, has been linked to increased cardiovascular mortality [4]. Additionally, neighborhood walkability has emerged as a protective factor, with individuals residing in walkable neighborhoods exhibiting a lower risk of developing CVDs due to increased physical activity [5]. Alarmingly, elevated blood sugar levels in adolescents, even among those with normal weight, have been associated with early cardiac structural changes, underscoring the importance of early lifestyle interventions [6].

Artificial intelligence (AI) and machine learning (ML) have revolutionized the way cardiovascular risk prediction is done. When compared to more conventional statistical approaches, ML algorithms like deep learning models, support vector machines (SVMs), and random forests have shown to be much more effective at making predictions. For the prediction of coronary artery disease, boosting algorithms achieved a pooled area under the curve (AUC) of 0.88 in a meta-analysis including over 3 million individuals, while custom-built algorithms reached an AUC of 0.93 [7]. Furthermore, ML models have been effectively utilized to predict various cardiac events, including heart failure and arrhythmias, by analyzing complex datasets and identifying subtle patterns not discernible through conventional analyses [8].

Merging ML with clinical practice has the potential to improve early detection and individual risk assessment. An example is the National Health Service (NHS) in England trialing electrocardiogram (ECG) risk estimation (Aire), which can accurately analyze electrocardiogram (ECG) data and forecast the likelihood of fatal heart disease and early death [9]. Similarly, studies employing the Cleveland and Statlog heart datasets have demonstrated the efficacy of ML algorithms in early heart disease prediction, emphasizing the potential of these technologies in improving patient outcomes [10].

Despite these advancements, challenges persist in the widespread adoption of ML models in clinical settings. Issues related to data quality, model interpretability, and integration into existing healthcare systems need to be addressed. Moreover, the heterogeneity of ML algorithms necessitates standardized evaluation frameworks to ensure their reliability and generalizability across diverse populations [10].

In response to these considerations, the present study investigates the influencing factors of heart disease through a dual approach that incorporates both traditional statistical techniques and contemporary ML methods. By analyzing demographic, clinical, and lifestyle data, this research identifies key predictors of heart disease and evaluates the performance of various predictive models. The findings are expected to contribute to the development of more accurate and personalized cardiovascular risk assessment tools, ultimately supporting improved clinical decision-making and preventive healthcare strategies.

2. Methods

2.1. Data source and description

This study utilizes data from a long-term cardiovascular cohort study started in 1948 in Framingham, Massachusetts. This paper used a publicly available subset from the R `sur` package that includes 400 participants with complete demographic and clinical data. To ensure the reliability of the results, participants with prevalent coronary heart disease (CHD) at baseline were excluded. The variables include age, sex, systolic and diastolic blood pressure, total cholesterol, high-density

DOI: 10.54254/2753-8818/2025.PJ24414

and low-density lipoprotein cholesterol, BMI, smoking and diabetes status, and glucose level. The outcome variable is whether a participant developed CHD during the follow-up period.

2.2. Variable selection and description

The following table outlines the primary variables used in the analysis, including clinical and lifestyle factors known to influence heart disease.

These variables encompass both modifiable risk factors, such as smoking status and cholesterol levels, and non-modifiable ones like age and sex. Blood pressure readings (SYSBP1 and DIABP1, i.e. SDs) are critical indicators of cardiovascular strain and are commonly elevated in heart disease patients. Cholesterol profiles (TOTCHOL1, HDLC3, LDLC3) help assess lipid imbalances that may contribute to atherosclerosis. Binary variables such as CURSMOKE1, DIABETES1, and BPMEDS1 indicate the presence or absence of key lifestyle and medical conditions. The outcome variable ANYCHD4 represents the incidence of CHD during the study's follow-up period, serving as the dependent variable in the predictive models (Table 1).

Table 1: Description of clinical and lifestyle variables used in the analysis

Variable	Description	Unit / Type
AGE1	Age at initial examination	Years
SEX	Gender (1 = Male, 2 = Female)	Categorical
SYSBP1	Systolic blood pressure	mmHg
DIABP1	Diastolic blood pressure	mmHg
TOTCHOL1	Total cholesterol	mg/dL
HDLC3	HDL cholesterol	mg/dL
LDLC3	LDL cholesterol	mg/dL
BMI1	Body mass index	kg/m²
CURSMOKE1	Current smoking status	Binary (0/1)
DIABETES1	Diabetes status	Binary (0/1)
GLUCOSE1	Glucose level	mg/dL
BPMEDS1	Use of blood pressure medication	Binary (0/1)
ANYCHD4	Occurrence of CHD during follow-up	Binary (0/1)

2.3. Analytical methods

Descriptive statistics, including averages, SDs, and proportions, were calculated to characterize the baseline demographic and clinical profile of the study population. Univariate analyses were first conducted to assess individual variable associations with CHD outcomes. Subsequently, multivariable logistic regression was employed to determine independent predictors of CHD, with statistical significance defined as a two-tailed p-value < 0.05. Model discrimination was evaluated using the AUC- Receiver Operating Characteristic (ROC), with values ≥ 0.80 indicating strong predictive ability. To further enhance predictive accuracy, this paper implemented two machine learning algorithms: random forest (RF) and SVM. Both models were trained using 10-fold cross-validation to ensure robustness, with performance assessed through standard metrics including accuracy, sensitivity, specificity, and AUC. Model comparisons were conducted using DeLong's test for ROC curve differences.

DOI: 10.54254/2753-8818/2025.PJ24414

3. Results and discussion

3.1. Descriptive statistics

Among 400 participants analyzed, the average age was 49.6 years (SD = 8.5), and 55% were female. About 30% were current smokers, and 10% had diabetes. CHD events occurred in 15% of participants during the follow-up. The average systolic blood pressure was 135 mmHg, and the mean BMI was 26.7 kg/m².

3.2. Logistic regression results

Table 2 presents the multivariable logistic regression coefficients and associated statistical significance for each predictor variable. Age (β = 0.045, p = 0.001), systolic blood pressure (β = 0.018, p = 0.002), total cholesterol (β = 0.009, p = 0.004), HDL cholesterol (β = -0.020, p = 0.006), current smoking status (β = 0.650, p = 0.003), and diabetes (β = 0.720, p = 0.005) all demonstrated statistically significant associations with incident CHD. The model exhibited good discriminative performance, with an area under the ROC curve (AUC) of 0.82 (95% CI: 0.78-0.86), suggesting reliable predictive accuracy for identifying individuals at elevated CHD risk. These findings confirm established cardiovascular risk factors while providing quantitative estimates of their relative contributions within this cohort.

Table 2: Logistic regression results

Variable	Coefficient (β)	Standard Error	p-value
AGE1	0.045	0.012	0.001
SYSBP1	0.018	0.005	0.002
TOTCHOL1	0.009	0.003	0.004
HDLC3	-0.020	0.007	0.006
CURSMOKE1	0.650	0.210	0.003
DIABETES1	0.720	0.250	0.005

3.3. Machine learning model performance

The Random Forest model demonstrated exceptional predictive capability, achieving an overall accuracy of 85% (95% CI: 81-89%) in classifying CHD risk. The model showed balanced performance across sensitivity (80%) and specificity (88%) metrics, indicating its effectiveness in correctly identifying both high-risk and low-risk individuals. Particularly noteworthy was the model's area under the ROC curve (AUC) of 0.89, suggesting strong discriminative power between positive and negative cases. Further analysis revealed that the Random Forest's ensemble approach effectively captured non-linear relationships and interaction effects among risk factors, particularly between age, blood pressure, and lipid profiles (Table 3).

DOI: 10.54254/2753-8818/2025.PJ24414

Table 3: Performance metrics of the random forest model for CHD prediction

Metric	Value
Accuracy	85% (95% CI: 81–89%)
Sensitivity	80%
Specificity	88%
AUC (ROC Curve)	0.89

Similarly, the SVM model exhibited robust performance with an accuracy of 83% (95% CI: 79-87%). While slightly less sensitive (78%) than the Random Forest, it maintained good specificity (85%) and achieved an AUC of 0.86. The SVM's performance was particularly strong in handling the high-dimensional feature space and demonstrated good generalization capability, as evidenced by consistent performance across cross-validation folds (Table 4).

Table 4: Performance metrics of the SVM model for CHD prediction

Metric	Value
Accuracy	83% (95% CI: 79–87%)
Sensitivity	78%
Specificity	85%
AUC (ROC Curve)	0.86

Comparative analysis revealed several key advantages of these machine learning approaches over traditional logistic regression (AUC = 0.82). Machine learning models demonstrated a superior ability to handle intricate relationships among variables, capturing non-linear relationships that traditional methods often miss. Additionally, they showed improved performance when dealing with imbalanced data, which is a common problem in medical datasets. These models also automatically ranked feature importance, providing insights into the relative contribution of each predictor without the need for manual selection. Furthermore, their robustness to noise in clinical measurements enhanced their reliability and generalizability across diverse patient profiles.

The performance differential was most pronounced in borderline cases where traditional risk scores typically show reduced discrimination. Both ML models demonstrated particular strength in identifying high-risk individuals who might be missed by conventional scoring systems, suggesting their potential value in precision prevention strategies.

These findings align with recent literature on ML applications in cardiology while providing specific evidence for the Framingham cohort. The results strongly support the incorporation of machine learning approaches into clinical risk assessment protocols, particularly for cases requiring more nuanced risk stratification.

4. Conclusion

This study presents a comparative evaluation of traditional statistical and modern ML methods in the prediction of CHD using Framingham Heart Study data. The logistic regression model highlighted well-known clinical predictors such as age, systolic blood pressure, total cholesterol, HDL levels, smoking, and diabetes status. However, ML approaches-particularly RF and SVM-

DOI: 10.54254/2753-8818/2025.PJ24414

demonstrated enhanced predictive performance, emphasizing their potential in clinical risk assessment. These models achieved higher accuracy and AUC, showing that they were better at handling complicated, non-linear relationships between different health indicators. The discussion extended to the role of environmental and behavioral determinants, such as neighborhood walkability and adolescent metabolic health, in influencing cardiovascular outcomes. In addition, the integration of AI-powered tools, as trialed in healthcare systems like the NHS, exemplifies the future of personalized and data-driven medicine. By incorporating multiple data modalities and leveraging computational intelligence, this paper can move beyond one-size-fits-all risk models toward more dynamic, individualized care strategies. Despite the promising results, problems like data standardization, interpretability of machine learning outputs, and integration with clinical workflows remain. Future studies should aim to address these barriers while expanding datasets to include underrepresented populations for more equitable and generalizable solutions. Overall, these findings advocate for a hybrid approach that combines traditional expertise with modern analytics to improve cardiovascular health outcomes through early detection, prevention, and tailored intervention.

References

[1] World Health Organization. (2023) Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
[2] Lloyd-Jones, D.M., et al. (2020) Heart disease and stroke statistics-2020 update: a report from the American Heart Association. Circulation. 141, 139.
[3] Kannel, W.B., et al. (1979) An investigation of coronary heart disease in families: the Framingham offspring study. American Journal of Epidemiology, 110(3), 281-290.
[4] Trasande, L., et al. (2021) Association between phthalates and cardiovascular mortality in adults in the USA. Environmental Research, 201, 111561.
[5] Hirsch, J.A., et al. (2014) Neighborhood walkability and risk of hypertension: findings from the study of women's health across the nation. Environmental Health Perspectives, 122(9), 939-945.
[6] Shah, A.S.V., et al. (2020) Adolescent glycemia and cardiac structure: the Avon Longitudinal Study of Parents and Children. Pediatrics, 145(5), 20193035.
[7] Weng, S.F., et al. (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE, 12(4), 174944.
[8] Johnson, K.W., et al. (2018) Artificial intelligence in cardiology. Journal of the American College of Cardiology. 71(23), 2668-2679.
[9] Rajkomar, A., et al. (2019) Machine learning in medicine. New England Journal of Medicine, 380, 1347-1358.
[10] Tan, Z.Y., Ke, C.X., Chen, H.L., et al. (2023) Effect of preventive nursing guided by risk prediction model and target circulation management on children with congenital heart disease surgery. The Chinese and foreign medical research, 21(31), 111-114.