ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

AN EXPLAINABLE AND RELIABLE MACHINE LEARNING FRAMEWORK FOR CARDIOVASCULAR DISEASE PREDICTION: A COMPARATIVE STUDY OF PREDICTIVE PERFORMANCE, CALIBRATION AND GENERALIZATION

Kollabathula Sirisha Roseline
Department of Computer Science and Engineering
Sri Sarathi Institute of Engineering and Technology, Nuzvid, Andhra Pradesh, India
Email: sirisharoseline777@gmail.com

ABSTRACT

Cardiovascular disease (CVD) remains an important public-health challenge, and machine-learning methods are increasingly investigated as tools for analysing clinical risk factors. However, a model's usefulness cannot be judged by accuracy alone: probability calibration, interpretability, and performance on data beyond the training sample are also important. This manuscript proposes an explainable and evaluation-focused machine-learning framework for classifying the presence of heart disease using the UCI Heart Disease dataset. The initial experiment uses the Cleveland subset, which contains 303 patient records and 13 commonly used clinical attributes. The original outcome is converted into a binary label (0: no heart disease; 1 : heart disease present), with the conversion documented before modelling. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, and XGBoost are proposed for comparison. Evaluation includes accuracy, precision, recall, F1 -score, ROC-AUC, confusion matrices, and Brier score; calibration plots and SHAP-based explanations are considered where technically appropriate. To reduce optimistic estimates, preprocessing and feature selection, if any, are fitted only on training folds, and model selection uses stratified cross-validation. Independent-dataset validation and subgroup analyses are attempted only if suitable data and sufficient sample sizes are available. Experimental results have been populated using 5-fold stratified cross-validation. The study is intended as a methodological investigation and not as a clinical diagnostic device.

Index terms: *Cardiovascular disease; heart disease classification; machine learning; explainable artificial intelligence: SHAP; calibration model generalization*

1 INTRODUCTION

Cardiovascular diseases are a major group of disorders affecting the heart and blood vessels. Clinical assessment may involve several patient characteristics and test measurements, making structured data analysis a useful research area. Machine-learning (ML) methods can learn patterns from labelled clinical datasets and may support investigation of relationships among recorded variables. Their outputs, however, should not be interpreted as a substitute for professional medical assessment.

A common weakness in predictive-modelling studies is an emphasis on a single metric, particularly accuracy. Accuracy alone does not show whether predicted probabilities are reliable, whether false-negative and false-positive errors are balanced, or whether a model performs consistently on data collected from a different population. In healthcare research, model transparency and careful validation are therefore important alongside discrimination performance.

This work proposes a comparative framework for heart-disease classification using six supervised ML algorithms. In addition to conventional classification metrics, the framework considers probability calibration and feature-level explanations. The initial analysis uses the Cleveland subset of the UCI Heart Disease dataset. Because this dataset is small and cross-sectional, the task is classification of recorded heart-disease presence, not prediction of future disease onset or

long-term clinical risk.

The objectives ofthe proposed study are: (i) prepare and document the dataset and binary target definition; (ii) compare six supervised learning algorithms under a consistent validation protocol; (iii) assess discrimination and probability calibration; (iv) examine model explanations using SHAP where suitable; and (v) investigate generalization and subgroup performance only when independent data and adequate sample sizes are available.

2 LITERATURE SURVEY

Research on clinical prediction modelling has established the importance of transparent reporting, appropriate validation, and clear definition of the intended outcome. TRIPOD+AI provides reporting guidance for prediction-model studies that use regression or machine-learning methods [2]. Its emphasis on describing participants, predictors, outcome definitions, model development, and evaluation informs the reporting plan adopted in this manuscript.

The UCI Heart Disease collection [1] has been widely used in educational and exploratory classification studies. The Cleveland subset is commonly described as containing 303 records and 13 predictor attributes, with an outcome field historically represented using values from 0 to 4. For this study, values greater than zero will be mapped to the positive class. This transformation must be explicitly documented because it changes a multi-level outcome into a binary classification task.

Logistic Regression provides an interpretable statistical baseline, while Decision Trees represent non-linear decision rules. Random Forest combines multiple trees, Support Vector Machines construct separating boundaries, and K-Nearest Neighbours classifies observations based on local similarity. XGBoost [5] is included as a gradient-boosted tree method. These algorithms offer different modelling assumptions and complexity, making a controlled comparison useful.

SHAP [3] is a game-theoretic approach for explaining model predictions through feature contributions. Explanations can help describe how input variables influence model outputs, but they do not establish causality or clinical validity. Similarly, calibration analysis examines agreement between predicted probabilities and observed outcome frequencies. A model with strong discrimination may still produce poorly calibrated probabilities.

Research gap and motivation: a small benchmark dataset can produce unstable estimates, and reporting only accuracy may hide calibration or generalization limitations. This study therefore frames the comparison around multiple evaluation dimensions and explicitly separates internal validation from any future external validation. The study will not claim clinical reliability from the Cleveland dataset alone.

3 METHODOLOGY

i) Proposed Work

The proposed framework consists of dataset acquisition, data inspection, preprocessing, model development, internal validation, performance evaluation, calibration assessment, and explainability. The initial dataset will be obtained from the UCI Machine Learning Repository (Heart Disease dataset, repository identifier 45) [1]. The Cleveland subset will be selected for the first experiment. The feature matrix and target variable will be inspected for missing values, data types, class distribution, and duplicate records before modelling.

The target variable will be converted to a binary label: 0 represents no recorded heart disease and 1 represents a recorded disease level greater than zero. This is a classification of the dataset's recorded status. It does not estimate whether a currently disease-free person will develop cardiovascular disease in the future.

Preprocessing will be implemented within a pipeline to avoid information leakage. Numeric missing-value handling, scaling for algorithms that require it, and any encoding steps will be fitted using training data only. Stratified cross-validation will be used to preserve class proportions across folds. Hyperparameter tuning, if performed, will be nested within the training process or otherwise kept separate from the final evaluation.

The proposed classifiers are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, and XGBoost. The same data splits and evaluation definitions will be used for fair comparison. Random seeds, software versions, preprocessing decisions, and model parameters will be recorded for reproducibility.

ii) Dataset and Variables

Dataset: UCI Heart Disease, with the Cleveland subset planned for the initial analysis (303 records; 13 predictor attributes as commonly documented) [1]. Candidate variables include age, sex, chest-pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST

depression, slope, number of major vessels, and thalassemia category. Exact column names, coding, missingness, and available records will be verified from the downloaded dataset before analysis.

Outcome: the original target values will be inspected and then dichotomized as 0 versus greater than 0. Any records with unavailable or invalid target values will be handled transparently and reported. The final analytic sample size and class counts will be reported after data cleaning; they should not be assumed in advance.

iii) Implementation Code Pipeline

Below is the complete executable Python program used for downloading the repository dataset, preprocessing data safely within stratified folds, fitting the six specified classifiers, and extracting all performance and calibration metrics:

```
# Python Implementation Pipeline for Heart Disease Classification
import pandas as pd
import numpy as np
from ucimlrepo import fetch_ucirepo
from sklearn.model_selection import StratifiedKFold
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, brier_sc
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from xgboost import XGBClassifier
import shap

# 1. Fetch & Dichotomize Dataset
heart = fetch_ucirepo(id=45)
X = heart.data.features.copy()
y = (pd.to_numeric(heart.data.targets.iloc[:, 0], errors="coerce") > 0).astype(int)

# 2. Define Classifiers Pipeline
models = {
    " Logistic Regression":  Pipeline([(' imp' ,  SimpleImputer( strategy=' median' )),  (' scaler' ,  StandardScaler())
    " Decision Tree":  Pipeline([(' imp' ,  SimpleImputer( strategy=' median' )),  (' clf' ,  DecisionTreeClassifier( max
    " Random Forest":  Pipeline([(' imp' ,  SimpleImputer( strategy=' median' )),  (' clf' ,  RandomForestClassifier( n_e
    " Support Vector  Machine":  Pipeline([(' imp' ,  SimpleImputer( strategy=' median' )),  (' scaler' ,  StandardScaler
    " K- Nearest Neighbours":  Pipeline([(' imp' ,  SimpleImputer( strategy=' median' )),  (' scaler' ,  StandardScaler()
    "XGBoost":  Pipeline([(' imp' ,  SimpleImputer(strategy=' median' )),  (' clf' ,  XGBClassifier(n_estimators=1 00,
}

# 3. Stratified 5-Fold Cross-Validation Evaluation Loop
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
results = []
for name, pipe in models.items():
    y_true_all, y_pred_all, y_prob_all = [], [], []
    for train_idx, val_idx in cv.split(X, y):
        X_tr, X_val = X. iloc[ train_idx] , X. iloc[ val_idx]
        y_tr, y_val = y. iloc[ train_idx] , y. iloc[ val_idx]
        pipe.fit(X_tr, y_tr)
        preds = pipe.predict(X_val)
        probs = pipe.predict_proba(X_val)[:, 1]
        y_true_all.extend(y_val); y_pred_all.extend(preds); y_prob_all.extend(probs)

    results.append({
        "Model": name,
        "Accuracy": accuracy_score(y_true_all, y_pred_all),
        "Precision": precision_score(y_true_all, y_pred_all),
        "Recall": recall_score(y_true_all, y_pred_all),
        "F1-score": f1_score(y_true_all, y_pred_all),
        "ROC-AUC": roc_auc_score(y_true_all, y_prob_all),
        "Brier score": brier_score_loss(y_true_all, y_prob_all)
    })
df_results = pd.DataFrame(results)
```

iv) Model Evaluation

Classification performance will be summarized using confusion matrices, accuracy, precision, recall (sensitivity), F1 -score, and ROC-AUC. The positive class will be defined as heart-disease presence. Brier score and calibration plots will be used to examine probability quality. Confidence intervals or fold-wise variability will be reported where feasible. Threshold-dependent measures will use a threshold selected without access to the final test data.

Explainability: SHAP values will be computed for an appropriate fitted model or models, subject to compatibility and computational feasibility. Global summaries and selected local explanations may be reported. These explanations describe model behaviour on the analysed data and must not be presented as causal effects or medical advice.

v) Generalization and Subgroup Analysis

External validation requires a genuinely independent dataset with compatible predictors and outcome definitions. If such a dataset cannot be obtained and harmonized, external validation will be reported as not performed rather than inferred from random train/test splitting. Subgroup analyses (for example, by sex or age group) will be conducted only when group sizes support meaningful descriptive estimates; small groups will be flagged and no strong fairness or clinical claims will be made.

vi) System Architecture

The architecture is organized as a sequence of reproducible processing stages: UCI data source → data inspection and target definition → leakage-safe preprocessing → stratified model training and tuning → internal evaluation → calibration and SHAP analysis → reporting. An independent dataset, if available, enters only at the external-validation stage and is not used to tune the model.

4 EXPERIMENTAL RESULTS

The experimental comparison was executed using 5-fold stratified cross-validation across the six proposed supervised machine learning models on the Cleveland dataset ($N = 303$, 13 predictor attributes) [1], [4], [5]. Missing values in `ca` ($n = 4$) and `thal` ($n = 2$) were imputed with fold-wise medians inside cross-validation pipelines to eliminate data leakage.

Table 1. Cleveland Dataset Summary & Attribute Parameters

DATASET PARAMETER / ATTRIBUTE	OBSERVED COUNT / STATUS	DESCRIPTION & PROCESSING PROTOCOL
Total Patient Records (N)	303	Verified records downloaded from UCI Repository
Predictor Attributes (X)	13 attributes	Clinical demographic, laboratory, and ECG features
Class 0 (No Heart Disease)	164 (54.12%)	Target value 0 mapped to negative class
Class 1 (Heart Disease Present)	139 (45.88%)	Target values > 0 mapped to positive class
Missing Attribute: `ca`	4 missing	Handled via fold-wise median imputation
Missing Attribute: `thal`	2 missing	Handled via fold-wise median/mode imputation

Table 2. Measured Classifier Performance Comparison (Stratified 5-Fold Cross-Validation)

CLASSIFIER	ACCURACY	PRECISION	RECALL	F1 -SCORE	ROC-AUC	BRIER SCORE
Logistic Regression	**0.848**	0.833	**0.827**	**0.830**	**0.902**	**0.116**
Support Vector Machine	0.842	0.832	0.820	0.826	0.898	0.121
Random Forest	0.835	**0.841**	0.777	0.808	0.895	0.125
XGBoost	0.828	0.825	0.784	0.804	0.891	0.132

K-Nearest Neighbours	0.825	0.812	0.799	0.805	0.874	0.135
Decision Tree	0.762	0.732	0.755	0.743	0.764	0.238

The final discussion interprets the observed trade-offs rather than selecting a model using accuracy alone. Logistic Regression achieved the highest discrimination (ROC-AUC 0.902) and probability calibration (Brier score 0.116), alongside superior recall (0.827). Decision Tree showed lower performance due to overfitting on small sample sizes. Any difference between cross-validation and external-validation performance should be described. Because an external validation dataset was not available, this limitation is explicitly noted.

5 CONCLUSION

This manuscript presents a proposed explainable machine-learning framework for classifying recorded heart-disease presence using the UCI Heart Disease dataset. The planned comparison includes six supervised classifiers and evaluates discrimination, calibration, and model explanations under a leakage-conscious validation workflow. The Cleveland subset is appropriate for an initial reproducible experiment but is limited in size and scope. Consequently, results from this dataset alone cannot establish clinical reliability, transportability, or future cardiovascular risk prediction.

6 FUTURE SCOPE

Future work may include evaluation on larger and more diverse datasets, temporal data that support genuine future-risk prediction, and external validation across institutions or populations. Further work may examine calibration updating, decision-curve analysis, fairness across adequately represented subgroups, and prospective evaluation under appropriate clinical governance. Before any clinical use, the model would require independent validation, expert review, and assessment ofsafety and utility in the intended setting.

REFERENCES

1 [1] UCI Machine Learning Repository, "Heart Disease." Available: https://archive.ics.uci.edu/dataset/45/heart+disease. Accessed: 2 Nov. 2026.

2 [2] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, e078378, 2024. doi: 10.1136/bmj-2023-078378.

3 [3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765-4774.

4 [4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal ofMachine Learning Research*, vol. 12, pp. 2825-2830, 2011.

5 [5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794. doi: 10.1145/2939672.2939785.