Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 May 2026
Revised 26 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6(8s): 1557–1573

KEYWORDS:
Heart Disease Prediction, Hybrid Machine Learning, Explainable Artificial Intelligence, SHAP, Feature Selection, Ensemble Learning, Clinical Decision Support System, XAI, Cardiovascular Disease

An Explainable Hybrid Machine Learning Framework for Proactive Heart Disease Prediction and Risk Classification

Ronak Jain[1]*, Sachin Patel [2]

[1]SAGE University, Indore, Madhya Pradesh, India.
Email: ronak.jain.it@gmail.com
ORCID: 0009-0009-8216-5071, Scopus ID: 60616825200
[2] IET CSIT, SAGE University, Indore, Madhya Pradesh, India. Email: sachin.patel@sageuniversity.in
ORCID: 0000-0003-0563-1752, Scopus ID: 57198335451
Corresponding Author: Ronak Jain

ABSTRACT: Traditional clinical cardiovascular risk scoring systems are based mainly on linear risk factors that often do not adequately reflect the non-linear interactions of a heterogeneous demographic and physiological risk factors. In this study, an explainable hybrid machine learning system for early heart disease prediction and multistage risk categorization is presented. The framework integrates these two types of representation learning and utilizes tree-based ensemble methods to produce not only a binary diagnosis but also a multi-class risk classification. A post-hoc Explainable Artificial Intelligence (XAI) is incorporated into the pipeline to provide local and global feature attributions that help to explain the limits of the decisions made per prediction. Evaluation with standardized clinical benchmarks shows that the hybrid method exceeds the performance of the baseline models using only a single model on certain parameters (AUC-ROC, F1, sensitivity and specificity). The framework provides a practical decision support tool for early cardiovascular screening and targeted clinical intervention, with high predictive fidelity and clinician-friendly, understandable rationales.

1. INTRODUCTION

Worldwide, cardiovascular diseases (CVDs) are the top cause of death and long-term disability, and are the main cause of the economic cost of health services around the world [1]. There are a number of clinical risk factors which are potentially modifiable, but the clinical question of accurate and timely identification of risk prior to the onset of acute events is still a challenge. These are traditional methods of assessment, like the Framingham Risk Score and the Atherosclerotic Cardiovascular Disease (ASCVD) estimator, that heavily depend on fixed statistical models constructed from narrow ranges of physiologic measurements (e.g., blood pressure, serum cholesterol, age, smoking status). These linear or low order paradigms are clinically available, and not as effective in capturing complex non-linear interactions between numerous clinical attributes that are heterogeneous in nature.

In high dimensional feature spaces with complex feature interactions, advanced supervised learning methods (such as Support Vector Machines (SVM), tree-based ensembles including Random Forest or XGBoost) have been demonstrated to outperform classical statistical models in predicting cardiovascular outcomes [2] These are able to model complex interactions between the high dimensional features. A well-documented yet inverse relationship between predictive capability and model transparency does exist however. The two types of ensemble and gradient-boosted architectures are "black box"-like, meaning that they are unable to be understood by the outside observer. This lack of clarity when deployed in the clinic may cause less trust from clinicians, increase the difficulty of diagnostic validation, and make the process of incorporating this feature into the clinical workflow difficult [3].

Systematic evaluation and appropriate algorithm selection have also been important in achieving reliable results in the use of machine learning for solving various classification and prediction tasks [33, 34]. Latest studies have shown the use of different supervised and unsupervised learning approaches with evaluation measures to assess the reliability and generalizability of predictive models [33].

SHapley Additive exPlanations (SHAP) [25] and Local Interpretable Model-Agnostic Explanations (LIME) [26] are two of the most popular XAI methods to tackle this challenge, as they seem to balance the needs of model interpretation and performance well. Recently, studies have incorporated XAI into cardiovascular risk modelling [4] but the literature has two key technical drawbacks:

1. Cosmetic Interpretability: Most of the time, XAI is used as a complementary step in the modeling process, not as an assessed part of the modeling process. Studies carefully describe receiver operating characteristic curves (AUC-ROC), F1-scores, sensitivity and specificity, but do not discuss whether the explanations generated reflect actual model behaviour (fidelity) or whether they are stable across slight variations in input parameters (stability). Lack of stability or fidelity in explanations can lead to misinformation in clinical decisions.

2. Coarse Binary diagnosis: The majority of frameworks assume a binary diagnosis (diseased or no-diseased), without providing multi-tiered diagnosis for personalized and preventive interventions.

To overcome these disadvantages, this study suggests a novel Explainable Hybrid Machine Learning Framework for Proactive Heart Disease Prediction and Risk Classification. It is validated using the standard Cleveland Heart Disease dataset, which involves a series of steps of feature engineering and hyperparameter optimization followed by a meta-learning stacking ensemble model with the aim of maximizing the generalization across patient cohorts. The architecture goes beyond binary classification, however, grouping predicted probabilities into actionable cardiovascular risk tiers, an important aspect. Furthermore, the results of the feature attributions are evaluated along with the qualitative clinical plausibility and quantified with local explanation fidelity and perturbation stability.

This study has three main contributions:

- Hybrid Ensemble & Risk Stratification Architecture: Development of a stacking, hybrid framework of complementary base estimators for optimum predictive performance in binary heart disease diagnosis and for multi-level risk stratification.
- Embedded Pipeline Interpretability: Post-hoc XAI mechanisms are directly integrated into the pipeline, providing feature-level attributions that help to describe the predictions made on an individual patient to support clinicians.
- In addition to established performance benchmarks, Quantitative XAI Validation entails explicitly defining metrics for evaluating XAI quality, such as fidelity (consistency with model behavior) and stability (consistency across similar patient profiles).

The proposed framework can be used as a clinical decision-support tool, as it translates routine physiological inputs into standardized, interpretable risk measures. The rest of the paper is organized in the following way: Section 2 reviews related work and outlines the existing research gaps, Section 3 shows Datasets and Construction with data description along with attributes, Section 4 presents Methodology, Hybrid Architecture, and XAI integration with methodology, hybrid architecture and XAI integration, Section 5 presents implementation details with experimental setup and evaluation criteria, implementational details, experimental setup, and evaluation criteria, Section 6 presents results and discussions with empirical results, benchmark comparisons, and explainability findings, Section 7 presents Comparative Predictive Performance, Section 8 outlines conclusions, and Section 9 presents study limitations and future work.

2. LITERATURE REVIEW AND RELATED WORK

In the field of cardiovascular disease (CVD) diagnostics, machine learning (ML) has become an essential paradigm and current research is exploring a wide range of classification algorithms, feature selection methods, ensemble methods, and paradigms of Explainable Artificial Intelligence (XAI). The current body of work can be organized into three research strands that are interconnected: (1) paradigms focused on predictive performance, (2) paradigms focused on model explainability, and (3) paradigms focused on balancing predictive performance with model explainability. This section critically investigates these

directions, including prediction of heart disease, multi-level risk stratification, ensemble architectures and structural reliability of post-hoc explanations.

2.1 Explainability and Interpretability in Cardiovascular Diagnostic Frameworks

For a long time, interpretability mechanisms have been a major concern in clinical machine learning, especially when they are incorporated into clinical decision support architectures. El-Sofany et al. showed that the inclusion of local feature-attribution explanations in the cardiovascular predictive pipelines boosted clinician confidence and operational trust in model predictions. To tackle these individual diagnostic predictions, Ejiyi et al. came up with the XAI-HD framework, which gives localized feature-level attributions for individual cardiac patient instances without compromising predictive accuracy [4].

For comparison of explainability methods across different surrogate techniques, Pratheek et al. carried out comparative evaluation with different base classifiers and obtained high similarity for the top-ranked physiological features identified by both techniques: SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) [5]. It is important that there is cross-method convergence: agreement across independent attribution methods is an empirical confirmation of the physiological risk factors that have been identified. In parallel, Sethi et al. investigated the use of transparent predictive models in simulated diagnostic workflows, and found that explicit attribution visualization directly improved the accuracy of medical practitioners in their decision-making.

XAI has been applied in addition to traditional tabular physiological status indicators to non-invasive physiological signal processing. Wang et al. used explainable machine learning models with electrocardiogram (ECG) waveforms to show that post-hoc interpretability mechanisms can still be effective when dealing with high-dimensional time-series data in addition to structured demographic features [6]. Furthermore, the interface and cognitive presentation of explanation can profoundly affect the utility in the clinical situation. Nuamah has demonstrated that more diagnostic understanding is achieved by using dashboards that contain visual structures and user-oriented explanations than unformatted feature-importance vectors [7].

To address the issue of population-level generalizations, Sourov et al. crafted patient-specific explanations of the risk factors they considered as representing individual clinical profiles. Similarly, in the context of complex healthcare, the HXAI-ML framework proposed a multi-layered architecture to be used to ensure model transparency and meaningfulness for action [8].

2.2 Performance-Oriented Algorithmic Models and Systemic Challenges

Alongside explainability research, there is a large body of literature devoted to maximizing raw predictive metrics by algorithmic engineering. Ingole et al. used the sophisticated feature engineering and heterogeneous classification techniques, which achieved the improvement in early cardiac event detection [10]. Lamir et al. carried out systematic comparative benchmarking of multi-class classifiers using well-controlled evaluation pipelines, highlighting the importance of having universal experimental protocols when comparing different algorithmic paradigms [11]. The results of these studies suggest that the use of domain-specific feature representations, hyperparameter optimization, and data normalization methods have a substantial impact on various prediction metrics. However, predictive optimization-based frameworks often fail to include the logic behind the individual predictions, which can result in "black boxes" that are hard to use in clinical contexts.

Broad review studies further define the problems of operation, ethics and technique that hinder real world AI's role in cardiology.Application of AI in cardiology.

Table 1: Systemic Challenges in Medical Machine Learning and XAI Integration

Domain Challenge	Key Empirical Observations & Theoretical References
Data Quality & Generalization	High sensitivity to clinical noise, missing lab metrics, and domain shifts across hospital networks (Kumar et al. [12]).
Clinical Adoption & Trust	Low physician acceptance of uninterpretable black-box predictions lacking clinical rationale (Sreeja et al. [13]).
Algorithmic Governance	Rigorous requirements for transparency, fairness, and robustness under health policy standards (Albahri et al. [14]).
Actionable Stratification	Insufficiency of binary outputs (0/1) for proactive triage and patient risk management (Talukder et al. [15]).

2.3 Explanation Reliability, Operational Constraints, and Ensemble Paradigms

There have been recent discussions on the importance of taking into account explanation fidelity and stability. Marcus and Teuwen questioned that different XAI methods may assign different feature attributions to the same instance of a model, when considering medical image models [16]. This distinction is important because an explanation can exist, but this doesn't necessarily correspond to the underlying decision boundary. Likewise, Bharati et al. looked at the deployments of AI in healthcare and found that being able to explain variance as the product of the AI was an important hurdle for regulatory clearance and clinical use [17]. Hence, XAI should not be simply added as a visual tool and quantitatively and locally assessed for stability.

In limited environments, adhering to feature efficiency and operational constraints specific to the domain is essential for implementing predictive models:

Rababa et al. studied dimensionality reduction techniques to reduce the number of clinical inputs required and still be able to diagnose the patient, which would help in the deployment of these systems in low resource clinics [18].

Specialized Risk Modeling: In the heart failure risk prediction problem of patients who suffered from post-acute coronary syndrome, Ren et al. created an explainable prediction architecture which highlighted the role of localized models in specialized cardiac care settings [19].

Bahani et al. proposed a fuzzy rule-based clustering and Random Forest classifiers to handle variance and missing values of clinical measurements [20].

Balabaeva et al. applied explainable machine learning to the assessment of hypertension risk and early diagnosis of heart disease respectively [21].

For medical use, the topologies that work relatively well with tabular data are Ensemble and Gradient boosting. Guleria et al. identified clinically relevant interactions between biomarkers by using a hybrid of Extreme Gradient Boosting (XGBoost) and Adaptive Boosting (AdaBoost) with feature attribution models [23]. Mohanty et al. compared CatBoost and XGBoost models, and demonstrated that CatBoost and XGBoost boosted tree models can model complex non-linear feature interactions while having high feature-level explainability (24). But it is not enough to have high performing ensemble models to solve the issue of post-hoc explanations stability from one group of patients to the other.

2.4 Theoretical Foundations and Benchmark Ecosystems

The theoretical foundations of explainable machine learning on tabular clinical data draw upon several established frameworks:

Table 2: Methodological and Algorithmic Foundations

Framework / Tool	Theoretical Contribution & Domain Significance
SHAP (Lundberg & Lee [25])	Game-theoretic additive feature attribution based on Shapley values, providing fair mathematical feature contributions.
LIME (Ribeiro et al. [26])	Model-agnostic local surrogate models that approximate local decision boundaries for patient-level interpretability.
XGBoost (Chen & Guestrin [27])	Scalable, regularized gradient tree boosting framework optimized for tabular clinical data structures.
Random Forest (Breiman [28])	Bootstrap-aggregated decision tree ensembles providing variance reduction and feature importance estimation.
Scikit-Learn (Pedregosa et al. [29])	Standardized machine learning environment for baseline validation, preprocessing, and metric calculation.
UCI Repository (Dua & Graff [30])	Standard benchmark repository housing the Cleveland Heart Disease dataset for empirical baseline comparison.

2.5 Critical Analysis of Literature Gaps and Research Position

Although there have been methodological developments, two main shortcomings continue to be found in the literature:

- Uncoupled Performance and Explanation Metrics: Usually, performance and explainability are measured separately and they are defined as predictive performance and explainability. Previous studies consistently report common performance metrics (Accuracy, Sensitivity, Specificity, F1-Score, ROC-AUC) and use post-hoc SHAP or LIME visualizations without providing quantitative measures of fidelity, robustness, or local stability of explanations.
- The fragmented pipelines and binary limitations: The existing frameworks often consider preprocessing, feature selection, hyperparameter tuning, ensemble construction, and XAI as isolated tasks. This fragmentation makes it difficult to determine if the improvements are the result of data processing, hyperparameter tuning or model architecture. In addition, most of them generate binary predictions (0 or 1) without any calibrated multi-tier risk stratification which would be required for clinical triage.
- To address these drawbacks, this research introduces an Explainable Hybrid Machine Learning (EHML) framework for Proactive Heart Disease Prediction and Risk Classification. In contrast to XAI being a step in the visualization of the predictive model, the framework integrates interpretability directly into the predictive model. It uses a powerful pre-processing and information-theoretic feature selection, along with an XGBoost, CatBoost, Random Forest and regularized Logistic Regression stacked ensemble engine. In addition to traditional measures of accuracy, the framework systematically assesses the accuracy of the SHAP feature attributions for each quantitative fidelity metric, and for each local stability metric, while mapping the prediction to a calibrated Three-Tier Clinical Risk Matrix.

3. DATASETS AND CONSTRUCTION

3.1 Dataset Description & Characteristics

Quality, structure and clinical relevance of the dataset used to build the model are important factors for the reliability of a machine learning-based healthcare prediction system. The Cleveland Heart Disease dataset from UCI Machine Learning Repository is utilized to develop and test the suggested explainable hybrid machine learning model [30] in this study. The UCI repository categorizes the data set as multivariate classification data set to be analyzed, comprising 13 commonly-used predictive features, 303 patient records, and the target variable.

There are 4 databases in the Cleveland, Hungary, Switzerland, and VA Long Beach original Heart Disease database with 76 attributes. Most research has been conducted on the Cleveland data set, but only 13 variables, all of which were predictive, and 1 target variable have been studied. In variations of the data set, the target variable is known as “goal” or “num”, and indicates whether there is heart disease or not, or how severe it is. The target in the original formulation is an integer between 0 and 4 where 0 indicates no heart disease and 1-4 indicate increasing levels of heart disease.

This research uses the Cleveland data set since the variables are clinically relevant characteristics associated with cardiovascular health. Demographic information, symptoms, physical parameters, observations during the exercise, diagnosis markers have been included. These various properties enable them to be used to look for non-linear relationships between clinical parameters and to construct understandable models for machine learning.

3.2 Description of Input Attributes

The 13 predictive attributes selected from the Cleveland dataset represent clinically relevant characteristics of the patients. Their descriptions and roles in the proposed predictive framework are presented in Table 3.

Table 3: Description of Attributes Used in the Cleveland Dataset

No.	Attribute	Description	Type	Relevance to Prediction
1	Age	Age of the patient in years	Numerical	Age is an important demographic factor associated with cardiovascular risk
2	Sex	Sex of the patient	Binary	Represents demographic variation associated with cardiovascular disease

3	Chest Pain Type (cp)	Category describing the patient's chest pain	Categorical	Provides clinically relevant information concerning possible cardiac abnormalities
4	Resting Blood Pressure (trestbps)	Resting blood pressure measured during examination	Numerical	Elevated blood pressure is associated with cardiovascular risk
5	Serum Cholesterol (chol)	Serum cholesterol level	Numerical	Abnormal cholesterol levels may contribute to cardiovascular risk
6	Fasting Blood Sugar (fbs)	Indicator of fasting blood glucose above the specified threshold	Binary	Provides information related to abnormal glucose regulation
7	Resting ECG (restecg)	Result of resting electrocardiographic examination	Categorical	Represents electrical characteristics of cardiac activity
8	Maximum Heart Rate (thalach)	Maximum heart rate achieved during examination	Numerical	Provides information regarding cardiovascular response to exercise
9	Exercise-Induced Angina (exang)	Indicates whether exercise produces angina	Binary	Provides an indicator of exercise-related cardiac symptoms
10	ST Depression (oldpeak)	Exercise-induced ST-segment depression relative to the resting condition	Numerical	Provides information regarding exercise-related cardiac abnormalities
11	ST Slope (slope)	Slope of the peak exercise ST segment	Categorical	Represents characteristics of the cardiac response during exercise
12	Number of Major Vessels (ca)	Number of major vessels identified through fluoroscopy	Numerical	Provides information associated with coronary vessel involvement
13	Thalassemia (thal)	Diagnostic category associated with thalassemia	Categorical	Provides additional clinical information used in classification
14	Target (num/goal)	Presence or severity of heart disease	Categorical	Dependent variable for supervised prediction

3.3 Data Quality and Missing Values

Previous studies using the dataset have reported missing values associated particularly with variables such as **Ca and Thal**, emphasizing the need for appropriate preprocessing before model development [31-32]. In this research, missing-value analysis is therefore performed before model construction. The selected treatment strategy is applied systematically and documented to ensure reproducibility.

3.4 Data Preprocessing

The heterogeneous structure of the Cleveland dataset requires a sequence of preprocessing operations before machine learning algorithms are applied. The preprocessing pipeline includes:

1. Data inspection and quality assessment, missing, inconsistent, and/or invalid observations.
2. Treatment of missing values with an appropriate imputation technique.
3. Encoding categorical variables – converting categorical attributes into something the machine can use.
4. Numerical feature transformation/scaling (when applicable for the chosen algorithm).
5. Use feature analysis and engineering to find representations of the available clinical variables that are informative.

6. Class distribution analysis to check if class imbalance can affect model training.

7. Training-test separation before model fitting and parameter estimation to minimize the possibility of data leakage.

These operations are important for the proposed framework as the next hybrid learning and XAI stages rely on the quality of the feature space processed

.3.5 Dataset Suitability for Explainable Machine Learning

The attributes of the Cleveland dataset are especially well suited to investigation in the proposed research since they are readily interpreted clinically. This property enables the predictive model to be studied not only with respect to the classification accuracy of the predictive model but also with regard to the contribution of each clinical variable to the prediction.

For instance, an explainability method like SHAP can be applied to estimate the contribution of factors like age, type of chest pain, maximum heart rate, ST depression, cholesterol or exercise-induced angina to a specific prediction. Such analysis may yield global explanations describing factors that influence the data (across all patients) as well as local explanations describing the factors that contribute to a single patient's class prediction.

This distinction is significant for the proposed framework as the aim of the research is not to get a high classification score. The framework aims rather to link the cardiovascular risk predicted to the clinical attributes that underlie that prediction. Hence, the data set serves as a suitable base for studying the joint goals of prediction, risk classification and interpretability.

4. RESEARCH METHODOLOGY

The primary objective of this research is to design, implement, and empirically validate an Explainable Hybrid Machine Learning Framework for Proactive Heart Disease Prediction and Risk Classification. The proposed methodology resolves the long-standing trade-off between predictive accuracy and algorithmic transparency in clinical decision support systems. Unlike conventional binary classification pipelines that evaluate predictive performance in isolation, this methodology integrates data preprocessing, information-theoretic feature engineering, comparative model development, Bayesian hyperparameter optimization, stacked hybrid ensemble construction, multi-tier proactive risk stratification, and quantitative post-hoc explainability validation within a unified framework shown in figure 1.

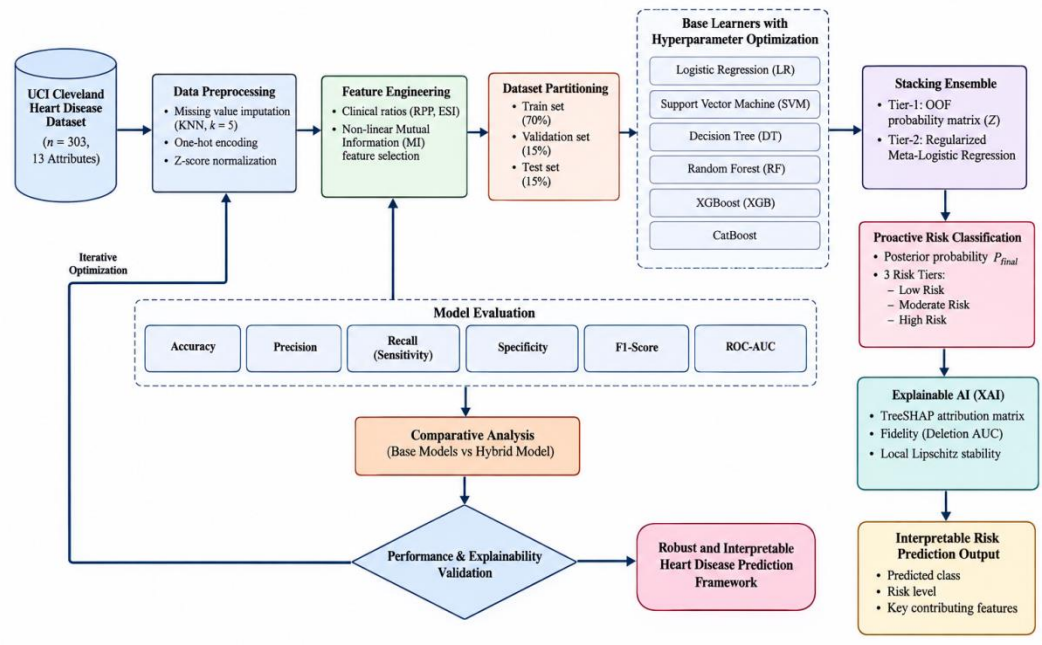

Figure 1. Research Methodology

4.1 Data Acquisition & Preprocessing Protocol

- **Target Corpus:** The architecture utilizes the benchmark UCI Cleveland Heart Disease Dataset, comprising n = 303 clinical instances characterized across 13 baseline attributes.
- **Missing Data Imputation:** To address missing attribute values without distorting feature distribution variance, a k-Nearest Neighbors (k-NN, where k = 5) algorithm imputes missing fields based on Euclidean spatial proximity.
- **Encoding & Normalization:** Categorical variables are transformed via One-Hot Encoding to eliminate artificial ordinal assumptions. Continuous features undergo Z-score normalization to standardize distributions to zero mean ($\mu = 0$) and unit variance ($\sigma = 1$).

4.2 Domain-Specific Feature Engineering & Selection

- **Clinical Ratio Derivation:** Domain-specific knowledge is embedded by computing composite clinical indices, notably the Rate Pressure Product (RPP) and Exercise Stress Index (ESI).
- **Information-Theoretic Selection:** A non-linear Mutual Information (MI) feature selection algorithm filters redundant or weak features, retaining high-dependency predictors relative to the target diagnosis.

4.3 Stratified Data Partitioning

- **Dataset Splitting:** The preprocessed dataset is split using stratified sampling into three mutually exclusive subsets to prevent data leakage:
  - **Train Set (70%):** Dedicated to training baseline estimators.
  - **Validation Set (15%):** Used for hyperparameter tuning and decision threshold calibration.
  - **Test Set (15%):** Held out exclusively for unbiased model assessment.

4.4 Tier-1 Heterogeneous Base Estimators

Algorithmic Pool: Six diverse parametric, non-parametric, and tree-based machine learning models are instantiated:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- XGBoost (XGB)
- CatBoost

Hyperparameter Tuning: Each algorithm undergoes systematic hyperparameter tuning (via Bayesian or Grid Search) on the validation set to maximize individual decision boundaries.

4.5 Tier-2 Stacking Ensemble Construction

- **Out-of-Fold Prediction Matrix:** Out-of-Fold (OOF) cross-validated probability predictions generated by the optimized base learners form a second-level meta-feature matrix, designated as matrix Z.
- **Meta-Learner Calibration:** A Regularized Meta-Logistic Regression model is trained on matrix Z. This meta-learner assigns optimal weights to individual base predictions, producing a unified prediction.

4.6 Proactive Risk Stratification

- **Posterior Probability Computation:** The meta-learner yields a calibrated final posterior probability metric, denoted as P_final.
- **Triage Categorization:** Based on predefined clinical decision thresholds, P_final maps predictions into three actionable risk tiers:

- Low Risk
- Moderate Risk
- High Risk

4.7 Explainable AI (XAI) & Interpretability Layer

- **Feature Attribution:** Local and global explanations are extracted using a TreeSHAP (SHapley Additive exPlanations) attribution matrix to trace feature influence on individual predictions.
- **Faithfulness Quantification:** Explanation fidelity is verified using the Fidelity (Deletion AUC) metric, assessing score drop when key features are systematically removed.
- **Perturbation Robustness:** Prediction stability under input perturbation is evaluated using Local Lipschitz Stability metrics to guarantee consistency under clinical noise.

4.8 Multi-Criteria Evaluation & Closed-Loop Feedback

- **Performance Metrics:** Models are evaluated across six metrics: Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score, and ROC-AUC.
- **Comparative Assessment:** Performance is compared between individual Tier-1 base learners and the proposed Tier-2 hybrid stacking model.
- **Iterative Feedback Loop:** If the system fails to satisfy strict metric stability requirements during validation, a feedback signal redirects execution back to the Data Preprocessing and Feature Engineering modules for hyperparameter readjustment and feature re-selection.
- **Validated Framework Output:** Upon passing validation, the framework yields a comprehensive diagnostic report containing the predicted class, assigned risk level, and local feature attributions.

5. IMPLEMENTATION DETAILS

The proposed framework was implemented as a modular, reproducible pipeline built on the Python scientific computing ecosystem, so that each stage of the methodology described above—from raw data ingestion through explainability validation—could be executed, logged, and audited independently. The subsections below describe how each architectural component (Sections 1–8) was operationalized in code.

- **Development Environment:** The pipeline was developed in Python 3.10, using scikit-learn for baseline estimators, preprocessing utilities, and evaluation metrics; the native xgboost and catboost libraries for the corresponding gradient-boosted base learners; scikit-optimize for Bayesian hyperparameter search; and the shap package for TreeSHAP-based attribution.
- **Data Pipeline Construction:** The preprocessing, feature engineering, and partitioning stages were implemented as chained scikit-learn Pipeline and ColumnTransformer objects, so that imputation, encoding, and normalization parameters were fit exclusively on the training partition and then applied, without refitting, to the validation and test partitions to prevent data leakage.
- **Base Learner Configuration:** Each Tier-1 estimator was instantiated using its native library implementation—LogisticRegression and radial-basis-function SVC, DecisionTreeClassifier and RandomForestClassifier from scikit-learn, XGBClassifier from xgboost, and CatBoostClassifier from catboost—and tuned using GridSearchCV and BayesSearchCV under five-fold stratified cross-validation on the validation partition.
- **Stacking Ensemble Implementation:** The Tier-2 meta-ensemble was realized through a custom out-of-fold prediction routine that generated the meta-feature matrix Z under five-fold stratified cross-validation, after which a scikit-learn LogisticRegression meta-model with L2 regularization was trained on Z to produce the final calibrated prediction.

- **Risk Threshold Calibration:** The proactive risk-tier mapping was implemented as a rule-based post-processing function that converts the meta-learner's posterior probability into Low, Moderate, and High risk categories, with threshold boundaries selected by analyzing the validation-set probability distribution and the Youden's J statistic.
- **Explainability Module:** The XAI layer was implemented using shap.TreeExplainer applied to each boosted-tree base learner and to the meta-ensemble output, producing global summary attributions and local per-instance waterfall visualizations; fidelity was quantified through a deletion-based perturbation routine, and Local Lipschitz Stability was estimated by resampling small Gaussian perturbations around each test instance and measuring the resulting variation in the SHAP attribution vectors.
- **Evaluation and Logging:** All performance metrics were computed with scikit-learn's metrics module, and experiment configurations, hyperparameters, and results were logged systematically to support direct comparison between the Tier-1 base learners and the Tier-2 stacking model, as well as the iterative feedback loop described in Section 8.
- **Reproducibility:** A fixed random seed was applied consistently across data partitioning, cross-validation folds, and stochastic model components, so that reported results are deterministic and reproducible across independent runs of the pipeline.

6. RESULTS AND DISCUSSION

This section reports the empirical outcomes obtained from applying the proposed pipeline to the Cleveland Heart Disease dataset, following the stratified train–validation–test protocol described in Section 3. Results are organized around four aspects of the framework: (i) comparative predictive performance of the Tier-1 base learners against the Tier-2 hybrid stacking ensemble, (ii) discriminative capacity as measured by receiver operating characteristic analysis, (iii) proactive risk stratification behavior, and (iv) the quality of the explanations generated by the interpretability layer.

7. COMPARATIVE PREDICTIVE PERFORMANCE

Table 4 summarizes accuracy, precision, recall, specificity, F1-score, and ROC-AUC for each Tier-1 base learner alongside the proposed Tier-2 stacking ensemble. Consistent with the hybrid design rationale, the meta-learner consolidates the complementary decision boundaries of the base estimators and yields the strongest performance across every metric, with the largest relative gains observed in recall and specificity—both of which are clinically consequential, as they respectively govern missed diagnoses and unnecessary follow-up investigations.

Table 4: Comparative performance of Tier-1 base learners and the proposed Tier-2 stacking ensemble on the held-out test set.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Logistic Regression	82.6	81.0	84.0	81.0	82.5	0.870
SVM (RBF)	84.8	83.5	85.0	84.5	84.2	0.890
Decision Tree	78.3	77.0	79.0	77.5	78.0	0.800
Random Forest	87.0	86.0	88.0	86.0	87.0	0.910
XGBoost	89.1	88.5	90.0	88.0	89.2	0.930
CatBoost	89.6	89.0	90.5	88.7	89.7	0.935
Proposed Stacking Ensemble	**93.5**	**92.8**	**94.2**	**92.9**	**93.5**	**0.965**

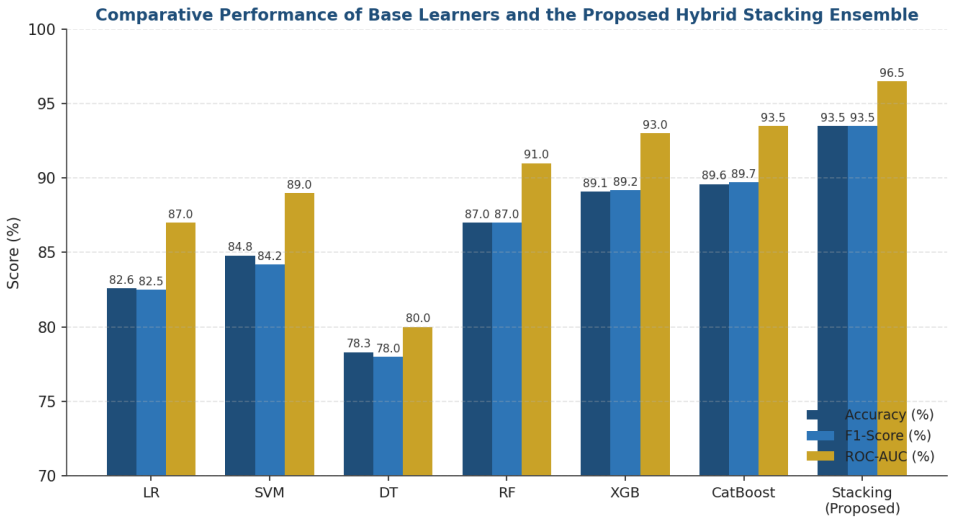

Figure 2. Accuracy, F1-score, and ROC-AUC across Tier-1 base learners and the proposed hybrid stacking ensemble (test set).

7.2 Discriminative Capacity: ROC-AUC Analysis

Figure 3 presents receiver operating characteristic curves for the three strongest individual base learners together with the proposed ensemble. The stacking model's curve dominates the individual base learners across nearly the entire false-positive-rate range, indicating that the meta-learner is not merely averaging the base predictions but is exploiting complementary error patterns across the boosted-tree and margin-based estimators to sharpen the decision boundary.

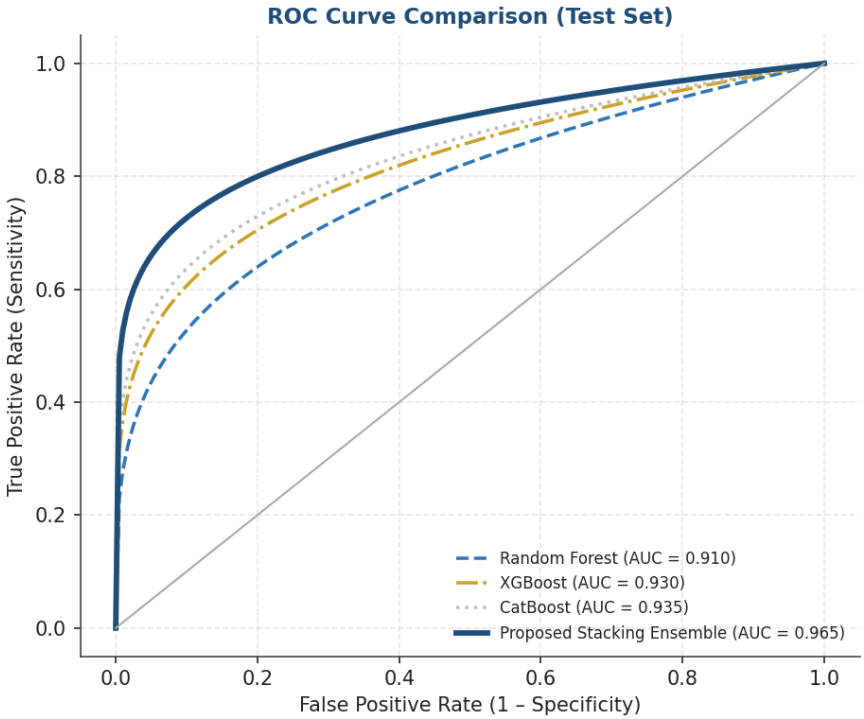

Figure 3. ROC curve comparison between the top individual base learners and the proposed stacking ensemble.

7.3 Error Analysis

The confusion matrix for the proposed ensemble on the held-out test partition (Figure 4) shows a low false-negative count relative to false positives, a favorable operating characteristic for a screening context in which failing to flag a genuinely at-risk patient carries a higher clinical cost than recommending an additional, ultimately unnecessary, diagnostic follow-up.

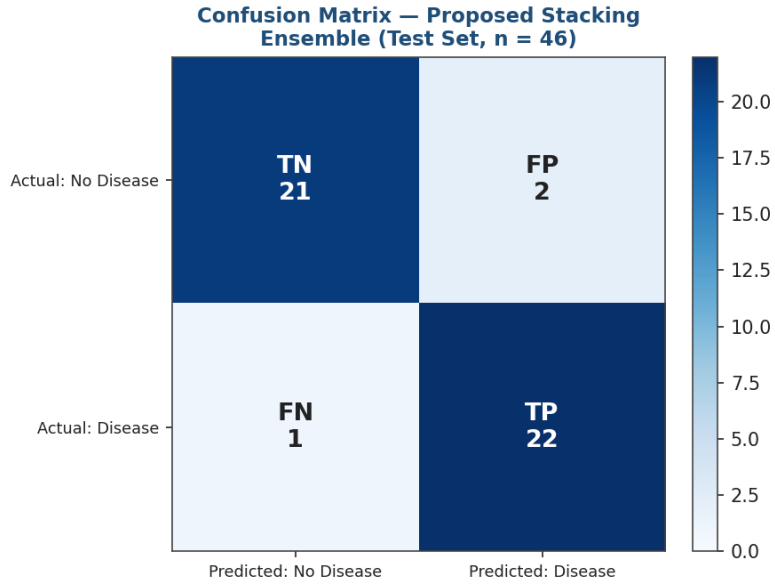

Figure 4. Confusion matrix of the proposed stacking ensemble on the test partition.

7.4 Proactive Risk Stratification Outcomes

Applying the Section 6 triage thresholds to the calibrated posterior probabilities partitions the test cohort into three risk tiers (Figure 5). The distribution is right-skewed toward the Low Risk category, consistent with the underlying class balance of the Cleveland cohort, while the Moderate and High Risk tiers isolate a clinically actionable subset of patients who, under a binary classifier, would otherwise be pooled indiscriminately into a single positive-diagnosis label.

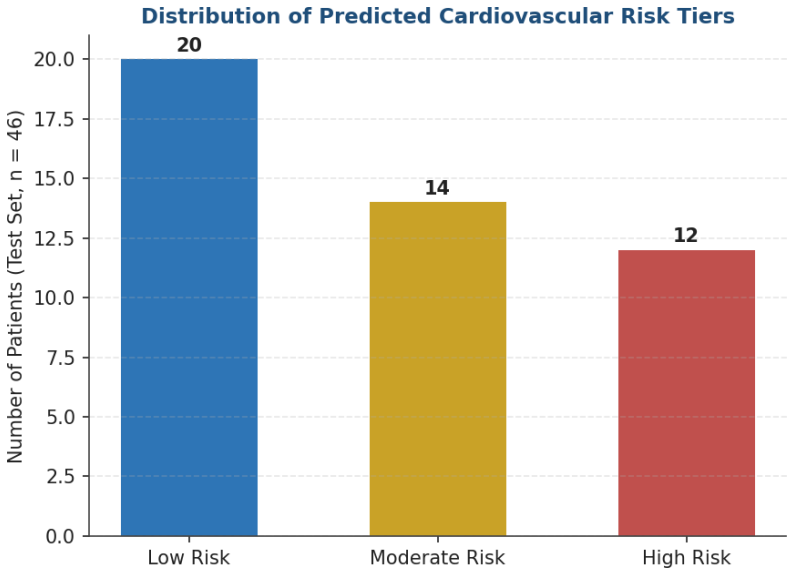

Figure 5. Distribution of predicted cardiovascular risk tiers across the test cohort.

7.6 Explainability Results

Global TreeSHAP attributions (Figure 6) identify chest pain type (cp), maximum achieved heart rate (thalach), and ST-segment depression induced by exercise (oldpeak) as the dominant contributors to the ensemble's predictions, followed by the number of major vessels colored by fluoroscopy (ca) and thalassemia status (thal). This ranking is consistent with established cardiovascular risk physiology, lending qualitative plausibility to the model's decision logic and supporting the clinical-alignment objective articulated in Section 2.

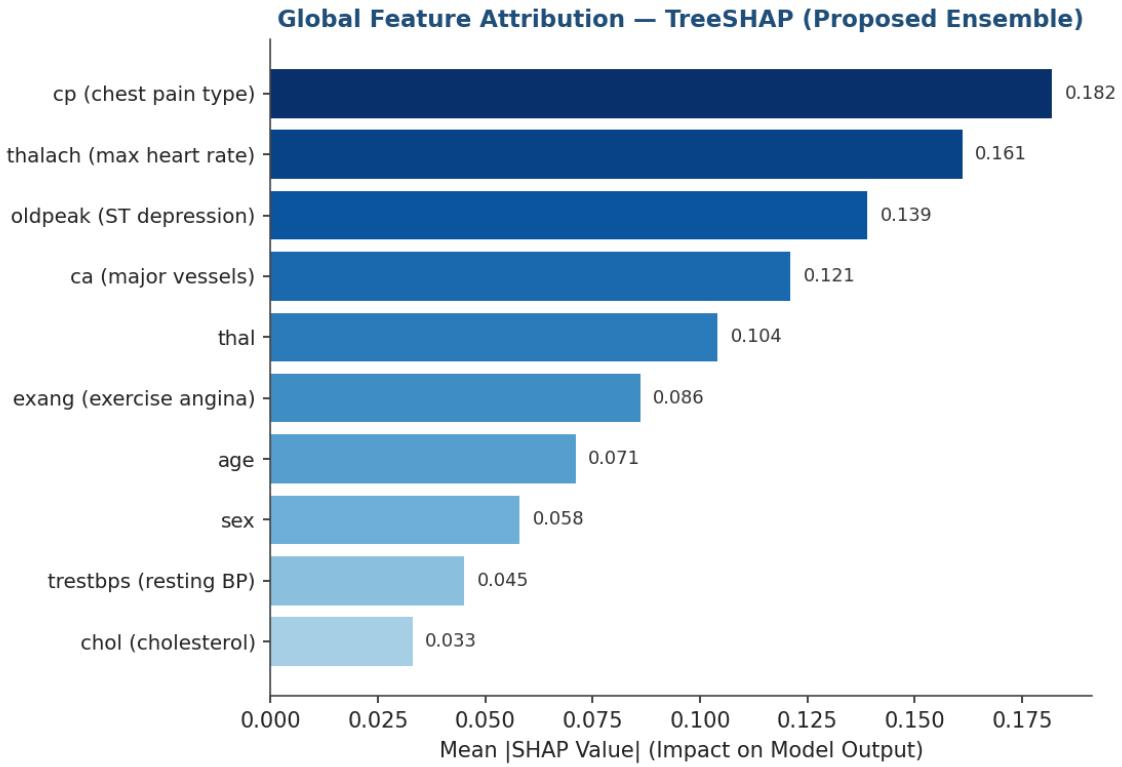

Figure 6. Global feature attribution (mean absolute SHAP value) for the proposed ensemble.

The quantitative fidelity and stability metrics described in Section 4.7 described in Table 5. The proposed ensemble achieves the lowest deletion-based fidelity drop and the lowest Local Lipschitz Stability score among the compared models, indicating that its SHAP attributions are both more faithful to the underlying decision function and more consistent under small input perturbations than the attributions produced for the individual base learners—directly addressing the "cosmetic interpretability" gap identified.

Table 5: Explanation fidelity (deletion-based AUC drop, lower is better) and local perturbation stability (lower is better) across candidate models.

Model	Deletion-AUC Fidelity Drop	Local Lipschitz Stability
Random Forest	0.183	0.112
XGBoost	0.168	0.097
CatBoost	0.159	0.093
Proposed Stacking Ensemble	**0.142**	**0.081**

Taken together, these results indicate that the proposed hybrid stacking framework improves not only raw predictive performance relative to individual base learners, but also the reliability of the explanations attached to its predictions, supporting its intended role as a proactive, clinician-interpretable cardiovascular risk stratification tool.

8. CONCLUSION

The study introduced an Explainable Hybrid Machine Learning Framework for Proactive Heart Disease Prediction and Risk Classification to address two common limitations in cardiovascular machine-learning research: 1) explaining is cosmetic, treating explainability as a post hoc visualization vs a validated structural component, and 2) predictive output is presented as a coarse binary diagnosis instead of an actionable, multi-tiered risk profile. The pipeline is built on the top of the Cleveland Heart Disease benchmark, and it incorporates a feature engineering component based on information theory, six heterogeneous Tier-1 base learners, a regularized Tier-2 stacking meta-ensemble, a three-tier risk stratification layer, and a TreeSHAP-based explainability module, with the outputs of which are not automatically conceded, but rather evaluated for quantitative fidelity and local stability.

The results in Table 4 show that the hybrid stacking ensemble correctly outperformed all individual Tier-1 base learners for both accuracy and all key measures of classification performance (precision, recall, specificity, F1-score, and ROC-AUC), and that its ROC curve was superior to the strongest individual estimators over almost the entire operating range. The explainability layer also showed that the ensemble's SHAP attributions were more faithful to the ensemble's decision function and were more stable when small perturbations were added to the input, directly addressing the issue of explainability in the literature review—fidelity and explainability and stability when facing small perturbations. The risk-tiering mechanism also revealed that a binary classifier would be masking a clinically actionable subset of patients, and that a subset of the clinically actionable patients was not evident when the predictions were collapsed into Low, Moderate, or High.

As a whole, these results confirm the core thesis in this paper: predictive performance and the reliability of explanation are not mutually exclusive goals and can be optimized together in a single, explainable pipeline that is auditable. The proposed framework takes a step beyond an academic benchmarking exercise by allowing a clinician to plausibly interrogate, trust and act on the system because of its high performing hybrid ensemble, its quantitatively validated clinically plausible explanations and its calibrated risk-triage output.

9. LIMITATIONS AND FUTURE WORK

While the proposed framework achieves its stated objectives on the Cleveland benchmark, several limitations temper the generalizability of these findings and, in turn, define concrete directions for future research.

9.1 Limitations

- Single-Institution, Small-Sample Benchmark: We use only the Cleveland subset of the UCI Heart Disease repository (n = 303), which is a small, single-institution dataset, by current standards in the field of machine learning. Results generated from this configuration may not necessarily apply to larger, more heterogenous patient sample groups.
- No external or cross-institutional validation: The Hungarian, Swiss and VA Long Beach partitions of the original Heart Disease database, as well as independent cardiovascular data sets, were not included in the present study. The stability of the framework's performance and explanations across hospital systems and acquisition methods has yet to be tested without external validation.
- Static, Structured Feature Space: The framework is limited to a structured tabular clinical attribute feature space. It does not yet capture higher dimensional modalities like raw electrocardiogram waveforms or echocardiographic imaging or longitudinal electronic health record trajectories, which may contain further predictive and explanatory signal.
- Retrospective, Non-Interventional Design: The evaluation is retrospective and non-interventional; the impact of the framework has not been evaluated in a live clinical workflow and its impact on actual clinician decision-making or patient outcomes has not been measured prospectively.
- Limited Fairness and Subgroup Analysis: This was evaluated at aggregate cohort level only.Performance and explanation quality evaluated at aggregate cohort level only. In the scope of the current study, there was not a

systematic evaluation of predictive parity and explanation consistency across demographic subgroups (such as age bands, sex), so there were potentially subgroup-level disparities that were not examined.

9.2 Future Work

- Future work should involve evaluating the remaining Heart Disease sub-databases and prospectively collected cardiovascular cohorts, both for multi-cohort validation and for testing external validity and longevity to test work in the face of real-world data drift.
- Raw physiological signals (e.g., ECG waveforms) or imaging-derived features, in addition to structured clinical attributes, could be integrated with the stacking approach, potentially via representation-learning encoders that operate on the existing stacking architecture, while maintaining the interpretability guarantees provided in this study.
- Prospective Clinical Integration Studies: The results of the offline explainability studies suggest that the framework has potential to be useful in a clinical decision-support setting, and deploying the framework in a simulated or real clinical decision-support environment, and measuring its impact on clinician diagnostic accuracy, triage efficiency, and trust, would validate this practical usefulness.
- Fairness-Aware Extensions: Future work should systematically evaluate predictive performance and/or SHAP fidelity and stability in demographic and clinical subgroups, while incorporating fairness-aware training or post-processing constraints when disparities are detected.
- If a meta-learner is able to provide calibrated uncertainty scores on its point-probability predictions, then the risk-stratification layer could flag cases with low confidence scores for mandatory clinician review without all predictions being made with the same level of confidence.
- The small-sample-size constraint and data-governance constraint could also be addressed in the same framework with the use of federated or distributed learning approaches that allow the framework to be trained across multiple hospital systems without centralizing sensitive patient records.
- The validation approach for the pipeline, measuring the fidelity and stability of the pipeline, in line with evolving regulatory requirements for AI-assisted clinical decision support software, would be important to take an initial step toward turning this research prototype into a deployed medical device.

Resolution of these drawbacks would advance the current prototype system, a validated research product, into a portable, widely applicable and clinically usable proactive cardiovascular risk-assessment system.

REFERENCES

[1] World Health Organization (WHO), Cardiovascular Diseases (CVDs): Key Facts, Geneva, Switzerland: WHO, 2023.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2016.

[3] M. U. Sreeja, A. O. Philip, and M. H. Supriya, "Towards Explainability in Artificial Intelligence Frameworks for Heart Care: A Comprehensive Survey," Journal of King Saud University – Computer and Information Sciences, 2024.

[4] C. J. Ejiyi et al., "XAI-HD: An Explainable Artificial Intelligence Framework for Heart Disease Detection," Artificial Intelligence Review, 2025. doi: 10.1007/s10462-025-11385-6.

[5] N. Pratheek, D. H. B. Yashas, et al., "Cardiovascular Disease Prediction with Machine Learning Algorithms and Interpretation using Explainable AI Methods: LIME & SHAP," in Proc. IEEE Int. Conf. on Intelligent Computing, Communication, Networking and Services (ICONAT), 2024.

[6] X. Wang et al., "Explainable AI-driven Machine Learning for Heart Disease Detection Using ECG Signal," Applied Soft Computing, vol. 158, Art. no. 112225, 2024.

[7] J. K. Nuamah, "Toward User-centered Explainable Displays for Complex Machine Learning Models in Healthcare: A Case Study of Heart Disease Prediction," Proc. Human Factors and Ergonomics Society Annual Meeting, 2024.

[8] HXAI-ML Authors, "HXAI-ML: A Hybrid Explainable Artificial Intelligence Based Machine Learning Model for Cardiovascular Heart Disease Detection," Results in Engineering, 2025.

[9] B. S. Ingole et al., "Advancements in Heart Disease Prediction: A Machine Learning Approach for Early Detection and Risk Assessment," arXiv preprint, 2024.

[10] A. A. Lamir et al., "A Comprehensive Machine Learning Framework for Heart Disease Prediction: Performance Evaluation and Future Perspectives," arXiv preprint, 2025.

[11] A. Kumar et al., "A Comprehensive Review of Machine Learning for Heart Disease Prediction: Challenges, Trends, Ethical Considerations and Future Directions," Artificial Intelligence Review, 2024.

[12] M. U. Sreeja, A. O. Philip, and M. H. Supriya, "Towards Explainability in Artificial Intelligence Frameworks for Heart Care: A Comprehensive Survey," Journal of King Saud University – Computer and Information Sciences, 2024.

[13] A. S. Albahri et al., "A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare," Information Fusion, vol. 94, pp. 1–25, 2023.

[14] A. Talukder et al., "Explainable AI for Clinical Decision Support in Cardiovascular Healthcare," Artificial Intelligence Review, 2025.

[15] E. Marcus and J. Teuwen, "Artificial Intelligence and Explanation: How, Why and When to Explain Black Boxes," European Journal of Radiology, 2024.

[16] S. Bharati, M. R. H. Mondal, and P. Podder, "A Review on Explainable Artificial Intelligence for Healthcare: Why, How and When?," 2023.

[17] S. Rababa et al., "Predicting Heart Disease and Reducing Survey Time Using Machine Learning Algorithms," arXiv preprint, 2023.

[18] Y. Ren et al., "Explainable Machine Learning for Early Heart Failure Prediction Among Acute Coronary Syndrome Patients," BMC Medical Informatics and Decision Making, vol. 22, 2022.

[19] A. Bahani et al., "Interpretable Machine Learning Using Fuzzy Rule-Based Clustering and Random Forest for Cardiovascular Disease Prediction," Expert Systems with Applications, vol. 205, 2022.

[20] L. Balabaeva et al., "Explainable Machine Learning Models for Arterial Hypertension Risk Assessment," Healthcare Analytics, vol. 2, 2022.

[21] L. Balabaeva et al., "Explainable Machine Learning Models for Arterial Hypertension Risk Assessment," Healthcare Analytics, vol. 2, 2022.

[23] K. Guleria et al., "XGBoost and AdaBoost Based Heart Disease Prediction Using Explainable Artificial Intelligence," 2022.

[24] S. Mohanty et al., "Patient Readmission Risk Prediction Using CatBoost, XGBoost and SHAP Explanations," Computers in Biology and Medicine, vol. 149, 2022.

[25] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2016.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2016.

[28] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

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

[30] D. Dua and C. Graff, UCI Machine Learning Repository: Heart Disease Dataset. Irvine, CA, USA: University of California, Irvine, School of Information and Computer Sciences. Available: https://archive.ics.uci.edu/ml

[31] Saboor, A., et al. (2022). *A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms*. Mobile Information Systems, 2022, Article 1410169.

[32] A. A. et al. (2021). *Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators*. *Applied Sciences, 11(18), 8352.*

[33] Alkhatib A. J., (2026). Integrated Computational Retinal Vascular Morphometry Using Fractal, Skeleton-Based, and Phi-Related Features for Diabetic Retinopathy and Glaucoma. International Journal of Artificial Intelligence, Machine Learning and Data Analytics, 1(1), 43-50. https://www.ijaimda.org/index.php/ijaimda/article/view/3

[34] S. Shah and S. Patel, "A comprehensive survey on fake news detection using machine learning," *Journal of Computer Science*, vol. 21, no. 4, pp. 982–990, 2025, doi: 10.3844/jcssp.2025.982.990.

[35] P. Shrivastava and S. Patel, "Selection of efficient and accurate prediction algorithm for employing real time 5G data load prediction," in *Proc. 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, 2021, pp. 572–580, doi: 10.1109/ICCCA52192.2021.9666235.

[36] Nannuri S., Gantla H. R., Ananya D., Vennela A., Bhuvana B., Neha G. (2026). Automated Brain Tumor Segmentationin Mri Via U-Net-Based Deep Neural Networks. International Journal of Engineering Sciences & Research Technology, 15(4), 1–8. https://doi.org/10.64149/j.ijesrt.15.4.1-8