**Introduction**

Cardiovascular disease (CVD) is a leading global cause of mortality, posing significant public health burdens across all socioeconomic strata. In the UK, approximately 7.6 million individuals live with CVD, contributing to over 174,000 deaths, representing 27.1% of all fatalities in 2022. A critical manifestation of CVD is coronary heart disease (CHD), characterised by luminal narrowing or obstruction of coronary arteries, primarily due to atherosclerosis. This can result in angina pectoris, myocardial infarction, and sudden arrhythmic death (1).

CHD arises from a complex interaction of demographic, metabolic, and physiological factors, including age, sex, blood pressure, lipid levels, glycaemic status, cardiac electrical function, exercise tolerance, and anginal features, all of which inform clinical risk stratification(2). Accurate detection of CHD is essential for timely intervention, yet standard diagnostics such as electrocardiography, angiography, and biochemical testing often lack sensitivity to early or subclinical disease. Symptoms including chest pain, dyspnoea, and palpitations typically reflect advanced pathology and may be atypical or absent in older adults and women(3).

The digitisation of healthcare records facilitates data-driven risk prediction using machine learning (ML). Supervised classification algorithms show promise in discerning complex, non-linear patterns within routine clinical data, potentially enhancing traditional risk assessment by identifying subtle associations. Given CHD's often asymptomatic progression, particularly in resource-constrained settings, improved prediction models can prioritise preventative measures, optimise resource allocation, and ultimately improve survival through earlier identification of at-risk individuals(3,4).

This study aims to develop and evaluate a ML Random Forest(RF) model using the StatLog Heart dataset encompassing demographic and clinical variables. The objectives are to predict CHD presence, identify key contributing factors, and thoroughly assess model performance. The findings may inform clinical decision-making and support timely, data-driven interventions for CHD.

**Methods:**

This study utilised the StatLog Heart dataset, comprising 270 patient records, to develop a supervised classification model for CHD detection. The dataset includes thirteen predictors spanning demographic characteristics, clinical symptoms, vital signs, laboratory measures, and diagnostic findings. The outcome variable was a binary indicator representing CHD, categorised as “Present” or “Absent”.

Random Forest (RF) was selected as the primary classification method due to its ability to capture complex, non-linear relationships without restrictive parametric assumptions. As an ensemble algorithm, it combines bootstrap aggregation with random feature selection at each node, reducing variance and mitigating overfitting, advantages particularly relevant given the dataset’s modest sample size (5). RF is also resilient to multicollinearity, robust to outliers, and insensitive to variable scaling, making it well-suited to heterogeneous clinical data. Furthermore, it provides intrinsic measures of variable importance and supports cost-sensitive learning, aligning with the clinical priority of minimising false negatives (6,7). Alternative methods such as logistic regression, support vector machines (SVMs), and gradient boosting were considered but not selected due to limitations in flexibility, interpretability, or risk of overfitting. Random Forest offers a favourable balance of predictive accuracy, generalisability, and interpretability (8).

Model development adhered to TRIPOD guidelines (9). The dataset was randomly partitioned into training (70%) and testing (30%) subsets, preserving class proportions. A Random Forest model was trained using the *randomForest*(10) package in R (v4.4.1)(11). Internal validation was conducted using five-fold cross-validation, with performance evaluated using accuracy, sensitivity, specificity, F1 score, and area under the receiver operating characteristic (ROC) curve. The model's convergence and stability were assessed using out-of-bag (OOB) error estimates, to provide an unbiased measure of prediction error during training. AUC was reported with 95% confidence intervals. Final model performance was assessed on the independent test set, and variable importance was quantified using the Mean Decrease in Gini index, with results presented using clinically interpretable labels(12).

**Results:**

Figure 1 illustrates the distribution of the outcome variable in the StatLog Heart dataset, with 150(55.6%) observations classified as 'Absent' and 120 (44.4%) as 'Present'. This indicates that there is a slight imbalance between the two outcome categories.

**Figure 1**

*Distribution of the Presence of Heart Disease*

A graph of a heart disease

AI-generated content may be incorrect.

Figure 2 shows the distribution of categorical clinical features by CHD status. Asymptomatic chest pain, exercise-induced angina, flat ST slopes, and reversible thalassemia defects were more common in individuals with heart disease. In contrast, upsloping ST segments and normal thalassemia readings were more prevalent among those without disease. Most individuals in both groups had fasting blood sugar ≤ 120 mg/dl. A higher number of major vessels (1–3) and left ventricular hypertrophy were more frequently observed in the disease group, while normal ECG readings and 0 vessels predominated in those without disease. Males were more common in both groups, with a slight predominance among those with heart disease.

**Figure 2**

*Exploration of Clinical Factors Associated with Heart Disease*

A chart of different types of heart disease

AI-generated content may be incorrect.

Figure 3 presents boxplots of continuous clinical variables grouped by heart disease status. Individuals with heart disease tended to be older, with a median age of 58 years compared to 52 years in those without disease. Cholesterol levels were slightly higher in the disease group, with median values of 250 mg/dL versus 235 mg/dL in the non-disease group. Maximum heart rate was lower among individuals with heart disease, with a median of 140 beats per minute compared to 165 in those without. Resting blood pressure differed modestly, with medians of 135 mmHg and 130 mmHg in the disease and non-disease groups, respectively. ST depression showed the most distinct separation, with a median of 1.5 mm in the disease group and 0.5 mm in those without.

**Figure 3**

*Distribution of Clinical Factors Associated with Heart Disease*

A screenshot of a graph

AI-generated content may be incorrect.

Figure 4 displays Pearson correlation coefficients, where values closer to 1 indicate stronger positive associations with CHD. The strongest correlations were observed for thalassemia type(r = 0.53), number of major vessels(r = 0.46), chest pain type(r = 0.42), ST depression(r = 0.42), and exercise-induced angina (r = 0.42). Notably, maximum heart rate was negatively associated with disease(r = –0.42).

**Figure 4**

*Correlation Matrix of Clinical Measures and Heart Disease Status*

A diagram of a matrix

AI-generated content may be incorrect.

Table 1 demonstrates Internal validation results using five-fold cross-validation. Average accuracy was 85.2%, with sensitivity of 76.8% and specificity of 91.6%. The F1 score was 0.82, reflecting balanced detection of true cases while minimising false positives. Discriminative ability was high, with an AUC of 0.897 (95% CI: 0.85–0.94).

**Table 1**

*Internal Validation Performance (5-Fold Cross-Validation)*

|  |  |
| --- | --- |
| **Metric** | **Estimate (95%CI)** |
| Accuracy | 0.852 |
| Sensitivity | 0.768 |
| Specificity | 0.916 |
| F1 Score | 0.818 |
| AUC | 0.897 (0.850 – 0.944) |

Figure 5 shows the model's misclassification rates over 500 trees. The overall out-of-bag (OOB) error stabilises around 16%, indicating consistent predictive performance. Class-specific errors differ, with lower error for patients without heart disease and higher variability for those with disease. The curves flatten after approximately 100 trees, suggesting model stability.

**Figure 5**

*Out-of-Bag Misclassification Rates Across 500 Trees in the Random Forest Model*

A graph of different colored lines

AI-generated content may be incorrect.

Test set performance metrics are presented in Table 2. The RF model showed strong predictive ability, with an accuracy of 80.2% and an AUC of 0.863(95% CI: 0.786–0.931), as illustrated in Figure 6. Sensitivity and specificity were 71.1% and 88.4%, respectively. The F1 score was 0.771, with positive and negative predictive values of 84.9% and 77.8%.

**Table 2**

*Test Set Performance Metrics*

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **95% CI** |
| Accuracy | 0.802 | 0.716 – 0.889 |
| Sensitivity | 0.711 | 0.568 – 0.850 |
| Specificity | 0.884 | 0.784 – 0.974 |
| F1 Score | 0.771 | 0.654 – 0.872 |
| AUC | 0.863 | 0.786 – 0.931 |

**Figure 6**

*Receiver Operating Characteristic Curve*

A graph showing a curve

AI-generated content may be incorrect.

Figure 7 shows the ranked importance of predictors based on the Mean Decrease in Gini index. Maximum heart rate, number of major vessels, and thalassemia type were the most influential features, followed by ST depression and chest pain characteristics. Variables such as fasting blood sugar, resting ECG, and sex contributed comparatively less to the model’s classification performance.

**Figure 7**

*Clinical predictors ranked by importance in the Random Forest model*

A graph of a number of patients with heart disease

AI-generated content may be incorrect.

**Discussion**

CHD arises from multifactorial interactions across clinical, genetic, demographic, and behavioural domains(2). This has driven increasing interest in machine learning (ML) methods capable of integrating heterogeneous clinical data to support more personalised risk prediction(13). In this study, a RF classification model was developed and validated to predict CHD using routinely collected clinical variables.

The Random Forest model demonstrated strong and consistent discriminative performance between the two classes, with an AUC of 0.897 under five-fold cross-validation and 0.863 on the independent test set, indicating strong generalisability. The slight discrepancy between cross-validation and test set metrics may suggest modest overfitting to the training data, a known risk in high-variance models like Random Forests. Test set sensitivity was 71.1% and specificity 88.4%, reflecting balanced classification with a tendency to favour disease-free predictions, consistent with out-of-bag (OOB) error estimates showing lower misclassification in the majority class.

The model’s high specificity suggests strong discriminatory capacity in identifying non-disease cases, which could reduce unnecessary investigations and associated healthcare costs. However, its lower sensitivity constrains standalone use in diagnostic pathways, particularly in early-stage or asymptomatic presentations. This trade-off likely reflects class imbalance, which can bias ensemble models like RF towards the majority class. To address this, the F1 score was adopted as a central performance metric. Unlike accuracy, which can be misleading in imbalanced datasets, the F1 score balances sensitivity and precision, providing a more reliable estimate of true case detection. In this study, the F1 score was 0.77, supporting the model’s adequacy under imbalance. Nonetheless, the modest sensitivity remains a clinical concern in CHD, where missed diagnoses may delay timely intervention and negatively impact outcomes(13,14).

Comparable studies employing Random Forest classifiers for cardiovascular disease prediction have reported marginally higher performance metrics. This is due to their use of larger sample sizes, expanded predictor sets, data resampling strategies for class imbalance, or greater proportions of data allocated for training(13,15, 16).

A notable strength of the Random Forest model is its capacity to rank predictors by importance(12). Maximum heart rate, number of major vessels, and ST depression emerged as the most influential features, consistent with established markers of myocardial ischemia and anatomical disease burden. Thalassemia type was also highly ranked, while not a direct causal factor, it may act as a proxy for underlying cardiovascular pathology or reflect dataset-specific patterns. Conversely, fasting blood sugar contributed minimally to prediction despite being a known risk factor. This likely reflects its binary encoding, which limits predictive granularity, and might also reflect potential treatment effects within the cohort. These results underscore the importance of contextual clinical interpretation when assessing model-derived feature rankings(17).

Previous studies have consistently shown that Random Forest outperforms other heart disease classifiers, including logistic regression, decision trees, and ensemble methods, across key metrics(13,15,16,18). One study reported RF achieved the highest precision (96.2%), perfect sensitivity (100%), an AUC of 0.989, and an F1 score of 97.7%, reflecting a strong balance between sensitivity and specificity. This reflects the promising potential of RFs as classification models(18).

A key limitation of this study is the relatively small sample, which may have restricted the Random Forest's capacity to fully leverage its ensemble learning advantages and led to overfitting, indicated by the slightly lower performance metrics on the test set. Moreover, class imbalance was not formally addressed, potentially biasing the model toward the majority class, leading to reduced sensitivity. Future work should incorporate class-balancing strategies such as the Synthetic Minority Over-sampling Technique (SMOTE) to enhance minority class representation and improve the model’s ability to detect true positive cases. External validation using independent, real-world datasets will also be necessary to assess generalisability and ensure reliable performance across heterogeneous clinical populations(14).

This study demonstrated that a Random Forest classifier can accurately CHD using routinely collected clinical features, with strong discriminatory performance and internal validity. While limitations such as class imbalance and absence of external validation may constrain generalisability, the model shows clinical promise. Future research should prioritise interpretability, enhanced data preprocessing, feature expansion, and external validation to support real-world implementation.

**REFERENCES**

1. Cheema KM, Dicks E, Pearson J, Samani NJ. Long-term trends in the epidemiology of cardiovascular diseases in the UK: insights from the British Heart Foundation statistical compendium. Cardiovasc Res. 2022 Jul 27;118(10):2267–80.
2. Waterall J. Health matters: preventing cardiovascular disease. Public Health Matters Public Health England. 2019.
3. Almansouri NE, Awe M, Rajavelu S, Jahnavi K, Shastry R, Hasan A, et al. Early Diagnosis of Cardiovascular Diseases in the Era of Artificial Intelligence: An In-Depth Review. Cureus [Internet]. 2024 Mar 9 [cited 2025 May 11]; Available from: <https://www.cureus.com/articles/236353-early-diagnosis-of-cardiovascular-diseases-in-the-era-of-artificial-intelligence-an-in-depth-review>
4. Liu T, Krentz AJ, Huo Z, Ćurčin V. Opportunities and Challenges of Cardiovascular Disease Risk Prediction for Primary Prevention Using Machine Learning and Electronic Health Records: A Systematic Review. Reviews in Cardiovascular Medicine. 2025 Apr 25;26(4).
5. Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst Appl. 2024 Mar;237:121549.
6. Petkovic D, Altman R, Wong M, Vigil A. Improving the explainability of Random Forest classifier – user centered approach. Biocomputing 2018. 2017 Nov 17;
7. Atefeh DARAEI, HAMIDI H. An Efficient Predictive Model for Myocardial Infarction Using Cost-sensitive J48 Model. Iranian Journal of Public Health [Internet]. 2017 May [cited 2025 May 11];46(5):682. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5442282/>
8. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019 Dec;19(1):281.
9. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;e078378.
10. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18–22. Available from: <https://CRAN.R-project.org/doc/Rnews/>
11. R Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2024. Available from: <https://www.R-project.org/>
12. Bustamante A, Giralt D, García-Berrocoso T, Rubiera M, Álvarez-Sabín J, Molina C, et al. The impact of post-stroke complications on in-hospital mortality depends on stroke severity. European Stroke Journal. 2016 Nov 28;2(1):54–63.
13. Sumwiza K, Twizere C, Rushingabigwi G, Bakunzibake P, Bamurigire P. Enhanced cardiovascular disease prediction model using random forest algorithm. Inform Med Unlocked. 2023;41:101316.
14. Gurcan F, Soylu A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. Cancers. 2024 Oct 8;16(19):3417.
15. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports. 2020 Mar 23;10(1).
16. Tao X. A random forest-based prediction of cardiovascular diseases. Applied and Computational Engineering [Internet]. 2024 Jul 16 ;76(1):311–6. Available from: <https://www.ewadirect.com/proceedings/ace/article/view/13779>
17. Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S. Significance of machine learning in healthcare: Features, pillars and applications. Int J Intell Netw. 2022;3:58–73.
18. Hossain S, Hasan MK, Faruk MO, Aktar N, Hossain R, Hossain K. Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023. BMC Cardiovasc Disord. 2024 Apr 18;24(1):214.