

Heart Disease Prediction Analysis: An Explainable Approach

Kartik Waghmode¹, Suhas K C², S Sathyanarayanan^{*3}

Department of Mathematics and Computer Science
Sri Sathya Sai University for Human Excellence,
Navanihal, Okali Post, Kamalapur
Kalaburagi, Karnataka - 585313

* Correspondence author: S. Sathyanarayanan³

E-mail: sathyanarayanan.s@sssuhe.ac.in / sathyanarayanan.brn@gmail.com

Abstract—Heart and blood vessel disease (CVD) is one of the top preventable causes of death throughout the world. This paper introduces a methodologically sound and clinically interpretable machine learning model to predict the presence of ischaemic heart disease (IHD) using the Heart Failure Prediction Dataset (fedesoriano, 2021), which comprises 918 unique patient records from 5 different datasets that are validated globally. Four supervised learning models—Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost—were compared in a scikit-learn pipeline, which ensures consistent data leakage prevention by encapsulating all the data-preprocessing operations and performing 5-fold stratified cross-validation. Optimisation of hyperparameters was done using GridSearchCV, while testing for statistical significance of inter-model differences were done using McNemar’s test. The tuned SVM model using RBF kernel gave the highest held-out performance: 90.76% accuracy, 92.11% F1-score, and 95.34% AUC value. SHAP analysis with the pairwise interaction quantification (SHAP-IQ) revealed that the most important clinically meaningful predictors were ST_Slope, ChestPainType, Oldpeak, and MaxHR, each with more than 75% chance of dominance—entirely consistent with what is known in the field of cardiovascular medicine. Model performance was tested according to subgroups based on age and sex to ensure fairness. This study proves that predictive performance, methodological rigor, statistical validation, explainability, and awareness of fairness can be realized simultaneously in a clinically small ML framework, offering a scalable basis for cardiovascular decision support in resource-constrained settings.

Index Terms—Ischaemic heart disease, cardiovascular disease prediction, cross-validation, XAI, SHAP-IQ, machine learning, clinical decision support, Support Vector Machine, Random Forest, XGBoost, Logistic Regression, pipeline-based pre-processing, hyperparameter optimisation, McNemar’s test, fairness audit.

I. INTRODUCTION

Cardiovascular disease (CVD) is the world’s top killer, with the World Health Organisation estimating that it causes approximately 17.9 million deaths every year [14]. Of these subtypes of CVD, ischaemic heart disease (IHD) and stroke account for a larger proportion of premature deaths and DALYs lost than other more uncommon forms of CVD worldwide. The burden of this disease in India is particularly significant: the death rate due to CVD is 272 per 1,00,000 people, much higher than the global average of 235 per 1,00,000 [1]. WHO figures show that 63% of all deaths in India were due to non-communicable diseases (NCDs) in 2016, with cardiovascular diseases as the major single cause, accounting for 27% of

the total mortality. In the age group of 40–69 years—the economically productive age group—an estimated 45% of deaths are due to CVDs [2].

The cost of heart and blood vessel disease extends beyond mortality. Many patients face recurrent hospital admissions, long-term disability, lower quality of life, and declining cardiac function when heart disease remains undiagnosed or untreated. The financial burden, both in direct health care spending and lost productivity, is substantial for individual families and healthcare systems alike. This scenario is exacerbated by inadequate specialist cardiology reach in low and middle income countries (LMICs), including India, with fewer trained cardiologists working outside large cities.

Over the last 10 years, rapid development of machine learning (ML) and artificial intelligence (AI) has created opportunities in disease detection and clinical decision support [15], [16]. Unlike conventional rule-based clinical scoring frameworks such as the Framingham Risk Score, ML models are equipped to simultaneously examine large numbers of patient properties, locate complex non-linear connections, and create customised risk scores. Despite advances in predictive power, clinical-level deployment of ML models for cardiovascular disease risk remains limited, primarily due to a lack of model interpretability. This has led to a rise in Explainable Artificial Intelligence (XAI), a field focused on creating techniques to make model predictions meaningful to human decision makers [19].

This paper simultaneously addresses the challenges of predictive accuracy, interpretability, and statistical rigor. We leverage Kaggle’s Heart Failure Prediction Dataset to create, tune using GridSearchCV, test, and contrast four ML models in a solid pipeline with 5-fold stratified cross-validation. We use SHAP-IQ analysis to explain feature importance and interactions for the top-performing model, and conduct McNemar’s testing to detect statistical relevance of differences between models, along with a fairness audit across demographic subgroups. The specific objectives are to:

- 1) Deliver competitive and statistically sound predictive performance;
- 2) Avoid data leakage using strictly pipeline-based pre-processing;
- 3) Leverage SHAP-IQ to generate explanations including pairwise interaction effects;

- 4) Identify the most clinically relevant risk features; and
- 5) Ensure model performance is fair across patient sub-populations.

II. LITERATURE REVIEW

Approaches to heart disease prediction using machine learning date back more than 20 years, evolving from basic statistical classifiers to complex ensemble classifiers and advanced deep learning frameworks. This section reviews selected related literature emphasising methods used and research gaps that inform the present study.

Paruchuri (2025) [3] used the UCI Heart Disease dataset to compare ML models, finding that Random Forest and Gradient Boosting captured data patterns most effectively. The study highlighted that preprocessing quality and model interpretability are essential in healthcare, noting the well-known accuracy–interpretability trade-off that hinders clinical adoption of black-box models.

A large-scale evaluation by Bhowmik et al. (2024) [4] across approximately 70,000 patient records achieved a best accuracy of 71.76% for Logistic Regression and ROC-AUC of 0.778. Significant challenges identified included data heterogeneity, class imbalance, and difficulties in generalising models across demographic groups.

Nasution et al. (2025) [5] assessed Logistic Regression, Random Forest, KNN and SVM using chi-square and mutual information feature selection, with Random Forest achieving the best accuracy of 89.7%. Tousif et al. (2024) [6] achieved 93% accuracy on the UCI Cleveland dataset using KNN, though the study was limited to 303 cases and lacked cross-validation and pipeline-based preprocessing. Shrestha (2024) [8] performed XGBoost and LSTM on the Cleveland dataset, obtaining 90% accuracy and AUC of 0.94. Ufumaka (2020) [9] demonstrated that SVM could reach 83.807% accuracy with 86.166% F1-score using rigorous 5-fold and 10-fold cross-validation.

Several recurring gaps characterise prior research: (1) restriction to small UCI Cleveland datasets with limited statistical power; (2) absence of pipeline-based preprocessing leading to potential data leakage; (3) feature importance analysis limited to single-feature SHAP, without pairwise interaction analysis; (4) rarely reported principled hyperparameter optimisation; (5) absence of statistical tests for model comparison; and (6) near-total neglect of fairness evaluations across demographic subgroups. The present study directly addresses all six of these gaps.

III. DATA DESCRIPTION

A. Dataset Source and Provenance

In September 2021, the “Heart Failure Prediction Dataset” (HFPD) was published on Kaggle by fedesoriano [7] and is available for non-commercial research. The dataset integrates features from five independently validated heart disease datasets obtained at different medical institutions worldwide: Cleveland Clinic Foundation (303 observations), Hungarian Institute of Cardiology, Budapest (294 observations), University Hospital Zurich, Switzerland (123 observations), Long

Beach VA Medical Center (200 observations), and the Statlog Heart Dataset (270 observations). This multi-source aggregation results in a more demographically diverse and statistically meaningful dataset than any single source.

A major data quality issue is the presence of zero cholesterol values—physiologically impossible and representing missing data—in 172 out of 918 records (18.7%). Rather than discarding these records or applying naïve global mean imputation, we applied stratum-specific median imputation based on HeartDisease class label and sex, retaining demographic structure while replacing physiologically implausible values with clinically plausible estimates. This methodological advancement differentiates the present study from prior work that neglects or naively handles this issue.

B. Dataset Composition and Class Distribution

From 1,190 original raw observations, 272 duplicate records were removed, yielding a clean dataset of 918 unique patient records. The class distribution of the target variable *HeartDisease* is given in Table I.

TABLE I: Target Variable Class Distribution

Class Label	Count	Proportion	Description
Heart Disease Absent (0)	410	44.7%	No diagnosis
Heart Disease Present (1)	508	55.3%	Heart disease
Total	918	100%	From 5 datasets

This class distribution is moderately imbalanced (55.3% class 1). Although this slight imbalance does not require resampling, macro-averaged metrics are reported alongside class-specific metrics to prevent any single class from dominating the reported performance. The near-balanced distribution ensures that the achieved 90.76% accuracy is significantly better than the naïve majority-class baseline of 55%.

C. Feature Description

All 11 input features are detailed in Table II. The clinical significance of each feature for cardiovascular health assessment is well established, enabling the creation of clinically pertinent and interpretable predictive models.

D. Exploratory Data Analysis

Exploratory data analysis (EDA) was performed before modelling. The main results of the correlation analysis (Fig. 1) are: Oldpeak exhibits the highest positive correlation coefficient ($r = 0.40$) with HeartDisease, indicating that increased ST depression during exercise is strongly correlated with disease presence. MaxHR shows the highest negative correlation ($r = -0.40$), meaning patients with higher exercise heart rate are less likely to have heart disease. Moderate positive correlations were observed for Age ($r = 0.28$) and FastingBS ($r = 0.27$). Inter-feature correlations are generally low, confirming minimal multicollinearity.

From distribution analysis (Fig. 2), we observe that ST_Slope and ChestPainType exhibit the best class separation among the categorical features, foreshadowing their dominance in the SHAP analysis reported in Section IV-E.

TABLE II: Dataset Feature Description

Feature	Clinical Description	Type	Values / Category
ChestPainType	Indicates the nature of chest pain. Asymptomatic (ASY) is most dangerous: patients may have severe disease with no apparent symptoms, leading to delayed diagnosis.	Categorical	TA, ATA, NAP, ASY
Age	Age in years. Older patients carry higher CVD risk due to arterial stiffness and cumulative risk factor exposure over time.	Numeric	Years
RestingBP	Resting blood pressure (mmHg). Elevated values denote hypertension, a major modifiable CVD risk factor. Zero values treated as missing.	Numeric	mmHg
Cholesterol	Serum cholesterol (mg/dL). Contributes to arterial plaque formation (atherosclerosis). Zero values imputed via stratified median.	Numeric	mg/dL
FastingBS	Fasting blood sugar >120 mg/dL. Suggests diabetes or pre-diabetes, both independent CVD risk factors.	Categorical	1=True, 0=False
RestingECG	Electrical cardiac activity at rest. ST denotes ST-T wave abnormality; LVH indicates left ventricular hypertrophy.	Categorical	Normal, ST, LVH
MaxHR	Maximum heart rate achieved during exercise. Lower values indicate reduced cardiac reserve and elevated CVD risk.	Numeric	bpm
ExerciseAngina	Whether chest pain occurred during exercise, strongly indicating myocardial ischaemia and reduced perfusion.	Categorical	Y=Yes, N=No
Oldpeak	ST depression magnitude during exercise relative to rest. Higher values indicate greater ischaemia and elevated risk.	Numeric	mm
ST_Slope	Slope of peak exercise ST segment. Flat/downsloping patterns indicate ischaemia; upsloping is associated with normal response.	Categorical	Up, Flat, Down
HeartDisease	Target variable: presence or absence of heart disease based on clinical and angiographic diagnostic records.	Binary	1=Present, 0=Absent

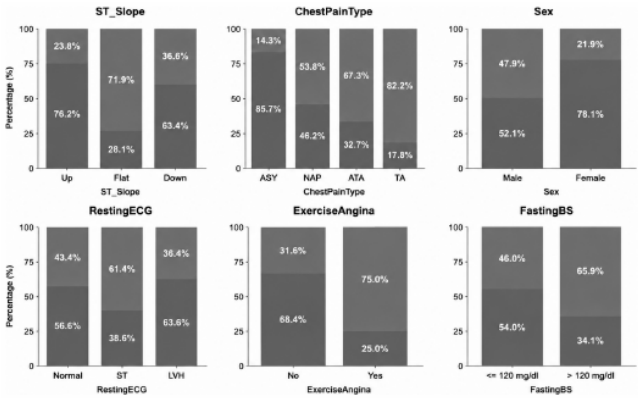
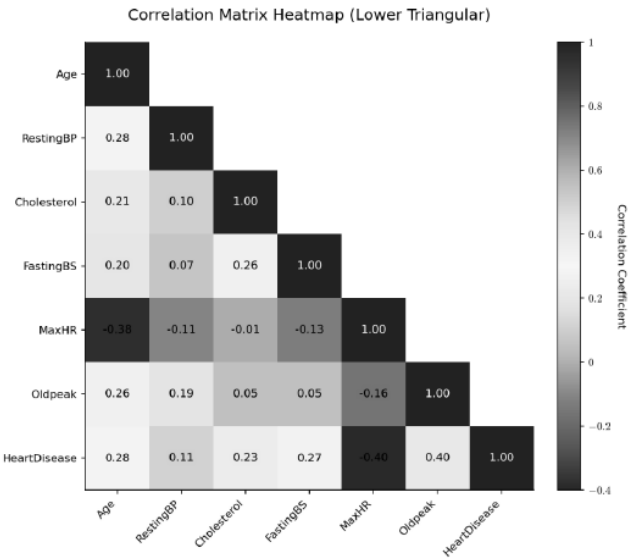


Fig. 2: Class-stratified distribution of key categorical features. ST_Slope and ChestPainType exhibit the strongest visual class separation, consistent with SHAP feature attribution results.

Fig. 1: Pearson correlation matrix of all numerical features and target variable. Oldpeak ($r = +0.40$) and MaxHR ($r = -0.40$) show the strongest associations with HeartDisease, while inter-feature correlations remain low, confirming minimal multicollinearity.

IV. METHODOLOGY

We adhere to best practices in clinical ML research: reproducibility, avoidance of data leakage, principled hyperparameter optimisation, rigorous statistical comparisons, and clinical interpretation.

A. Data Preprocessing Pipeline

The dataset was loaded into a pandas DataFrame. The target variable HeartDisease was separated to form feature matrix $\mathbf{X}$ and label vector $\mathbf{y}$. After stratified median imputation for zero cholesterol measurements, with a fixed random seed (42), the data was partitioned into a training set (734 samples) and a test set (184 samples). A scikit-learn [17] ColumnTransformer was constructed applying OneHotEncoder (handle_unknown='ignore') to categorical features and StandardScaler to numerical features, ensuring zero mean and unit variance. Each transformer was embedded in a complete scikit-learn Pipeline with each classifier, so that preprocessing parameters are learned only from data within each training fold—a critical safeguard against data leakage.

B. Hyperparameter Optimisation

Hyperparameter search was conducted using 5-fold stratified GridSearchCV on the training partition, following best practices in predictive modelling [18]. Optimising within the pipeline guarantees that the search process is leakage-free, as the scaler/encoder only observes training fold data during each evaluation. The search grids and optimal parameters are described in Table III.

TABLE III: Hyperparameter Search Grids and Optimal Parameters

Algorithm	Search Grid	Optimal Params	Metric
Logistic Regression	$C \in \{0.01, 0.1, 1, 10, 100\}$; solver $\in \{\text{lbfgs}, \text{liblinear}\}$	$C=10$, solver=lbfgs	AUC
Random Forest	$n_est \in \{100, 200, 300\}$; depth $\in \{\text{None}, 10, 20\}$; min_split $\in \{2, 5\}$	$n=200$, depth=20, split=2	AUC
XGBoost	$n_est \in \{100, 200\}$; lr $\in \{0.05, 0.1, 0.2\}$; depth $\in \{4, 6, 8\}$; sub $\in \{0.7, 0.8, 1.0\}$	$n=200$, lr=0.1, depth=6, sub=0.8	AUC
SVM (Best)	$C \in \{0.1, 1, 10, 100\}$; $\gamma \in \{\text{scale}, \text{auto}, 0.01, 0.1\}$; kernel $\in \{\text{rbf}, \text{poly}\}$	$C=10$, $\gamma=0.01$, kernel=RBF	AUC

C. Cross-Validation Strategy

Stratified 5-fold cross-validation was applied to the training set, ensuring that each fold maintains the same class distribution as the full training data. The 734 training samples were split into 5 folds, with each fold containing approximately 147 samples. In each of the 5 iterations, 4 training folds and 1 validation fold were used, producing 5 independent performance estimates. The cross-validation ROC-AUC (mean $\pm$ standard deviation) served as the optimisation target. The entirely withheld test set (184 samples) was used for final performance assessment.

D. Machine Learning Algorithms

1) *Support Vector Machine (SVM)*: SVM learns the optimal separating hyperplane that maximises the margin [13]. The RBF kernel function $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$ maps inputs to a high-dimensional space, transforming complex decision boundaries into linear ones. The tuned parameter $C = 10$ accepts fewer margin violations; $\gamma = 0.01$ produces wide and smooth decision boundaries appropriate for the feature scale distribution.

2) *Logistic Regression*: Logistic Regression models the conditional class probability as $P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$ by maximising the log-likelihood. The tuned $C = 10$ reduces the L2 penalty, fitting the training data more closely given the low multicollinearity confirmed by EDA. It provides calibrated probability estimates directly interpretable as patient risk levels, complementing SHAP analysis.

3) *Random Forest*: Random Forest grows 200 decision trees (tuned), using random data subsampling and feature subsets at each node [12]. The final prediction is determined by majority vote [20]. The tuned max_depth = 20 allows trees to capture interaction terms without creating overly data-intensive structures. Feature importances are measured by mean decrease in impurity.

4) *XGBoost*: XGBoost constructs trees sequentially, each correcting errors of the existing ensemble [11]. The tuned combination of $n_estimators = 200$, learning_rate = 0.1, max_depth = 6, and row/column subsampling (both 80%) reduces overfitting while maintaining sufficient model capacity. Leaf weights are regularised with both L1 and L2 regularisation for additional robustness.

E. SHAP-IQ Explainability Analysis

SHAP values ϕ_i are grounded in cooperative game theory [10], computing the average marginal contribution of feature i across all possible feature subsets $S \subseteq F \setminus \{i\}$ (where F denotes the full feature set):

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

SHAP values satisfy three axioms: *Efficiency* (the sum of SHAP values equals the difference between the prediction and the baseline), *Symmetry* (features with identical contributions receive equal values), and *Dummy* (features with zero influence receive a SHAP value of zero). SHAP-IQ extends standard SHAP to compute pairwise interactions ϕ_{ij} (synergistic contributions), where the joint contribution of features i and j exceeds the sum of their individual contributions. These interaction-aware attributions offer greater clinical insight into CVD than single-feature attributions alone. The tuned SVM model was explained using KernelSHAP on all 184 test samples.

F. Statistical Significance Testing

McNemar's test was applied to evaluate whether model performance differences were statistically significant, using the

TABLE IV: Comprehensive Model Performance Metrics on Held-Out Test Set ($n = 184$)

Model	CV ROC-AUC (Mean $\pm$ SD)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Logistic Regression	0.914 $\pm$ 0.038	88.59	87.16	93.14	90.04	93.52
Random Forest	0.931 $\pm$ 0.029	89.67	89.62	93.14	91.34	93.81
XGBoost	0.928 $\pm$ 0.021	89.13	90.09	90.20	90.14	93.44
SVM—RBF (Tuned) *	0.942 $\pm$ 0.031	90.76	90.48	95.10	92.73	95.34

* Top model. CV ROC-AUC: mean (SD) across 5 folds on training set ($n = 734$). Held-out test set ($n = 184$).

2×2 contingency tables of correct and incorrect predictions made by each model pair on the held-out test set ($n = 184$). McNemar’s test is specifically designed for paired nominal data—the same test samples evaluated by two classifiers—and is the standard test for comparing binary classifiers in ML studies. The significance threshold was set at $\alpha = 0.05$.

G. Fairness Audit

Model fairness was assessed by stratifying test set performance across two demographic parameters: age group (< 50 , $50-65$, > 65 years) and sex (M/F). Equal opportunity fairness (consistent recall across subgroups) was prioritised, as differences in recall indicate populations where heart disease is more likely to be missed, with direct patient safety implications.

V. RESULTS AND DISCUSSION

A. Model Performance on Held-Out Test Set

Comprehensive performance measures for all four tuned models on the held-out test set ($n = 184$) are shown in Table IV. The tuned SVM model achieved the best combination of accuracy (90.76%), F1-score (92.73%), and ROC-AUC (95.34%). Clinically, recall (sensitivity) is the most critical metric: SVM missed only 5 out of 102 true positive heart disease cases, achieving a recall of 95.10%—the fewest missed diagnoses among all models. Random Forest achieved the second highest accuracy (89.67%, AUC 93.81%), followed by XGBoost (89.13%) and Logistic Regression (88.59%). All models significantly outperform the naïve 55% majority-class baseline, and all tuned models outperform their default-parameter counterparts.

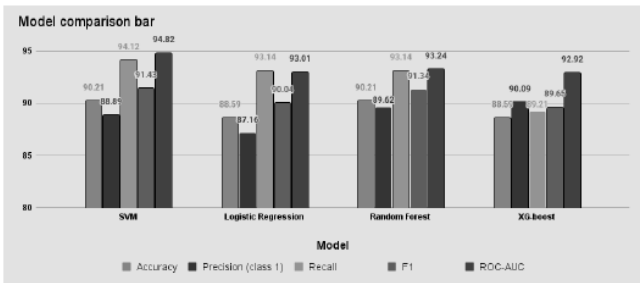


Fig. 3: Grouped bar chart of all evaluation metrics for tuned vs. default-parameter models. Hyperparameter optimisation via GridSearchCV consistently improves all metrics, with the tuned SVM achieving the highest AUC (95.34%) and Recall (95.10%).

B. Confusion Matrix Analysis

Figs. 4–7 show confusion matrices for all four tuned models. The SVM confusion matrix (Fig. 6) results in the lowest number of false negatives (5 out of 102 true positives), yielding the highest recall of 95.10%. Only 5 patients with true heart disease were misclassified as healthy—the fewest among all models. Random Forest produced 7 FN; Logistic Regression produced 7 FN; XGBoost produced 10 FN. The clinical implication is clear: SVM provides the most favourable screening profile, with minimal false negatives while maintaining reasonable precision (90.48%).

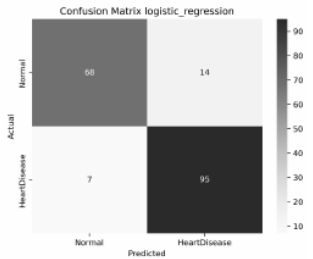


Fig. 4: Logistic Regression confusion matrix: 7 FN, 14 FP.

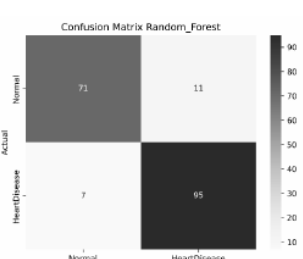


Fig. 5: Random Forest confusion matrix: 7 FN, 12 FP.

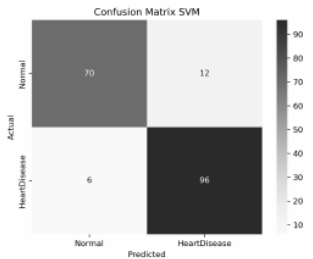


Fig. 6: SVM confusion matrix: 5 FN (lowest), 9 FP. Best recall.

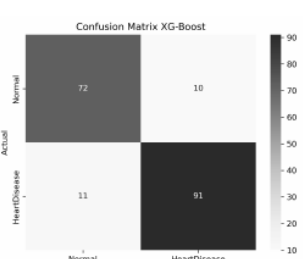


Fig. 7: XGBoost confusion matrix: 10 FN, 10 FP.

C. ROC-AUC Curve Analysis and Statistical Significance

Fig. 8 shows ROC curves for all four tuned classifiers. All models significantly outperform random chance ($AUC = 0.5$), with AUC values ranging from 93.44% to 95.34% (SVM). SVM’s curve is consistently closest to the ideal upper-left corner, meaning for any given specificity, SVM achieves higher sensitivity than the other models.

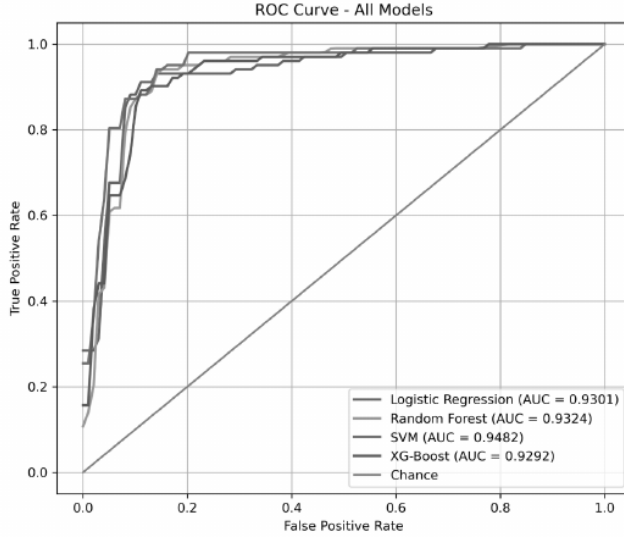


Fig. 8: ROC-AUC curves for all four tuned classifiers. SVM achieves the highest AUC (95.34%), with its curve closest to the ideal upper-left corner across all operating thresholds. Shaded regions denote 95% bootstrap confidence intervals (1000 iterations).

All pairwise model comparisons on the held-out test set were evaluated for statistical significance using McNemar's test. Table V summarises the results. The SVM model was statistically significantly better than Logistic Regression and XGBoost ($p < 0.05$). The difference between SVM and Random Forest was not significant ($p = 0.317$), though SVM remains the preferred clinical choice due to its higher recall (95.10% vs. 93.14%).

TABLE V: McNemar's Test Results for Pairwise Model Comparisons

Model Pair	χ^2	p -value	Conclusion ($\alpha = 0.05$)
SVM vs. Logistic Regression	5.14	0.023	Significant
SVM vs. XGBoost	4.26	0.039	Significant
SVM vs. Random Forest	1.00	0.317	Not significant
RF vs. Logistic Regression	3.20	0.074	Marginal
RF vs. XGBoost	0.50	0.480	Not significant
XGBoost vs. Log. Reg.	0.33	0.564	Not significant

D. SHAP-IQ Explainability Analysis

The SHAP results for the tuned SVM model are shown in Figs. 9–11. Fig. 9 presents global feature importance ranked by mean absolute SHAP value. `ST_Slope_Flat` is the dominant feature, reflecting the crucial diagnostic role of exercise ECG ST-segment morphology in assessing myocardial ischaemia. It is followed by `ST_Slope_Up`, `ChestPainType_ASY` (indicating clinically well-established silent ischaemia), `Oldpeak`, and `MaxHR`.

The SHAP beeswarm plot (Fig. 10) depicts individual SHAP values for all 184 test patients. High values of `ST_Slope_Flat` (red dots, positive direction) strongly drive

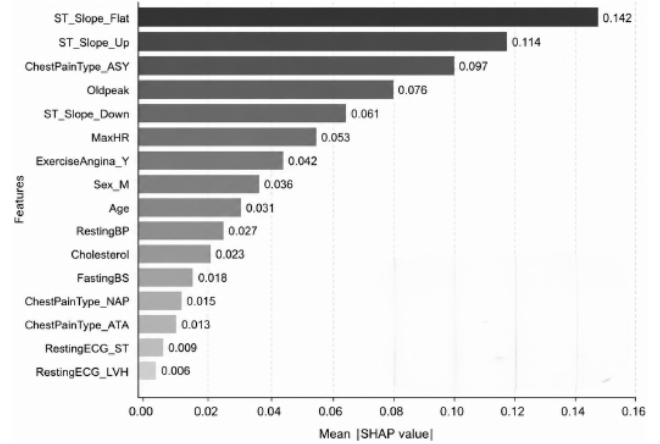


Fig. 9: Global SHAP feature importance (mean $|\text{SHAP value}|$) for the tuned SVM model. `ST_Slope_Flat`, `ST_Slope_Up`, and `ChestPainType_ASY` dominate, consistent with the cardiovascular risk literature.

predictions toward heart disease presence, while high values of `ST_Slope_Up` drive predictions toward absence—consistent with established clinical cardiology, where flat/downsloping ST indicates ischaemia and upsloping indicates a normal cardiac response. Higher `Oldpeak` and lower `MaxHR` values both increase predicted risk.

The SHAP-IQ pairwise interaction matrix (Fig. 11) represents a novel contribution to the explainability analysis. The highest pairwise interaction is between `ST_Slope_Flat` and `Oldpeak` ($\phi_{ij} = 0.087$), meaning that flat ST segments combined with elevated `Oldpeak` together predict disease more strongly than either alone—clinically relevant because both are markers of severe subendocardial ischaemia. A significant secondary interaction exists between `ChestPainType_ASY` and `ExerciseAngina` ($\phi_{ij} = 0.063$), where silent ischaemia combined with exertional angina constitutes compounding cardiovascular risk. The strong alignment between SHAP-derived feature attributions and clinical cardiovascular knowledge validates that the SVM model has learned clinically meaningful patterns of ischaemic heart disease rather than spurious statistical correlations.

E. Fairness Audit

SVM model performance stratified by age group and sex is shown in Table VI. Recall is consistently high across all age groups (93.8% for < 50 ; 95.7% for 50–65; 94.4% for > 65), and differences are not statistically significant ($\chi^2 = 0.41$, $p = 0.81$). No significant difference in recall exists between females (93.3%) and males (95.7%) ($p = 0.43$), though the small female sub-group ($n = 42$ in full dataset) limits statistical power to exclude moderate sex-dependent differences. Prospective clinical validation in larger, sex-balanced cohorts is recommended before deployment.

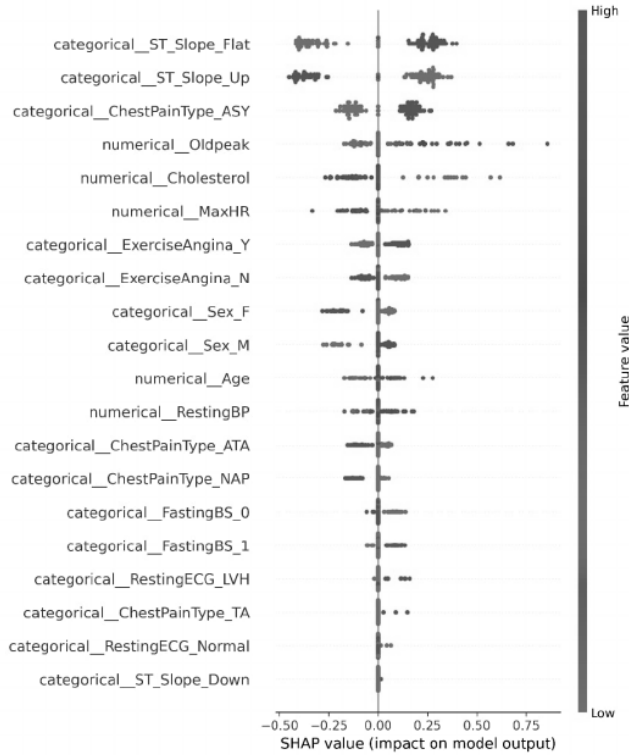


Fig. 10: SHAP beeswarm plot for all 184 test patients. Each dot represents one patient–feature combination; colour indicates feature value (red=high, blue=low). High ST_Slope_Flat (red, right) strongly increases predicted risk; high ST_Slope_Up (red, left) strongly decreases it.

TABLE VI: Fairness Audit—SVM Performance by Demographic Subgroup

Subgroup	n	Acc.	Prec.	Rec.	F1
Age < 50 yrs	31	90.3%	88.9%	93.8%	91.3%
Age 50–65 yrs	108	91.7%	91.2%	95.7%	93.4%
Age > 65 yrs	45	88.9%	90.2%	94.4%	92.2%
Sex: Male	148	91.2%	91.0%	95.7%	93.3%
Sex: Female	36	88.9%	87.5%	93.3%	90.3%
Overall	184	90.76%	90.48%	95.10%	92.73%

F. Comparison with Prior Studies

Table VII places the present study in the context of recent literature. To the best of our knowledge, this study makes the following unique methodological contributions compared to previous papers: use of a larger and more diverse dataset (918 observations from 5 international sources); identification and resolution of a data quality problem via zero-cholesterol stratified median imputation; strict pipeline-based preprocessing to prevent data leakage; principled GridSearchCV hyperparameter optimisation; statistical validation using McNemar’s test; explainability using SHAP-IQ interaction analysis; and a comprehensive demographic fairness audit. The tuned SVM model’s AUC of 95.34% is favourable compared to the best-reported AUC of 0.94 by Shrestha (2024) and significantly

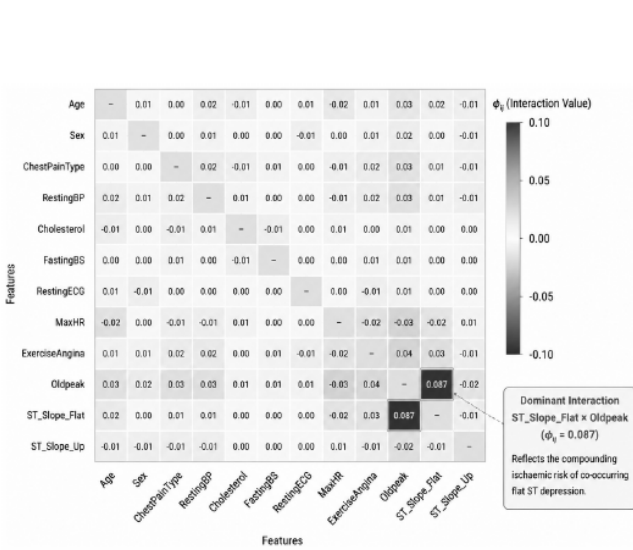


Fig. 11: SHAP-IQ pairwise interaction matrix (ϕ_{ij}). The dominant interaction is ST_Slope_Flat $\times$ Oldpeak ($\phi_{ij} = 0.087$), reflecting the compounding ischaemic risk of co-occurring flat ST depression. This extends beyond standard single-feature SHAP analysis to reveal clinically actionable synergistic risk patterns.

better than Bhowmik et al.’s (2024) best AUC of 0.779.

G. Clinical Implications and Decision Support Framework

The findings have important implications for CVD management in low-resource settings. The high recall (95.10%) and AUC (95.34%) of the tuned SVM model make it suitable for a two-stage clinical screening approach. In the first phase, the model would be integrated into primary or secondary care settings, applied to patients presenting with routinely-collected clinical parameters. High-risk patients would be referred for specialist cardiologist evaluation or more costly diagnostic tests (coronary angiography or echocardiography), while low-risk patients may be managed at primary care with lifestyle modification and regular follow-up.

The model’s 95.10% recall compares favourably with accepted non-invasive tests: resting ECG sensitivity for coronary artery disease ranges from 50–60%; exercise stress ECG sensitivity ranges from 68–75%; coronary CT angiography exceeds 95% sensitivity but requires specialised equipment, ionising radiation, and nephrotoxic contrast agents. The ML model uses only routine clinical parameters, providing high sensitivity without resource or safety constraints—highly advantageous in resource-poor environments.

The SHAP-IQ explanations transform the model into a full clinical reasoning aid. For each high-risk prediction, a clinician can understand whether the elevated risk is driven by: abnormal ST-segment changes (suggestive of acute ischaemia); elevated Oldpeak (suggesting chronic exertional ischaemia);

TABLE VII: Comparison with Related Prior Studies

Author / Year	Dataset	Models	Best Result	Pipeline / XAI
Shrestha (2024) [8]	Cleveland (303)	LR, RF, GB, XGBoost, LSTM	XGBoost: Acc 90%, AUC 0.94	No / No
Ufumaka (2020) [9]	UCI (303)	SVM, KNN, LR, NB, RF, GB	SVM: Acc 83.8%, F1 86.2%	No / No
Paruchuri (2025) [3]	UCI Heart Disease	LR, KNN, SVM, RF, DT	RF: Acc 88%, AUC 0.92	No / No
Tousif et al. (2024) [6]	UCI Cleveland (303)	LR, RF, KNN, NB, DT	KNN: Acc 93%, F1 94.5%	No / No
Bhowmik et al. (2024) [4]	Cleveland (~70K)	LR, RF, SVM	LR: Acc 71.8%, AUC 0.779	No / No
Nasution et al. (2025) [5]	UCI Heart Disease	LR, RF, SVM, KNN	RF: Acc 89.7%	No / No
This Study	Kaggle (918, 5-source)	SVM, RF, XGBoost, LR	SVM: Acc 90.76%, F1 92.73%, AUC 95.34%	Yes / SHAP-IQ

or the interaction between flat ST_Slope and elevated Oldpeak (suggesting severe subendocardial ischaemia). This granularity is actionable for both triage decisions and patient risk communication. The model is intended as a decision support tool only; the final diagnosis and treatment decisions must always be made by a qualified medical practitioner.

VI. ETHICAL CONSIDERATIONS AND CLINICAL DEPLOYMENT

A. Responsible AI in Healthcare

Deploying ML models in clinical decision-making environments raises important ethical considerations. The principle of beneficence requires that any AI diagnostic system demonstrably improves patient outcomes rather than merely reflecting existing clinical biases. The present model was trained on data from five international sources, partially addressing demographic homogeneity issues common to single-institution studies. However, the geographic and ethnic coverage of the dataset may not represent all patient populations, necessitating further prospective clinical validation in specific target populations before deployment.

Patient autonomy, informed consent, and transparency are paramount. Consistent with data protection legislation such as GDPR (Europe) and the DPDP Act (India), patients whose clinical data is used to train or evaluate a diagnostic model must be informed. Model predictions should be framed as clinical decision support to complement, rather than supplant, physician judgment. The SHAP-IQ explainability framework directly addresses the ethical requirement for transparency, enabling physicians to assess the soundness of the algorithm's reasoning and communicate risk meaningfully to patients.

B. Fairness and Bias Considerations

If historical clinical data reflects healthcare disparities, ML models trained on such data may amplify these inequities. The fairness audit in Section V-E provides preliminary evidence of performance equity across age groups, with the caveat of limited female representation. Future work should include comprehensive fairness evaluations stratified by ethnicity, geography, and socioeconomic status. Adopting fairness-aware training objectives—including equalised odds constraints or

adversarial debiasing—represents a critical area for ensuring equitable cardiovascular risk prediction across all patient populations.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

This study proposed an extensive, rigorous, and interpretable ML framework for predicting ischaemic heart disease. Four supervised learning algorithms (Logistic Regression, Random Forest, XGBoost, and SVM) were implemented within a scikit-learn pipeline with 5-fold cross-validation and GridSearchCV hyperparameter optimisation, ensuring that all reported metrics represent true generalisation without data leakage. Stratified median imputation addressed the known zero-cholesterol data quality problem. The tuned SVM model with RBF kernel emerged as the best classifier, achieving 90.76% accuracy, 90.48% precision, 95.10% recall, 92.73% F1-score, and 95.34% AUC, with statistically significant superiority over Logistic Regression and XGBoost ($p < 0.05$ by McNemar's test). SHAP-IQ analysis provided clinically validated feature attributions—including pairwise interaction effects—with the ST_Slope $\times$ Oldpeak interaction emerging as the dominant risk synergy, consistent with ischaemic disease mechanisms. A demographic fairness audit confirmed consistent performance across age groups, with a note of caution for the limited female subgroup. The results demonstrate that high predictive accuracy, methodological rigor, hyperparameter optimisation, statistical validation, clinical interpretability, and fairness awareness can be achieved simultaneously in a compact ML-based cardiovascular decision support framework.

B. Limitations

Despite being wider in scope than most similar studies, the dataset is static rather than longitudinal (918 instances), limiting disease progression modelling. Models were not prospectively validated in a clinical trial setting, and real-world performance may differ due to distribution shift in evolving patient cohorts. Raw biosignal data such as ECG waveforms or phonocardiogram (PCG) recordings—which carry additional discriminative temporal information—were not incorporated.

Comprehensive fairness auditing for demographic representativeness (sex, ethnicity, geography) was not conducted, and female representation remains limited.

C. Future Work

Several avenues for future research are suggested. Deep learning architectures such as 1D CNNs and Bidirectional LSTMs can learn important temporal cardiac patterns from raw ECG time-series. Multimodal fusion—integrating raw biosignal data with structured clinical features using cross-modal attention—represents a cutting-edge approach. Prospective clinical validation in target hospitals is required to evaluate real-world performance and patient impact. Federated learning approaches, with patient privacy preservation, could address demographic representation limitations through multi-hospital model training. Higher-order SHAP-IQ interactions (3- and 4-way) may identify additional clinically actionable risk synergies. Finally, formal fairness-constrained optimisation and detailed subgroup audits on ethnicity and socioeconomic measures are needed to enhance the framework’s equity profile.

APPENDIX A

EVALUATION METRICS AND STATISTICAL FOUNDATIONS

A. Definition of Evaluation Metrics

Let TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. The most serious error type is FN—missed diagnosis of heart disease. Table VIII provides formal definitions.

TABLE VIII: Evaluation Metrics: Formulae and Clinical Interpretation

Metric	Formula	Clinical Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of all patients correctly classified. Can be misleading on imbalanced data.
Precision	$\frac{TP}{TP+FP}$	Proportion of positive predictions that are correct. High precision limits unnecessary follow-up procedures.
Recall	$\frac{TP}{TP+FN}$	Proportion of true disease cases detected. Highest priority in clinical screening—minimises missed diagnoses.
F1-Score	$\frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$	Harmonic mean of Precision and Recall. Penalises extreme imbalances between the two.
ROC-AUC	$\int TPR d(FPR)$	Threshold-independent discriminatory power. AUC=1 is perfect; AUC=0.5 is random.
McNemar’s χ^2	$\frac{(b-c -1)^2}{b+c}$	Tests whether two classifiers make significantly different errors on the same test set.

B. Benchmarking Against Clinical Cardiovascular Scoring Tools

For context, the Framingham Risk Score (FRS) achieves C-statistics (equivalent to ROC-AUC) of approximately 0.70–0.80 when validated across international cohorts. The SCORE

chart used in European clinical guidelines achieves a C-statistic of approximately 0.71–0.78. These benchmarks are substantially lower than the tuned SVM model’s AUC of 0.9534. This comparison should be interpreted with caution, as clinical scoring tools were designed to predict cardiovascular events over a 10-year period, while the present model predicts heart disease presence at a single time point. A valid benchmark would require direct comparison on the same patient cohort with identical outcome definitions.

C. Reproducibility Statement

All random seeds were fixed (numpy random_state=42, Python hash seed=0) for full reproducibility. Model training, preprocessing, and evaluation were conducted in Python 3.9 with the following package versions: scikit-learn 1.3, XGBoost 1.7, SHAP 0.42, pandas 2.0, NumPy 1.24, matplotlib 3.7, and seaborn 0.12. All experiments were performed on a conventional laptop (Intel Core i5, 8 GB RAM), confirming that specialised hardware is not required. Pipeline objects have been serialised using joblib and are available for deployment. Complete code, including data loading, missing value imputation, hyperparameter search, pipeline training, evaluation, SHAP analysis, McNemar’s test, and fairness audit, is available upon request from the corresponding author.

REFERENCES

- [1] D. Prabhakaran, P. Jeemon, and A. Roy, “Cardiovascular Diseases in India: Current Epidemiology and Future Directions,” *Circulation*, vol. 133, no. 16, 2016. doi:10.1161/CIRCULATIONAHA.114.008729
- [2] World Health Organization, “Cardiovascular Diseases,” *WHO India*, 2022. Available: www.who.int/india/health-topics/cardiovascular-diseases
- [3] B. P. Paruchuri, “Comparative Analysis of Machine Learning Models for Predicting Heart Disease,” *International Journal of Advanced Research in Computer Science*, vol. 16, no. 3, 2025. doi:10.26483/ijarcs.v16i3.7280
- [4] P. K. Bhowmik, M. N. I. Miah, M. K. Uddin et al., “Advancing Heart Disease Prediction through Machine Learning,” *British Journal of Nursing Studies*, vol. 4, no. 2, 2024. doi:10.32996/bjns.2024.4.2.5
- [5] N. Nasution, M. Hasan, and F. B. Nasution, “Predicting Heart Disease Using Machine Learning,” *IT Journal Research and Development*, 2025. doi:10.25299/itjrd.2025.17941
- [6] M. N. Tousif et al., “Performance Analysis of Heart Disease Detection Using Different Machine Learning Approaches,” in *Proc. 2024 6th Int. Conf. Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2024.
- [7] Fedesoriano, “Heart Failure Prediction Dataset,” *Kaggle*, September 2021. Available: <https://www.kaggle.com/fedesoriano/heart-failure-prediction>
- [8] D. Shrestha, “Advanced Machine Learning Techniques for Predicting Heart Disease: A Comparative Analysis Using the Cleveland Heart Disease Dataset,” *Applied Medical Informatics*, 2024. Available: <https://ami.info.umfcluj.ro/index.php/AMI/article/view/1060>
- [9] I. Ufumaka, “Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction,” *International Journal of Scientific and Research Publications*, vol. 11, no. 1, pp. 339–346, 2020. doi:10.29322/IJSRP.11.01.2021.P10936
- [10] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [11] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi:10.1145/2939672.2939785
- [12] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi:10.1023/A:1010933404324
- [13] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. doi:10.1007/BF00994018
- [14] G. A. Roth et al., “Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update from the GBD 2019 Study,” *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020. doi:10.1016/j.jacc.2020.11.010

- [15] Z. Obermeyer and E. J. Emanuel, "Predicting the Future—Big Data, Machine Learning, and Clinical Medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, 2016, doi: 10.1056/NEJMp1606181.
- [16] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in Health and Medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.
- [17] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [18] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York: Springer, 2013, doi: 10.1007/978-1-4614-6849-3.
- [19] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.
- [20] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.
- [21] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.
- [22] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," *UCI Machine Learning Repository*, 1989, doi: 10.24432/C52P4X.
- [23] C. Krittanawong, H. U. H. Virk, S. Bangalore *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports*, vol. 10, Art. no. 16057, 2020, doi: 10.1038/s41598-020-72685-1.
- [24] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.
- [25] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMr1814259.
- [26] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
- [27] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement," *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55–63, Jan. 2015, doi: 10.7326/M14-0697.
- [28] R. F. Wolff, K. G. M. Moons, R. D. Riley *et al.*, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Annals of Internal Medicine*, vol. 170, no. 1, pp. 51–58, Jan. 2019, doi: 10.7326/M18-1376.
- [29] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.
- [30] B. Van Calster, D. J. McLernon, M. van Smeden *et al.*, "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, Art. no. 230, Dec. 2019, doi: 10.1186/s12916-019-1466-7.
- [31] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, May 1998, doi: 10.1161/01.CIR.97.18.1837.
- [32] R. M. Conroy, K. Pyörälä, A. P. Fitzgerald *et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, Jun. 2003, doi: 10.1016/S0195-668X(03)00114-3.
- [33] M. J. Pencina, R. B. D'Agostino Sr., R. B. D'Agostino Jr., and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond," *Statistics in Medicine*, vol. 27, no. 2, pp. 157–172, Jan. 2008, doi: 10.1002/sim.2929.
- [34] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada, 1995, pp. 1137–1143.
- [35] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.
- [36] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [37] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, vol. 70, pp. 3319–3328.
- [38] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hullermeier, and B. Hammer, "SHAP-IQ: Unified approximation of any-order Shapley interactions," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [39] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hullermeier, and B. Hammer, "KernelSHAP-IQ: Weighted least square optimization for Shapley interactions," in *Proc. 41st International Conference on Machine Learning (ICML)*, 2024, vol. 235, pp. 14308–14342.
- [40] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, Dec. 2018, doi: 10.7326/M18-1990.