Medical Journal of Public Health
ISSN:
DOI:

RESEARCH ARTICLE OPEN ACCESS

Predicting Heart Disease Using Demographic, Behavioural, and Clinical Risk Factors: A Public Health Approach

Dr Rajkiran Tiku[1], Dr Anupam Biswas[2], Dr.S.Senthilkumar[3], P. Krishnan[4]

[1]*PT, HOD/Professor, Physiotherapy, Cardiorespiratory Physiotherapy, Suryadatta Institute of Health Sciences - College of Physiotherapy, City / Pincode: Pune /411021, ORCID ID : 0009-0008-8201-6976, Email ID: rajkiran.tiku@gmail.com*
[2]*Assistant Professor , Department of Kayachikitsa, Dayanand Ayurvedic Medical College and Hospital, Siwan, Bihar, BAMS, MD, The West Bengal University of Health Sciences, orcid id- 0009000845480366, email- dranupam9378@gmail.com*
[3]*Professor, School of physiotherapy, Dhanalakshmi Srinivasan University Trichirappalli Tamilnadu India, Orcid Id: 0000-0003-4604-6488, Email: drsenmpt@gmail.com*
[4]*PT, Research Scholar & Assistant Professor, School of Physiotherapy, Dhanalakshmi Srinivasan University, Samayapuram, Trichy - 621112, Orcid Id: 0009-0003-8312-7436, Email Id: drkrishnan.mpt@gmail.com*

Received:- 14/06/2026, Revised:- 15/07/2026, Accepted:- 21/07/2026, Published:- 29/07/2026

Abstract:

Identifying those at high risk of heart disease early is critical to population-level prevention and effective use of health care resources. In this study, machine learning models were created based on demographic, behavioral and clinical risk factors to assist public health screening. The data were used for a cross-sectional analysis. 5,510 participants were included after removing duplicate observations, with 527 having heart disease. The models were trained with Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost on 80% of the data and tested on the remaining 20% of the data which was a stratified split. The Synthetic Minority Over-sampling Technique was employed on the Training data to address class imbalance. Accuracy, precision, recall, F1 score, ROC–AUC, confusion matrix and five-fold cross-validation were used to assess the performance. The prevalence of heart disease was 9.56%. Participants with heart disease were older with higher rates of smoking history, hypertension, diabetes, high cholesterol, previous stroke, and family history of heart attack. Logistic Regression performed the best results in terms of overall performance with an accuracy of 0.790, a recall of 0.829, an F1 score of 0.430, an ROC–AUC of 0.876, and a cross-validated ROC–AUC of 0.865 ± 0.018. The most important predictors were age, followed by hypertension, smoking history, use of cholesterol-lowering drugs, history of stroke, sex, family history and diabetes. Interpretable machine learning plays a crucial role in population-based heart disease risk stratification and targeted preventive interventions.

Keywords: heart disease; machine learning; public health; risk prediction; cardiovascular prevention

1. Introduction

Cardiovascular diseases (CVDs) are still the leading cause of death globally and are a major burden on the healthcare system in terms of premature deaths, disability, and increased healthcare costs. Heart disease has been a leading cause of morbidity, productivity loss, and high healthcare utilization for years to come and causes millions of deaths each year. Despite the improved survival rates due to advances in preventive medicine and clinical management, the global burden of cardiovascular disease (CVD) has remained unchanged due to the increase in ageing population, obesity, diabetes mellitus, hypertension, and other cardiometabolic diseases (Benjamin et al., 2019; Virani et al., 2020). Many cardiovascular diseases are driven by the key pathological process of atherosclerosis, which may develop over decades before symptoms appear, emphasizing the need for early detection of high-risk individuals (Libby, 2021). As a result, predicting the risk of heart disease has now become a major public health priority, as early detection opens the opportunity to improve lifestyle changes, target screening, prevent disease, and allocate health-care resources (Arnett et al., 2019).

The development of heart disease is the result of a combination of demographic, behavioural and clinical determinants, which all contribute to cardiovascular risk. Older age continues to be an important risk factor because the buildup of vascular injury and metabolic changes are known to speed up over time. Sex differences also play a role in disease incidence, and racial and ethnic differences remain to create disparities in cardiovascular outcomes that result from differences in socioeconomic factors, access to health care, and environmental exposures (Velarde et al., 2023). Cigarette smoking, physical inactivity, unhealthy diets and other adverse lifestyle behaviors also contribute to cardiovascular risk by impairing endothelial function, inflammation, obesity, and metabolic abnormalities (Arnett et al., 2019). Furthermore, established clinical factors such as hypertension, diabetes mellitus, hypercholesterolaemia, prior stroke and family history of cardiovascular disease significantly increase the risk of developing cardiovascular disease (Benjamin et al., 2019). Recently, evidence of the role of biological and epigenetic mechanisms in the susceptibility to disease has emerged, offering further insights into cardiovascular risk stratification (Kremer et al., 2025).

The increasing development of electronic health records, national health surveys, and large-scale biomedical datasets has led to a growing use of machine learning techniques in cardiovascular risk prediction. Traditional statistical methods, however, cannot model complex, nonlinear relationships, high-dimensional interactions, and heterogeneous patient characteristics as well, which may be important (Beam & Kohane, 2018). Many algorithms have shown good predictive accuracy in various cardiovascular applications, such as decision trees, random forests, support vector machines, deep learning, and gradient boosting algorithms (Krittanawong et al., 2019). AI has already demonstrated significant promise in clinical screening, with automated ECG analysis and cardiac dysfunction detection already considered valuable innovations (Attia et al., 2019). Recent reviews and bibliometric analyses have also highlighted and confirmed the widespread application of machine learning in cardiovascular medicine, heart failure research, wearable health technologies, and precision medicine, underscoring its potential in the context of modern practice of public health (Ahsan & Siddique, 2022; Dutta et al., 2019; Yu et al., 2023).

In addition, many prediction models focus on prediction accuracy rather than interpretability of the prediction model and external applicability and implementation in public health. Statutory statements in science in recent years have highlighted the need for prediction models to be both clinically useful and statistically sound in population health-based decision-making (Khan et al., 2026). Thus, the need for interpretable machine learning models built on nationally representative data sets that combine both demographic and behavioural and clinical risk factors with evidence-based findings for population-level cardiovascular screening and prevention persists (Sritharan et al., 2026).

The current study seeks to create and assess machine learning models to predict the presence of heart disease based on demographic, behavioural and clinical risk factors from a nationally representative health survey dataset. The study aims to evaluate the predictive ability of the Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Extreme Gradient Boosting algorithms on several evaluation metrics and compare the results. In addition, the study seeks to uncover the most powerful predictors associated with cardiovascular disease risk and to provide evidence for early public health screening, preventive interventions and evidence-based cardiovascular disease risk management.

2. Materials and Methods

2.1 Study Design and Data Source

A cross-sectional analytical study with data from NHANES 2017-2018. The National Center for Health Statistics conducts NHANES to determine the health and nutritional status of the civilian, noninstitutionalized population of the United States. The analytical dataset comprised 5,568 participants with 19 variables including demographics, behaviours and clinical parameters. This was a

secondary analysis of publicly available and de-identified data, so no further ethical approval was required (Belbakhouche, 2026).

2.2 Study Population

Those who had complete data for heart disease status and all predictor variables selected were included. Records missing outcome data, incomplete demographic, behavioural or clinical data and duplicate data were excluded. 5,568 participants were included in the final analytical sample.

2.3 Study Variables

The outcome variable was heart disease status, coded as a binary measure indicating the presence or absence of heart disease. Demographic factors were age, gender, race/ethnicity, education, and poverty-income ratio. Predictors were smoking history, current smoking status and physical activity (behavioural). The clinical predictors included hypertension, diabetes, high cholesterol, stroke history, prediabetes, taking blood pressure medication, cholesterol-lowering medication, insulin, diabetes prescription and family history of heart attack.

2.4 Data Preprocessing

Data set was analysed for missing data, duplicate data and inconsistent data. Categorical variables were coded numerically and the continuous variables were left in their original format. Stratified sampling was used to partition data into 80% training and 20% testing sets. Standardization techniques were used for Logistic Regression and Support Vector Machine models. The class imbalance was solved by class weighting and only the Synthetic Minority Over-sampling Technique was applied to the training data.

2.5 Machine Learning Models

Four classifiers were developed: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost and LightGBM. Logistic Regression was used as the base model and tree-based and boosting algorithms were used to model nonlinear relationships and interactions among the predictors. For model optimization, cross-validation was used, for hyperparameters.

2.6 Model Evaluation

The accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve were used to evaluate the model performance. Confusion matrices were created to assess classification results. Five-fold stratified cross-validation was used to assess model stability. Due to the unbalanced outcome distribution, the model selection focused on more than accuracy—ROC-AUC, recall, and F1-score were the metrics used.

2.7 Statistical Analysis

The data for continuous variables were summarized in the form of means and standard deviations, while for categorical data, frequencies and percentages were used. The chi-square test was used to evaluate associations between categorical predictors and heart disease status. To examine the relationships among the predictors and for possible multicollinearity, correlation analysis was conducted. Random Forest and gradient boosting importance scores were calculated for the features, with SHAP values used to aid interpretation of the model. All analyses were done with Python and Pandas, NumPy, SciPy, Scikit-learn, XGBoost, LightGBM, imbalanced-learn, SHAP and Matplotlib.

3. Results

3.1 Participant Characteristics

A total of 5,510 participants were included in the final analytical sample, with 527 of them having heart disease and 4,983 not having heart disease. The ages of the participants with heart disease were significantly higher than those of the participants without heart disease (heart disease: 67.46 ± 11.79 years; no heart disease: 49.95 ± 17.50 years). Heart disease patients were 63.2% male, while the non-heart disease patients were 47.0% male. Participants suffering from heart disease also had less favourable behavioural and clinical profile. 62.6% of heart disease patients and 40.1% of healthy patients reported a history of smoking. Fifty-five percent (55.4%) of those who had heart disease and 68.2% of those who did not reported physical activity. The clinical risk factors were significantly higher in participants who had heart disease. 76.1% of those who had heart disease had hypertension; 42.1% had diabetes; 61.7% had high cholesterol; and 20.7% had previously had a stroke. The proportions of the same four groups among participants without heart disease were, respectively, 34.3%, 13.1%, 32.5%, and 3.3%. Other risk factors that were more common among participants with heart disease included: cholesterol medication, blood-pressure medication, insulin use, diabetes-pill use, and family history of heart attack. Demographic, behavioural, and clinical details are given in Table 1.

Table 1. Baseline Characteristics of Study Participants

Characteristic	Overall, N = 5,510	No heart disease, n = 4,983	Heart disease, n = 527
Age, years	51.62 ± 17.80	49.95 ± 17.50	67.46 ± 11.79
Poverty-income ratio	2.48 ± 1.49	2.51 ± 1.50	2.25 ± 1.35
Male sex	2,676 (48.6%)	2,343 (47.0%)	333 (63.2%)
Non-Hispanic White	1,911 (34.7%)	1,637 (32.9%)	274 (52.0%)
College graduate or above	1,298 (23.6%)	1,208 (24.2%)	90 (17.1%)
Ever smoked	2,329 (42.3%)	1,999 (40.1%)	330 (62.6%)
Current smoking	1,002 (18.2%)	909 (18.2%)	93 (17.6%)
Physically active	3,688 (66.9%)	3,396 (68.2%)	292 (55.4%)
Hypertension	2,111 (38.3%)	1,710 (34.3%)	401 (76.1%)
Diabetes	876 (15.9%)	654 (13.1%)	222 (42.1%)
High cholesterol	1,943 (35.3%)	1,618 (32.5%)	325 (61.7%)
Previous stroke	273 (5.0%)	164 (3.3%)	109 (20.7%)
Prediabetes	571 (10.4%)	507 (10.2%)	64 (12.1%)
Blood-pressure medication use	1,642 (29.8%)	1,293 (25.9%)	349 (66.2%)
Cholesterol medication use	1,198 (21.7%)	893 (17.9%)	305 (57.9%)
Insulin use	244 (4.4%)	155 (3.1%)	89 (16.9%)
Diabetes medication use	714 (13.0%)	560 (11.2%)	154 (29.2%)
Family history of heart attack	709 (12.9%)	557 (11.2%)	152 (28.8%)

3.2 Distribution of Heart Disease

In the total population of 5,510, 527 participants were diagnosed with heart disease, giving a prevalence of 9.56%. There were 4,983 participants (90.44% of the analytical sample) who did not have heart disease. As shown in Figure 1, the class imbalance was significant, with a large difference between the two categories, supporting the use of SMOTE in the training set.

Older, male, non-Hispanic White participants, those with fewer years of education, and those who had a

cardiovascular and metabolic risk factor were more likely to have heart disease. In particular, the heart disease group had significantly more people who were hypertensive, diabetic, hyperlipidaemic, had previously had a stroke and who had a family history of heart attack.

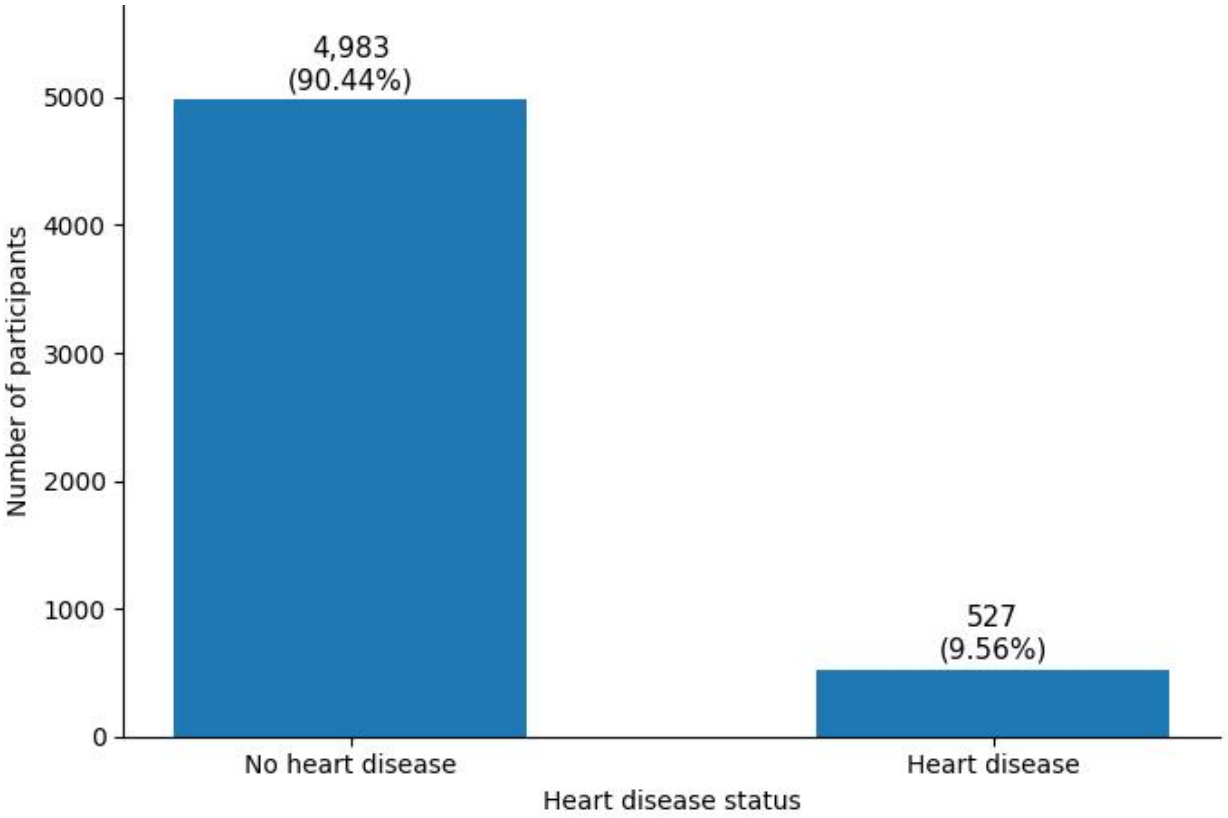

Figure 1. Distribution of Heart Disease Cases

3.3 Association Between Risk Factors and Heart Disease

Statistically significant relationships between heart disease and the majority of demographic, behavioural and clinical predictors were found in the bivariate analyses. The participants with heart disease were significantly older than the other participants ($t = -30.716$, Cohen's $d = 1.174$, $p < 0.001$). There was also a significant difference in poverty-income ratio between the participants with heart disease and those without, $t = 4.131$, Cohen's $d = -0.181$, $p < 0.001$, though this difference was small.

Sex, race/ethnicity and educational attainment were significant demographic factors associated with heart disease. Smoking history was also associated with the outcome, $\chi^2 = 97.974$, Cramér's $V = 0.133$, $p < 0.001$. Current smoking status was not however significantly associated with heart disease ($\chi^2 = 0.077$, $p = 0.782$). Physical activity was found to be significantly associated, and those with heart disease had lower levels of physical activity.

The highest category relationships were found for use of cholesterol meds, blood-pressure meds, hypertension, prior stroke and diabetes. Cholesterol medication use showed the largest effect size, $\chi^2 = 444.786$, Cramér's $V = 0.284$, $p < 0.001$. Next were the use of blood-pressure medicines, hypertension, previous stroke and diabetes. Unadjusted analysis did not show a significant relationship between prediabetes and heart disease ($\chi^2 = 1.784$, $p = 0.182$). Table 2 shows the association results.

Table 2. Association of Demographic, Behavioural, and Clinical Factors with Heart Disease

Risk factor	Statistical test	Test statistic	Effect size	p-value
Age, years	Welch t-test	−30.716	Cohen's d = 1.174	<0.001
Poverty-income ratio	Welch t-test	4.131	Cohen's d = −0.181	<0.001

Sex	Chi-square test	49.229	Cramér's V = 0.095	<0.001
Race/ethnicity	Chi-square test	91.006	Cramér's V = 0.129	<0.001
Education	Chi-square test	23.002	Cramér's V = 0.065	<0.001
Ever smoked	Chi-square test	97.974	Cramér's V = 0.133	<0.001
Current smoking	Chi-square test	0.077	Cramér's V = 0.004	0.782
Physically active	Chi-square test	34.398	Cramér's V = 0.079	<0.001
Hypertension	Chi-square test	350.147	Cramér's V = 0.252	<0.001
Diabetes	Chi-square test	297.618	Cramér's V = 0.232	<0.001
High cholesterol	Chi-square test	176.726	Cramér's V = 0.179	<0.001
Previous stroke	Chi-square test	302.446	Cramér's V = 0.234	<0.001
Prediabetes	Chi-square test	1.784	Cramér's V = 0.018	0.182
Blood-pressure medication	Chi-square test	367.632	Cramér's V = 0.258	<0.001
Cholesterol medication	Chi-square test	444.786	Cramér's V = 0.284	<0.001
Insulin use	Chi-square test	210.515	Cramér's V = 0.195	<0.001
Diabetes-pill use	Chi-square test	135.069	Cramér's V = 0.157	<0.001
Family history of heart attack	Chi-square test	131.070	Cramér's V = 0.154	<0.001

3.4 Machine Learning Model Performance

The testing and training subsets were the same and constituted five strata. Logistic Regression was the best model with ROC–AUC 0.876, accuracy 0.790, recall 0.829, precision 0.290 and F1-score 0.430. It was found that it performed stably with the cross-validated ROC–AUC of 0.865 ± 0.018. The highest accuracy of 0.896 and a ROC–AUC of 0.870 were achieved by XGBoost, which has a low recall of 0.276. Random Forest gave the maximum recall of 0.429 with a ROC–AUC of 0.864 whereas Support Vector Machine gave maximum recall of 0.552 with a ROC–AUC of 0.848. The lowest ROC–AUC was Decision Tree with 0.825 with comparatively high recall of 0.705. In the overall evaluation, Logistic Regression showed the best discrimination and heart disease case detection, reflecting that merely accuracy is not enough to evaluate in an imbalanced dataset, as shown in Table 3.

Table 3. Performance Comparison of Heart Disease Prediction Models

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC	CV ROC–AUC, mean	CV ROC–AUC, SD	CV F1-score, mean
Logistic Regression	0.790	0.290	0.829	0.430	0.876	0.865	0.018	0.412
XGBoost	0.896	0.426	0.276	0.335	0.870	0.851	0.015	0.352
Random Forest	0.884	0.398	0.429	0.413	0.864	0.855	0.012	0.381
Support Vector	0.848	0.326	0.552	0.410	0.848	0.818	0.017	0.391

Machine								
Decision Tree	0.781	0.261	0.705	0.380	0.825	0.792	0.043	0.357

3.5 ROC Curve Comparison

All five models had an ROC curve analysis that showed their performance was better than random classification. The largest area under the curve was obtained for Logistic Regression (ROC–AUC 0.876). The discriminatory performance of XGBoost and Random Forest was 0.870 and 0.864 (ROC–AUC), respectively. The Support Vector Machine had the highest ROC–AUC score (0.848) while the lowest ROC–AUC score was for the Decision Tree (0.825).

As illustrated in Figure 2, Logistic Regression curve was consistently in the upper left corner of the ROC space over a wide range of classification thresholds. This led to its choice as the most discriminative model to distinguish whether people had heart disease or not.

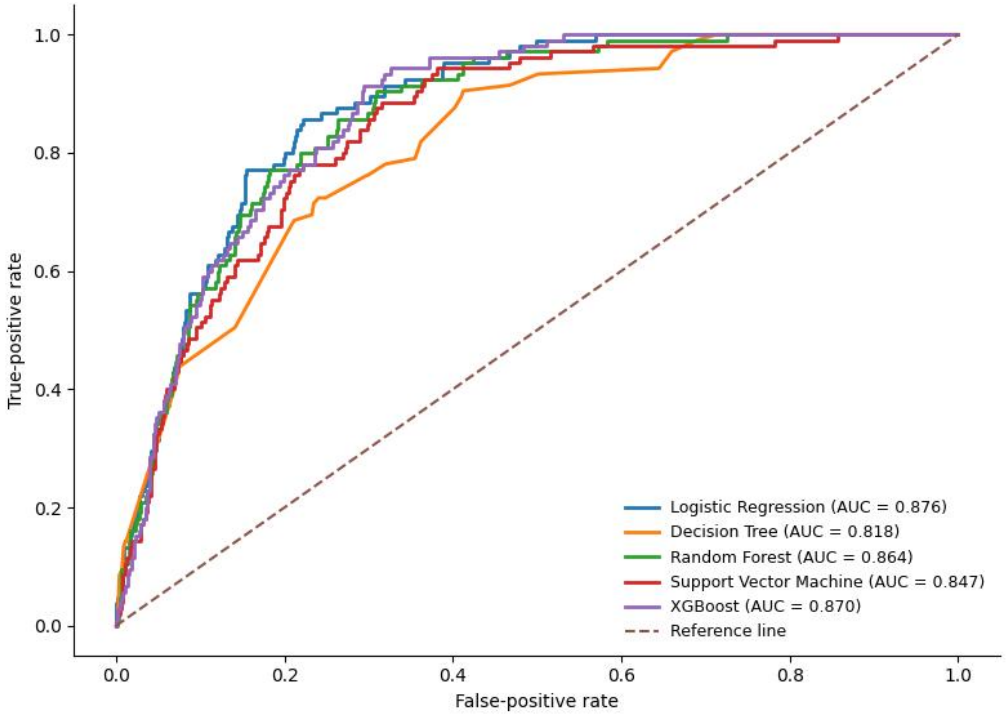

Figure 2. Receiver Operating Characteristic Curves of Machine Learning Models

Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and XGBoost receiver operating characteristic curves. Logistic Regression had the highest test ROC–AUC of 0.876 followed by XGBoost with 0.870 and Random Forest with 0.864.

3.6 Confusion Matrix of the Best-Performing Model

The Logistic Regression model correctly classified 784 (without heart disease) and 87 (with heart disease). This misclassified 213 people as positive who did not have heart disease and missed 18 people with heart disease. The model showed a recall of 0.829, which means that it recalled about 83% of the heart disease cases in the test sample. The relatively low number of undetected high-risk participants (false-negative predictions) is crucial in the public health screening context, as this implies that persons may not undergo timely preventive assessment. The 213 false positive predictions however, led to the relatively low precision value of 0.290. As such, this model may be more suitable for use as a screening or risk-stratifying tool, rather than as a stand-alone diagnostic system. The classification pattern is presented in Figure 3.

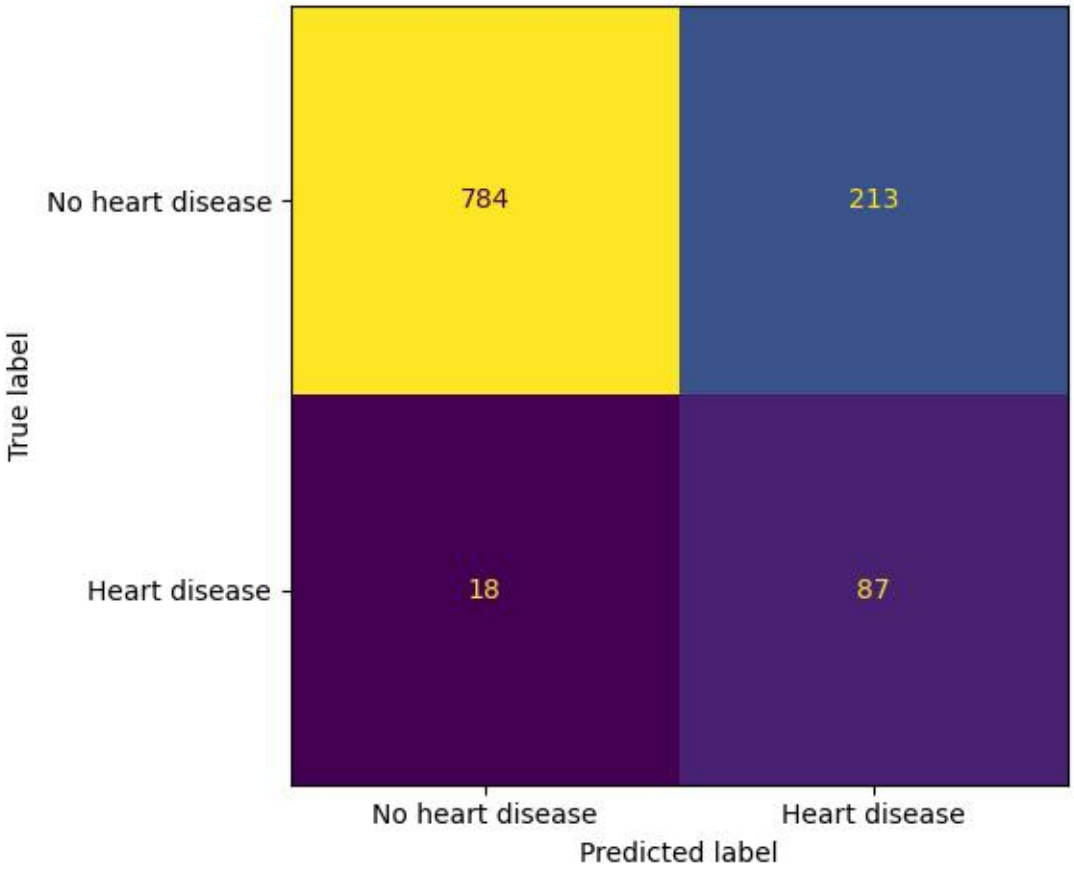

Figure 3. Confusion Matrix of the Best-Performing Logistic Regression Model

3.7 Feature Importance Analysis

The permutation importance analysis revealed that the most influential predictor of heart disease was age with a mean decrease in ROC–AUC of 0.0819 upon random permutation of its values. This contribution was significant, much more so than the other predictors.

The second most important factor was hypertension, followed by smoking history, cholesterol medication use, previous stroke, sex, family history of heart attack, diabetes, race/ethnicity, and diabetes medication use. In conclusion, both non-modifiable demographic factors, known clinical conditions, treatment indicators, hereditary susceptibility as well as behavioural exposures, were all involved in informing the model's predictions.

Age, hypertension, smoking history, use of cholesterol medications, previous stroke, diabetes, and family history were all highly correlated with the well-known cardiovascular risk areas. After all the other predictors were considered current smoking and physical activity had comparatively low permutation importance, indicating that the contribution of these predictors to the fitted model was comparatively less. The top 10 predictors are shown in Table 4 and plotted in Figure 4.

Table 4. Leading Predictors Based on Permutation Feature Importance

Rank	Predictor	Mean decrease in ROC–AUC
1	Age	0.0819
2	Hypertension	0.0162
3	Ever smoked	0.0155
4	Cholesterol medication use	0.0145
5	Previous stroke	0.0143
6	Sex	0.0136
7	Family history of heart attack	0.0131
8	Diabetes	0.0112

9	Race/ethnicity	0.0102
10	Diabetes medication use	0.0058

Feature importance ranking for Logistic Regression model with the highest performance using permutation. The dominant predictors were age followed by hypertension, smoking history, cholesterol medication use, previous stroke, sex, family history of heart attack, diabetes, race and ethnicity, and diabetes medication use.

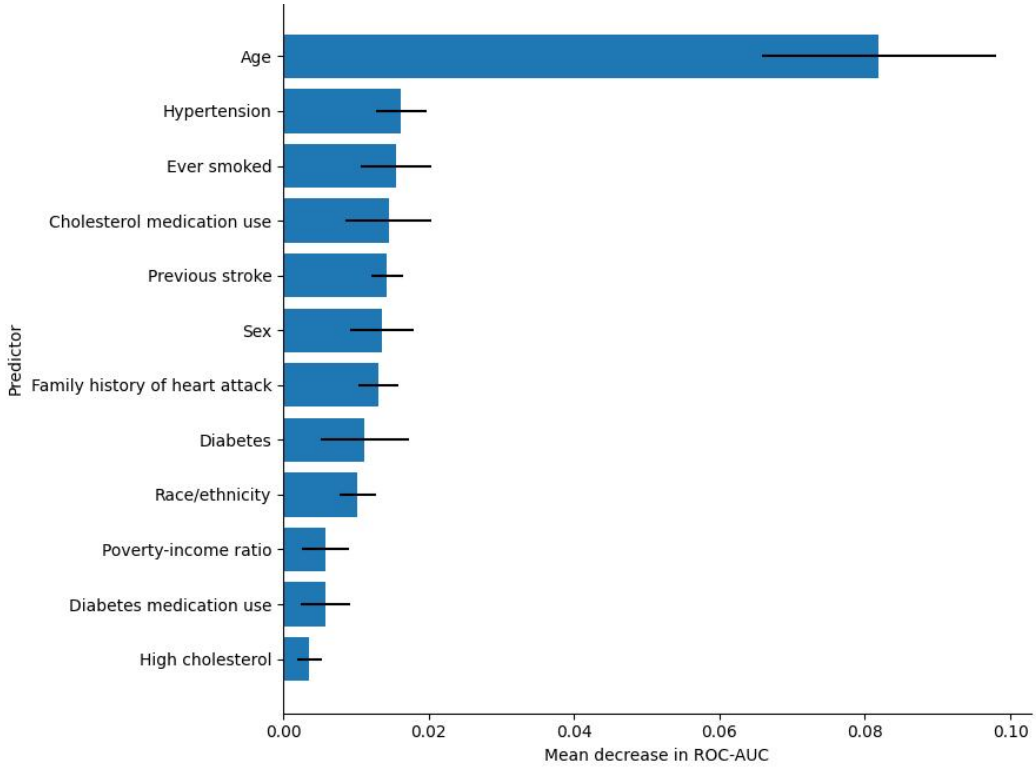

Figure 4. Feature Importance Ranking for Heart Disease Prediction

3.8 Public Health Interpretation of Prediction Results

The results identified a unique high-risk group including those who were older, had a history of previous stroke, hypertension, diabetes, high cholesterol, current smokers and those who used heart attack medication, along with a family history of heart attack. The average age of participants with heart disease was about 17.5 years older than the average age of participants without heart disease, and had a significantly higher prevalence of hypertension, diabetes, prior stroke and high cholesterol.

Logistic Regression was associated with a high recall, which may indicate that this model applied in population-level screening as a first step. The model used in community health assessments or national surveillance systems based on surveys to identify people who benefit from more detailed cardiovascular assessment. Moderate precision suggests that the positive predictions should be looked upon as pointers to clinical confirmation and not as diagnoses.

An added advantage of the model is that it can also be used to plan for preventive interventions through identification of modifiable risk domains. Relevant intervention opportunities include smoking cessation, blood-pressure control, diabetes management, cholesterol reduction, promotion of physical activity, and medication adherence. In the population context, demographic, behavioural and clinical data help public health professionals better prioritise prevention resources to those populations at highest expected cardiovascular risk.

4. Discussion

Five machine learning algorithms were created and compared to predict heart disease from demographic, behavioural and clinical variables from a nationally representative health survey in the present study. Logistic Regression among the evaluated models showed the best overall performance as it had high discriminatory ability (ROC – AUC = 0.876) and excellent sensitivity in identifying heart disease patients. The feature importance analysis also

showed that age, hypertension, smoking history, cholesterol medication use, previous stroke, sex, family history of heart attack, diabetes, race/ethnicity and diabetes medication use were the most important factors to the model predictions. The results suggest that, as always, cardiovascular risk factors remain relevant, and that machine learning methods that can be understood and interpreted can yield accurate and clinically useful risk stratification for population-based screening (Rajkomar et al., 2019; Topol, 2019).

Past studies have demonstrated that demographic factors significantly impact the estimated cardiovascular risk, yet can be altered by variation in healthcare access, socioeconomic factors and environmental exposures (Vasan & van den Heuvel, 2022). In the same vein, age and sex are strong and consistent predictors of future cardiovascular events across various populations and therefore are dominant predictors in the World Health Organization cardiovascular risk charts (Kaptoge et al., 2019).

However, behavioural characteristics also made a significant contribution to prediction performance. Among the strongest were smoking history, with physical activity having a relatively smaller independent contribution after accounting for other variables. These results indicate that exposure to these behaviours in the long term may have a larger effect on cumulative cardiovascular risk than current behaviour status alone. This is in line with earlier findings that smoking is known to cause endothelial damage, chronic inflammation and atherosclerosis, which leads to a higher risk of cardiovascular disease in the long term despite quitting. Therefore, behaviour change is still a critical element of cardiovascular prevention programs and population health initiatives that aim to decrease future burden of disease (Mensah et al., 2019; Roth et al., 2020).

Of the clinical factors, hypertension, diabetes, past stroke, hypercholesterolemia and use of cardiovascular drugs were repeatedly found to be important factors in determining the accuracy of the prediction. Moreover, the significance of medication-related factors simply be a reflection of the severity of underlying cardiometabolic disease itself, and not of the treatment effects. This has been noted in recent cardiovascular risk prediction trials where the use of traditional clinical risk factors and novel biomarkers significantly enhanced risk prediction over single risk factors alone (Martins et al., 2021). These findings further support the existing recommendations of using multiple interdependent clinical determinants in a comprehensive cardiovascular risk assessment (Tsao et al., 2022).

Logistic Regression provides clear estimation of the effects of the predictors, making it easier to interpret clinically and integrate into healthcare decision-making processes, unlike complex black-box models. Model transparency is also a key factor in fostering trust and responsible use of AI in healthcare systems, as highlighted in recent studies (Lundberg et al., 2020; Rajkomar et al., 2019).

Results of the present study are broadly similar to other population-based cardiovascular prediction studies. Age, hypertension, diabetes, smoking, and lipid abnormalities have long been found to have a dominant effect on cardiovascular outcomes in large epidemiological studies in a variety of populations (Tsao et al., 2022; Victor et al., 2024). Similarly, most recent studies that have used machine learning to predict cardiovascular disease have yielded better discrimination than standard prediction methods and argued for the need for nationally representative data to improve model generalizability (Patton, 2024). The observed performance of Logistic Regression is also consistent with previous studies suggesting that simpler, interpretable models can perform at a similar level in comparison to more complex algorithms in the presence of high-quality population health data (Rajkomar et al., 2019; Topol, 2019).

The current model is relevant to population-level cardiovascular prevention. The use of routinely collected demographic, behavioural and clinical information to identify individuals at high risk helps to enable earlier preventive interventions, targeted screening and more efficient allocation of healthcare resources. Given the model's high sensitivity, it is a useful tool in national health surveys and primary healthcare programmes as a first screening tool. These findings are consistent with current public health efforts to prevent cardiovascular disease around the world by proactively assessing and managing risk, and using data to guide healthcare planning (Roth et al., 2020; Victor et al., 2024). Future research should involve assessing longitudinal data, integrating with laboratory markers, imaging, and data captured by wearable devices as well as explainable artificial intelligence approaches to enhance prediction performance, clinical interpretability, and applicability to precision cardiovascular medicine.

5. Conclusion

Machine learning methods were shown to be effective at predicting heart disease in the population based on demographic, behavioural and clinical risk factors collected in a nationally representative health survey. Logistic Regression had the highest discriminatory ability with excellent sensitivity for heart disease among the five algorithms evaluated in this study and had the best overall predictive performance. The results also verified that the major factors that contributed to cardiovascular risk were age, history of hypertension, smoking, use of cholesterol medication, prior stroke, sex, family history of heart attack, diabetes, race/ethnicity and diabetes

medication use. These findings confirm the well-known fact that heart disease is a result of a combination of factors (demographic, lifestyle, and clinical conditions) and not just risk factors. The study also showed that interpretable machine learning models can be equally competitive in predicting outcomes and offer clear decision support to clinicians and public health officials. These models hold great promise for inclusion in population screening programmes, primary health care and preventive cardiovascular programs for early detection of high-risk individuals, with the expectation of timely intervention. Future studies should confirm these results with longitudinal and multi-centre data, further integrate laboratory biomarkers, imaging data, genomics and wearable-device data, and make use of explainable artificial intelligence techniques to further enhance the predictive performance, clinical understandability, and use in precision cardiovascular medicine. In general, the suggested structure is a feasible and research-informed way to improve cardiovascular risk assessment and facilitate data-driven public health decision making.

References

1. Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 128, 102289.
2. Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Muñoz, D., Smith, S. C., Virani, S. S., Williams, K. A., Yeboah, J., & Ziaeian, B. (2019). 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease. Journal of the American College of Cardiology, 74(10), e177–e232. https://doi.org/10.1016/j.jacc.2019.03.010
3. Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., & Munger, T. M. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70–74.
4. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. Jama, 319(13), 1317–1318.
5. Belbakhouche, A. (2026). Heart Disease Prediction. https://www.kaggle.com/datasets/akrambelbakhouche/heart-disease-prediction
6. Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., … On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation, 139(10). https://doi.org/10.1161/CIR.0000000000000659
7. Dutta, A., Batabyal, T., Basu, M., & Acton, S. (2019). An efficient convolutional neural network for coronary heart disease prediction. Available at SSRN 3514078. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3514078
8. Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M. T., Kavousi, M., Stevens, G., Riley, L. M., Savin, S., Khan, T., & Altay, S. (2019). World Health Organization cardiovascular disease risk charts: Revised models to estimate risk in 21 global regions. The Lancet Global Health, 7(10), e1332–e1345.
9. Khan, S. S., Greenland, P., Hayman, L. L., Khera, R., Navar, A. M., Pencina, M. J., Reza, N., Shah, S. H., Shanbhag, S., & Weber, B. (2026). Criteria to assess the predictive and clinical utility of novel models, biomarkers, and tools for risk of cardiovascular disease: A scientific statement from the American Heart Association. Circulation, 153(11), e953–e970.
10. Kremer, T. L., Grimaldi, D. A., Fleischer, H., Schwarz, E., Meyer-Lindenberg, A., Braun, U., & Tost, H. (2025). Intermediate endophenotypes and epigenetic mechanisms of psychiatric disorders. Imaging Neuroscience, 3, IMAG-a.
11. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., Min, J. K., Tang, W. W., Halperin, J. L., & Narayan, S. M. (2019). Deep learning for cardiovascular medicine: A practical primer. European Heart Journal, 40(25), 2058–2073.
12. Libby, P. (2021). The changing landscape of atherosclerosis. Nature, 592(7855), 524–533.
13. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.
14. Martins, A. M., Paiva, M. U., Paiva, D. V., de Oliveira, R. M., Machado, H. L., Alves, L. J., Picossi, C. R., Faccio, A. T., Tavares, M. F., & Barbas, C. (2021). Innovative approaches to assess intermediate cardiovascular risk subjects: A review from clinical to metabolomics strategies. Frontiers in Cardiovascular Medicine, 8, 788062.
15. Mensah, G. A., Roth, G. A., & Fuster, V. (2019). The Global Burden of Cardiovascular Diseases and Risk Factors. Journal of the American College of Cardiology, 74(20), 2529–2532. https://doi.org/10.1016/j.jacc.2019.10.009
16. Patton, M. J. (2024). Biomedical Informatic Analysis of Critical Care Outcomes & Chronic Sequelae of SARS-CoV-2 Infection. The University of Alabama at

Birmingham. https://search.proquest.com/openview/db53d2191be65c9c42cfbfef270b8426/1?pq-origsite=gscholar&cbl=18750&diss=y

17. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
18. Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A., Benjamin, E. J., & Benziger, C. P. (2020). Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 study. Journal of the American College of Cardiology, 76(25), 2982–3021.
19. Sritharan, H. P., Nguyen, H., Allahwala, U. K., & Bhindi, R. (2026). Machine learning for cardiovascular risk prediction: A practical primer for clinicians. Heart, Lung and Circulation. https://www.sciencedirect.com/science/article/pii/S1443950626000430
20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
21. Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., Boehme, A. K., Buxton, A. E., Carson, A. P., Commodore-Mensah, Y., Elkind, M. S. V., Evenson, K. R., Eze-Nliam, C., Ferguson, J. F., Generoso, G., Ho, J. E., Kalani, R., Khan, S. S., Kissela, B. M., … on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2022). Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. Circulation, 145(8). https://doi.org/10.1161/CIR.0000000000001052
22. Vasan, R. S., & van den Heuvel, E. (2022). Differences in estimates for 10-year risk of cardiovascular disease in Black versus White individuals with identical risk factor profiles using pooled cohort equations: An in silico cohort study. The Lancet Digital Health, 4(1), e55–e63.
23. Velarde, G., Bravo-Jaimes, K., Brandt, E. J., Wang, D., Douglass, P., Castellanos, L. R., Rodriguez, F., Palaniappan, L., Ibebuogu, U., Bond, R., Ferdinand, K., Lundberg, G., Thamman, R., Vijayaraghavan, K., & Watson, K. (2023). Locking the Revolving Door: Racial Disparities in Cardiovascular Disease. Journal of the American Heart Association, 12(8), e025271. https://doi.org/10.1161/JAHA.122.025271
24. Victor, G., Shishani, K., Vellone, E., & Froelicher, E. S. (2024). The global burden of cardiovascular disease in adults: A mapping review. Journal of Cardiovascular Nursing, 10–1097.
25. Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Khan, S. S., Kissela, B. M., Knutson, K. L., Kwan, T. W., Lackland, D. T., … On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2020). Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. Circulation, 141(9). https://doi.org/10.1161/CIR.0000000000000757
26. Yu, D., Yang, S., Wang, R., Wang, K., Han, W., Wu, H., Wang, W., & Wang, X. (2023). Machine Learning in Heart Failure Research: A Bibliometric Analysis from 2003 to 2023. https://assets-eu.researchsquare.com/files/rs-3024872/v1_covered_1f031e61-af43-42b0-a4de-c971d7c1f6e2.pdf