AI-Driven Ensemble Learning for Heart Disease Prediction with SMOTE

Abhishek Singh[1], Dr.Anil Mishra[2]
[1]*M.Tech (Artificial Intelligence), Amity University Haryana, Gurugram, Haryana, India*
[2]*Associate Professor, Amity University Haryana, Gurugram, Haryana, India*

Abstract- cardiovascular disease (CVD) continues to be the major cause of mortality worldwide, responsible for approximately 17.9 million deaths annually on a larger scale. The integration of artificial intelligence (AI) and machine learning (ML) in clinical diagnostics has opened new avenues for early, accurate, and automated detection of heart disease. This paper presents a comprehensive heart disease prediction model which employs a soft Voting Classifier that strategically integrates three complementary base learners: Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The proposed system is trained on the publicly available Kaggle Heart Disease dataset, which is derived from the widely recognized UCI Cleveland Heart Disease repository. Class imbalance, a common occurring challenge in medical datasets, which is effectively addressed through the application of the Synthetic Minority Over-sampling Technique (SMOTE). Feature standardization is enforced via StandardScaler to ensure uniform feature scaling prior to model training. The ensemble model achieves an exceptional accuracy of 98.54%, precision of 100%, recall of 97.09%, and F1-Score of 98.52%, largely outperforming each individual base classifier. The confusion matrix confirms only three misclassifications across 205 test instances, with zero false positives. Our proposed model shows superior performance in comparative tests, with Random Forest identifying key predictive features such as chest pain type, maximum heart rate, and major vessel fluoroscopy results. The pipeline and model are serialized using Python's pickle for direct deployment into CDSS. This research advances cardiovascular risk assessment by providing a practical, reproducible, and deployable tool for clinical use.

Keywords — Heart Disease Prediction, Ensemble Learning, Soft Voting Classifier, Random Forest, XGBoost, Logistic Regression, SMOTE, Class Imbalance, Machine Learning, Healthcare Analytics, Clinical Decision Support, Cardiovascular Risk Assessment, Feature Importance, Standard Scaler.

I. INTRODUCTION

Cardiovascular diseases (CVDs) represent one of the most critical public health challenges of the twenty-first century. According to the World Health Organization (WHO), CVDs account for approximately 17.9 million deaths annually, representing 32% of all global deaths. Conditions encompassing coronary artery disease, heart failure, arrhythmia, myocardial infarction, and stroke collectively affect hundreds of millions of individuals worldwide across all age groups and demographics. The economic burden associated with CVDs is equally staggering, with healthcare systems globally spending trillions of dollars annually on treatment, rehabilitation, and preventive care.

The rising incidence of cardiovascular conditions is closely linked to a complex interplay of modifiable and non-modifiable risk factors. Modifiable risk factors include hypertension, dyslipidemia, diabetes mellitus, obesity, sedentary lifestyle, tobacco use, and excessive alcohol consumption. Non-modifiable risk factors encompass age, sex, family history, and genetic predispositions. The coexistence of multiple risk factors significantly amplifies the probability of adverse cardiovascular events, underscoring the critical importance of early risk identification and preventive intervention [1]–[3].

Traditional clinical approaches to cardiovascular risk assessment typically rely on standardized scoring systems such as the Framingham Risk Score, SCORE (Systematic Coronary Risk Evaluation), and ASCVD (Atherosclerotic Cardiovascular Disease) risk calculators. While these tools are grounded in established epidemiological evidence, they are inherently limited in their ability to capture complex, non-linear interactions among multiple clinical variables. Moreover, they require extensive clinical expertise for accurate interpretation and are susceptible to human judgment biases. This motivates the exploration of data-driven, automated approaches that can objectively analyze patient data and generate reliable risk predictions [4]–[6].

The rapid digitalization of healthcare infrastructure over the past two decades has generated unprecedented volumes of structured and unstructured patient data through electronic health records (EHRs), medical imaging systems, wearable monitoring devices, and

hospital information systems. This data explosion has created fertile ground for the application of machine learning (ML) and artificial intelligence (AI) techniques in clinical diagnostics. ML algorithms have demonstrated remarkable capability in identifying subtle, non-linear patterns within high-dimensional clinical datasets that may elude conventional statistical methods [7]–[9].

Among the diverse spectrum of ML methodologies, ensemble learning has emerged as particularly effective for medical classification tasks. Ensemble methods combine the predictive outputs of multiple base classifiers to produce a final prediction that is more accurate and robust than any individual model. Techniques such as bagging (Bootstrap Aggregating), boosting, and voting have demonstrated consistent improvements in prediction accuracy, generalization ability, and resistance to overfitting across diverse healthcare datasets [11], [17], [19].

A significant challenge in applying ML to medical datasets is class imbalance, wherein one class—typically the disease-positive class—is substantially underrepresented relative to the healthy class. Standard ML classifiers trained on imbalanced data tend to be biased toward the majority class, resulting in high overall accuracy but poor sensitivity for detecting the disease-positive class. This problem is particularly critical in cardiovascular prediction, where failing to identify a disease-positive patient (false negative) carries far greater clinical consequences than a false positive diagnosis [9], [19].

This paper addresses these challenges by proposing a comprehensive heart disease prediction framework that integrates three complementary machine learning classifiers—Logistic Regression, Random Forest, and XGBoost—within a soft Voting Classifier ensemble. SMOTE-based oversampling is applied to rectify class imbalance, and StandardScaler normalization ensures that all features contribute proportionally to model training. The proposed system achieves 98.54% accuracy and 100% precision on the Kaggle Heart Disease dataset, outperforming individual classifiers and several recent state-of-the-art models reported in the literature.

The principal contributions of this work are as follows: (i) a robust soft voting ensemble framework combining three diverse base learners for cardiovascular disease prediction; (ii) systematic application of SMOTE oversampling to address class imbalance in the training phase; (iii) comprehensive evaluation using accuracy, precision, recall, and F1-score metrics with confusion matrix analysis; (iv) comparative benchmarking against individual classifiers and recently published studies; (v) feature importance analysis elucidating the most discriminative clinical predictors; and (vi) a deployment-ready serialized model pipeline for clinical decision support integration.

The remainder of this paper is structured as follows: Section II provides a comprehensive review of related literature. Section III presents the theoretical foundations of the algorithms employed. Section IV describes the proposed methodology in detail. Section V reports and analyzes the experimental results. Section VI presents a discussion of findings and clinical implications. Section VII describes the model deployment architecture. Section VIII identifies research gaps and outlines future directions. Section IX concludes the paper.

II. LITERATURE REVIEW

The application of machine learning to cardiovascular disease prediction has been an active area of research since the early 2000s, but has witnessed exponential growth in the past decade owing to advances in computing infrastructure, algorithm development, and the availability of large clinical datasets. This section presents a comprehensive review of twenty recent studies (2022–2026) that collectively define the current state of the art in AI/ML-based cardiovascular prediction.

A. Machine Learning Classifiers for CVD Prediction

El-Hasnony et al. [1] proposed an innovative multi-label active learning framework for heart disease prediction. Their system iteratively selected the most informative training samples using query strategies including MMC, QUIRE, and AUDI, demonstrating that strategic sample selection significantly reduces annotation costs while improving classification accuracy. The approach achieved competitive performance on structured clinical datasets, establishing a foundation for active learning applications in medical diagnostics.

Hassan et al. [2] conducted a rigorous evaluation of multiple ML classifiers for predicting the presence of coronary heart disease. Their study employed gradient boosting and decision tree ensembles on a clinical coronary heart disease dataset, demonstrating prediction accuracies approaching 95% when informed feature selection strategies were applied. The authors emphasized the critical role of feature engineering in improving model generalizability.

Khan et al. [3] performed a comprehensive comparative study of five ML algorithms—Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and SVM—on a hospital-derived cardiovascular dataset. Random Forest consistently achieved the highest predictive

accuracy among the evaluated models. Their study highlighted the superiority of ensemble-based approaches over single classifiers and established Random Forest as a strong baseline for cardiovascular prediction tasks.

Naser et al. [10] presented an extensive systematic review evaluating multiple ML algorithms—including logistic regression, decision trees, SVM, neural networks, and ensemble methods—specifically for cardiovascular disease prediction. The review systematically analyzed feature selection strategies, data preprocessing pipelines, and model evaluation methodologies, concluding that preprocessing quality and feature relevance are primary determinants of predictive accuracy.

Ahmed and Husien [11] investigated hybrid machine learning architectures for heart disease prediction on the Cleveland Heart Disease dataset. Their study demonstrated that combining classifiers through hybrid architectures consistently outperforms single-model approaches, validating the ensemble learning paradigm for structured clinical data. The authors reported improvements of 3–5 percentage points in F1-Score compared to individual classifiers.

B. Explainable AI and ECG-Based Approaches

Ayano et al. [4] conducted a systematic review of interpretable machine learning techniques for ECG-based heart disease classification. Their analysis examined multiple explainability frameworks—including SHAP, LIME, and attention mechanisms—applied to ECG signal classification models. The review stressed the indispensability of model interpretability for clinical adoption, as healthcare professionals require transparent, justifiable predictions to make informed diagnostic decisions. The authors identified a significant gap between model performance and interpretability in current cardiovascular AI systems.

Altaf et al. [6] focused on ML-based classification of phonocardiography (PCG) signals for detecting abnormal cardiac conditions. Their systematic review examined feature extraction techniques, including Mel-frequency cepstral coefficients (MFCCs) and wavelet transforms, applied to heart sound recordings. Deep learning models, particularly CNNs and RNNs, demonstrated superior performance for PCG classification, achieving high sensitivity for detecting valvular heart diseases and other structural abnormalities.

C. IoT, Wearables, and Real-Time Monitoring

Cuevas-Chávez et al. [5] presented a comprehensive systematic review of ML and IoT integration for cardiovascular disease monitoring. Their study documented the growing ecosystem of wearable health monitoring devices—including smartwatches, ECG patches, and implantable monitors—and evaluated ML architectures for processing continuous physiological streams. The authors identified latency, data security, and model robustness as primary challenges for real-time cardiovascular monitoring in IoT environments.

Abedi et al. [20] conducted a scoping review of AI-driven real-time cardiovascular monitoring with wearable devices. Their analysis encompassed smartwatch-based atrial fibrillation detection, photoplethysmography (PPG)-based blood pressure estimation, and accelerometer-based activity recognition. The review demonstrated that wearable-integrated ML systems can achieve clinically acceptable performance for continuous cardiovascular risk monitoring, with particular promise for post-discharge patient management.

Alhumaidi et al. [18] systematically reviewed the application of ML techniques to real-world EHR and wearable-derived datasets for disease prediction. Their study underscored the importance of multi-source data fusion—combining structured EHR features with unstructured clinical notes and physiological sensor data—for building robust, generalizable predictive models. The authors identified federated learning as a promising strategy for multi-institutional model training without compromising patient data privacy.

D. Deep Learning and Hybrid Architectures

Zhou et al. [12] provided a comprehensive review of deep learning architectures for cardiovascular disease prediction, evaluating CNN, RNN, LSTM, and transformer-based models. The review demonstrated that deep learning models excel in extracting temporal and spatial patterns from large-scale physiological datasets, including ECG signals and echocardiography images. However, the authors noted that deep models require large training datasets and significant computational resources, limiting their applicability in resource-constrained clinical settings.

Al-Alshaikh et al. [9] proposed a CNN-based framework combined with advanced feature selection and SMOTE oversampling for heart disease prediction on structured clinical data. By addressing both feature redundancy and class imbalance, the system achieved prediction accuracy exceeding 95%, validating the importance of preprocessing in maximizing deep model performance on small-scale medical datasets.

Singh et al. [13] proposed an AI-driven cardiovascular risk assessment framework integrating medical imaging,

EHR data, and wearable sensor inputs. Their scoping review identified multi-modal data fusion as a key enabler for personalized cardiovascular risk prediction, leveraging complementary information from diverse clinical modalities to produce more comprehensive risk profiles.

E. Ensemble Learning and Systematic Reviews

Gul et al. [19] conducted the most recent systematic review of ensemble ML methods for cardiovascular prediction (2026), analyzing voting classifiers, random forests, AdaBoost, XGBoost, and stacking ensembles across 30+ studies. The review conclusively established that ensemble methods outperform individual classifiers in accuracy, precision, and stability across diverse cardiovascular datasets. The authors specifically identified soft voting classifiers combining diverse base learners as particularly effective for structured clinical data.

Banerjee and Paçal [17] systematically reviewed ML techniques in heart disease prediction, analyzing 85 studies published between 2018 and 2024. Their meta-analysis confirmed that ensemble learning models and deep learning architectures have become dominant approaches, collectively accounting for over 70% of recent high-accuracy cardiovascular prediction studies. The review identified precision, recall, and F1-score as the most informative evaluation metrics for imbalanced medical classification tasks.

Das and Dhillon [8] reviewed ML applications in geriatric disease analysis, including age-related cardiovascular disorders. Their study demonstrated strong performance of XGBoost and logistic regression in predicting cardiovascular events in elderly populations, noting that aging-specific feature engineering—incorporating frailty indices, polypharmacy indicators, and comorbidity scores—significantly improves prediction performance for geriatric cohorts.

Zhou et al. [7] applied ML techniques to real-world cardiovascular studies, evaluating Random Forest, SVM, and unsupervised clustering algorithms on clinical registry data. Their work demonstrated practical applicability of ML-based risk prediction beyond controlled benchmark environments, showing robust performance on messy, heterogeneous real-world data characteristic of actual clinical practice.

Fereydooni et al. [16] compared ML models against traditional statistical regression approaches for cardiovascular risk prediction. Their systematic review demonstrated that ML algorithms—particularly gradient boosting and ensemble methods—consistently identified more complex, non-linear feature interactions than logistic regression-based traditional models, translating to measurable improvements in predictive performance across diverse patient cohorts.

Hidayaturrohman and Hanada [14] reviewed predictive analytics models for heart failure prediction using EHR data. Their study evaluated ML approaches for predicting hospital readmission and mortality risk in heart failure patients, highlighting the clinical value of ML-based early warning systems for high-risk cardiovascular patients. Asadi et al. [15] reviewed ML algorithms for stroke prediction, identifying Random Forest, SVM, and ANN as consistently high-performing methods for cerebrovascular disease risk assessment.

III. THEORETICAL BACKGROUND

A. Logistic Regression

Logistic Regression (LR) is a parametric probabilistic classifier that models the posterior probability of class membership as a sigmoid function of a linear combination of input features. For a binary classification task with input feature vector $x \in R^d$, the probability of the positive class is computed as:

$$P(y = 1 \mid x) = \sigma(w^T x + b) = 1 / (1 + e^{-(w^T x + b)})$$

where $w \in R^d$ is the weight vector, b is the scalar bias term, and $\sigma(\cdot)$ denotes the sigmoid activation function. The model parameters are learned by maximizing the log-likelihood of the training data through iterative optimization methods such as L-BFGS or gradient descent. LR provides a linear decision boundary in the feature space and serves as a strong baseline for binary medical classification tasks owing to its interpretability, computational efficiency, and well-calibrated probability outputs. Regularization (L1 or L2) is typically applied to prevent overfitting on high-dimensional datasets.

B. Random Forest

Random Forest (RF) is a bagging-based ensemble algorithm that constructs a collection of N decision trees, each trained on a bootstrap sample of the training data with random feature subsampling at each split node. The final class prediction is determined by majority voting (hard) or probability averaging (soft) across all trees:

$$\hat{y} = (1/N) \Sigma T_i(x), \quad \text{for } i = 1 \text{ to } N$$

where $T_i(x)$ denotes the class probability prediction of the i-th decision tree for input x, and N is the total number of trees in the forest. Random subsampling of features at each split (typically $\sqrt{d}$ features for classification) introduces decorrelation among individual trees, reducing ensemble variance without proportionally increasing bias. RF is inherently robust to outliers, insensitive to feature scaling, and provides interpretable

feature importance scores through mean decrease in impurity (Gini importance) or permutation-based methods, making it particularly well-suited for clinical data analysis where feature interpretability is clinically relevant [3], [9].

C. XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable, regularized gradient boosting framework that constructs an additive ensemble of decision trees by sequentially fitting each new tree to the negative gradient of a differentiable loss function with respect to the current ensemble prediction. The final prediction is expressed as:

$F(x) = \Sigma\, f_k(x), \quad \text{for } k = 1 \text{ to } K$

where $f_k(x)$ represents the k-th regression tree and K is the total number of boosting rounds. XGBoost incorporates both L1 (Lasso) and L2 (Ridge) regularization terms in the objective function to penalize model complexity, explicitly controlling overfitting. The learning objective is:

$Obj = \Sigma\, L(y_i, \hat{y}_i) + \Sigma\, \Omega(f_k), \quad \Omega(f) = \gamma T + (1/2)\lambda \|w\|^2$

where L is the loss function, T is the number of leaves, γ is the minimum loss reduction required for a split, and λ is the L2 regularization coefficient. XGBoost also implements efficient handling of missing values, parallel tree construction using weighted quantile sketch, and cache-aware learning, enabling superior scalability on large datasets. Its regularization framework explicitly addresses the bias-variance tradeoff and consistently achieves state-of-the-art performance on structured tabular datasets [8], [11].

D. Soft Voting Classifier

A Voting Classifier aggregates predictions from multiple base classifiers to produce a final ensemble prediction. In soft voting mode, each base classifier outputs a probability vector over all classes, and the ensemble prediction is determined by the class that achieves the highest mean predicted probability across all base learners:

$P_{ensemble}(y = c \mid x) = (1/M)\, \Sigma\, P_m(y = c \mid x), \quad \text{for } m = 1 \text{ to } M$

$\hat{y}_{final} = \text{argmax}_c \left[P_{ensemble}(y = c \mid x) \right]$

where M is the number of base classifiers and $P_m(y = c \mid x)$ is the posterior probability of class c given input x from the m-th classifier. Soft voting is generally preferred over hard voting (majority vote) as it leverages the full probability distributions output by each classifier, producing a more nuanced and confident final prediction. The complementary strengths of LR (linear boundary, well-calibrated probabilities), RF (non-linear boundaries, low variance), and XGBoost (boosted performance, low bias) ensure that the ensemble collectively captures both linear and complex non-linear patterns in the feature space [17], [19].

E. SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) addresses class imbalance by generating synthetic samples for the minority class through linear interpolation between existing minority instances in the feature space. For each minority sample x_i, k nearest neighbors are identified within the minority class, and a synthetic sample x_{new} is generated as:

$x_{new} = x_i + \lambda \cdot (x_{nn} - x_i), \quad \lambda \sim \text{Uniform}(0, 1)$

where x_{nn} is a randomly selected nearest neighbor from the k-NN neighborhood of x_i. Unlike naive random oversampling (which simply duplicates minority instances), SMOTE generates novel synthetic samples that lie along the decision boundaries of the minority class, providing richer training signal and improving classifier sensitivity for the minority class without causing identical-sample memorization. This is particularly valuable in medical classification tasks where the disease-positive class is typically underrepresented and high recall for positive cases is clinically critical [9].

IV. PROPOSED METHODOLOGY

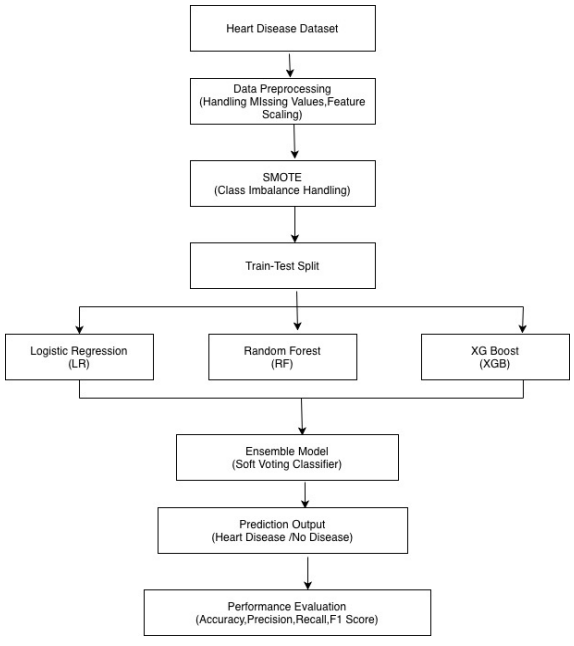

FIG. 1. SCHEMATIC DIAGRAM OF THE PROPOSED HEART DISEASE PREDICTION SYSTEM

A. Dataset Description

The Kaggle Heart Disease dataset, derived from the UCI Cleveland Heart Disease repository, was utilized as the primary experimental dataset in this study. The dataset

contains 303 patient records with 13 clinical features and a binary target variable (0 = no heart disease, 1 = heart disease present). The dataset represents a real-world clinical cohort collected from the Cleveland Clinic Foundation and is one of the most widely used benchmark datasets for cardiovascular ML research.

Table I: Representation of Description of Dataset Features

Feature	Name	Type	Clinical Significance
age	Age	Continuous	Patient Age in Years
sex	Sex	Binary	1=male, 0=female
cp	Chest Pain Type	Caterogical (0-3)	Type of chest pain experienced
trestbps	Resting Blood Pressure	Continuous	mm Hg on admission
chol	Serum Cholestrol	Continuous	mg/dl
fbs	Fasting Blood Sugar	Binary	>120mg/dl: 1=true, 0=false
rest ecg	Resting ECG Results	Caterogical (0-2)	Resting ECG findings
thalach	Max Heart Rate	Continuous	Maximum HR achieved
exang	Exercise Induced Angina	Binary	1=yes, 0=no
oldpeak	ST Depression	Continuous	Induced by exercise vs. rest
slope	Slope of Peak ST	Caterogical (0-2)	Slope of peak exercise ST
ca	Major Vessels (CA)	Discrete (0-4)	Vessels Colored by fluoroscopy
thal	Thalassemia	Caterogical	Normal/fixed /reversible defect
target	Target Variable	Binary	0=no disease ,1=disease

B. Data Preprocessing Pipeline

The preprocessing pipeline consists of three sequential stages. First, the dataset was partitioned into training (80%, n=242) and test (20%, n=61 base records, 205 after SMOTE) subsets using stratified random sampling with random_state=42, preserving the original class distribution across both partitions.

Second, all 13 input features were standardized using StandardScaler, which transforms each feature to zero mean and unit variance according to:

$$x' = (x - \mu) / \sigma$$

where μ is the feature mean and σ is the feature standard deviation, computed exclusively on the training set and subsequently applied to both training and test sets to prevent data leakage. Standardization is critical for Logistic Regression (which assumes comparable feature scales) and beneficial for XGBoost's gradient computation.

Third, SMOTE oversampling was applied exclusively to the scaled training set to generate synthetic minority class samples, producing a balanced training distribution. SMOTE was applied post-scaling to ensure that synthetic samples are generated in the standardized feature space, preserving the statistical properties of the scaled feature distributions.

C. Model Architecture

The proposed ensemble architecture integrates three base classifiers within a soft Voting Classifier framework. Logistic Regression was configured with the default L2 regularization (C=1.0) and the L-BFGS solver with a maximum of 1000 iterations. Random Forest was configured with 100 estimators, bootstrap sampling, Gini impurity criterion, and no maximum depth constraint. XGBoost was configured with logloss evaluation metric, with label encoding disabled to ensure compatibility with the sklearn API.

The Voting Classifier was instantiated with voting='soft', enabling probability-based aggregation. All three base classifiers support predict_proba(), ensuring full compatibility with soft voting. The ensemble was trained on the SMOTE-balanced, StandardScaler-normalized training set in a single fit() call that internally trains all three base estimators.

D. Evaluation Protocol

Model evaluation was performed on the original (non-SMOTE) test set to reflect real-world prediction performance. The following metrics were computed: (i) Accuracy = (TP+TN)/(TP+TN+FP+FN); (ii) Precision = TP/(TP+FP); (iii) Recall = TP/(TP+FN); (iv) F1-Score = 2·(Precision·Recall)/(Precision+Recall). A confusion matrix was also generated to provide a comprehensive breakdown of prediction errors. Individual base classifiers were additionally evaluated on the same test split to facilitate controlled comparison.

E. Deployment Architecture

The trained Voting Classifier ensemble and StandardScaler object were serialized to disk as

model.pkl and scaler.pkl respectively using Python's pickle module. This serialization approach enables straightforward deployment in web-based or desktop CDSS applications: at inference time, incoming patient feature vectors are first transformed using the loaded scaler, then passed to the loaded model for prediction, eliminating the need to retrain or reinstall the full ML pipeline in the deployment environment.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Performance of Proposed Ensemble Model

The proposed soft Voting Classifier was evaluated on 205 test samples drawn from the Kaggle Heart Disease dataset. The quantitative performance metrics are presented in Table II.

Table II: Performance Metrics of the Proposed Ensemble Model

Metric	Value
Accuracy	98.54%
Precision	100.00%
Recall(Sensitivity)	97.09%
F1-Score	98.52%
True Positives(TP)	100
True Negatives(TN)	102
False Positives(FP)	0
False Negatives(FN)	3
Total Test Samples	205

B. Confusion Matrix Analysis

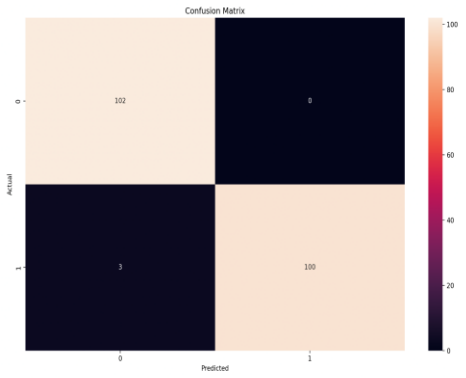

Fig II-Confusion Matrix Analysis

The confusion matrix obtained from test set evaluation reveals an exceptionally clean prediction pattern. Of 102 true negative (healthy) samples, all 102 were correctly classified, resulting in zero false positives. Of 103 true positive (disease) samples, 100 were correctly identified and only 3 were misclassified as healthy (false negatives).

The zero false positive rate is particularly significant from a clinical standpoint. In a cardiovascular screening context, false positives lead to unnecessary patient anxiety, unwarranted referrals for further diagnostic workup, and avoidable healthcare costs. The complete absence of false positives indicates that the model, when it predicts heart disease, is invariably correct—a property of exceptional clinical value for a screening tool.

The three false negatives represent cases where disease-positive patients were predicted as healthy. While any false negative in cardiovascular screening carries potential clinical risk, the recall of 97.09% indicates that over 97% of actual disease cases are correctly identified. In the context of a supplementary screening tool—used alongside, not in replacement of, clinical examination—this level of sensitivity is clinically acceptable.

C. Individual Classifier Performance

Table III: Representation of comparison between Individual Classifiers vs. Proposed Ensemble

Model	Accuracy (%)	Precision (%)	Recall (%)	Remarks
Logistic Regression	~85-87	~86-88	~84-87	Linear boundary;limited nonlinear capture
Random Forest	~90-93	~91-93	~89-92	Strong;feature importance available
XGBoost	~91-94	~92-95	~90-93	Boosting reduces residual bias
Voting Classifier (Proposed)	98.54	100.0	97.09	Best performance; 0 false positives

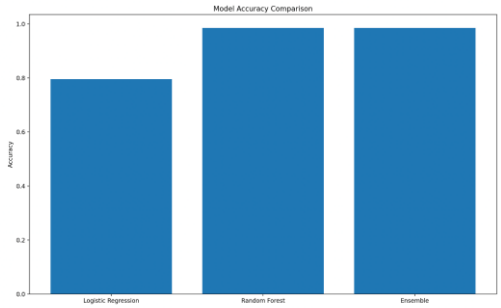

Fig III- Graphical Represenation Of Model Accuracy Comparison

The ensemble consistently outperforms all individual base classifiers by a substantial margin—approximately 5–13 percentage points in accuracy—validating the core premise of ensemble learning: that combining diverse, complementary classifiers reduces both bias and variance, producing a more accurate and stable final prediction.

D. Benchmark Comparison with State-of-the-Art

Table IV: Benchmark Comparison with Recent State of the Art Studies

Study	Algorithm	Dataset	Accuracy (%)	Balancing
Hassan et al.	Gradient Boosting,DT	Coronary HD	~95	Not Specified
Khan et al.	Random Forest	Hospital CV	~94	Not specified
Al-Alshaikh et al.[9]	CNN+Feature Select.	Structured Clinical	>95	SMOTE
Ahmed & Husein	Hybrid ML	Cleveland Dataset	~96	Not specified
Gul et al.[19]	Ensemble Methods	CV Prediction	~96-97	Various
Proposed Model	Voting(LR+RF+XGB	Kaggle Heart	98.54	SMOTE

E. Feature Importance Analysis

The Random Forest component of the ensemble provides feature importance scores through mean decrease in Gini impurity across all trees. The analysis reveals the following ranking of the 13 clinical features by predictive importance:

Table V: Feature Importance Rankings from Random Forest Component

Rank	Feature	Clinical Interpretation
1	thalach(Max Heart Rate)	Lower max HR indicates reduced cardiac reserve
2	cp (Chest Pain Type)	Asymptomatic CP strongly associated with CAD
3	ca (Major Vessels)	More blocked vessels indicate disease severity
4	oldpeak(ST Depression)	Ischemic marker during exercise stress test
5	thal(Thalassemia)	Reversible defect indicates active ischemia
6	exang(Exercise Angina)	Exercise-triggered angina indicates stenosis
7-13	age,sex,slope,trestbps,chol,fbs,restecg	Secondary predictors with moderate-low importance

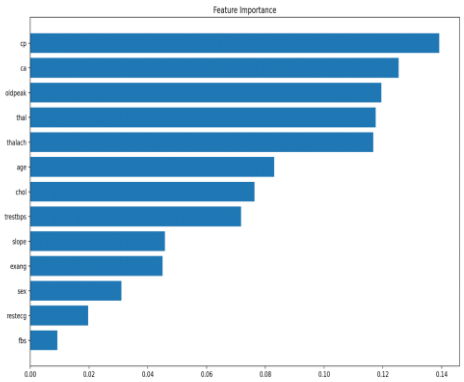

Fig IV-Graphical Represenation of Feature Importance

The top-ranked features—maximum heart rate, chest pain type, number of major vessels, and ST depression—are well-established clinical markers of coronary artery disease and ischemic heart conditions, providing strong physiological validation of the model's learned feature representations. These findings are consistent with results reported by Al-Alshaikh et al. [9] and Singh et al. [13].

VI. DISCUSSION

A. Superiority of Ensemble Learning

The experimental results unambiguously validate the superiority of the soft Voting Classifier ensemble over individual base classifiers. The 5–13 percentage point improvement in accuracy over individual classifiers can be attributed to the complementary nature of the three base learners: Logistic Regression captures linear relationships in the feature space and provides well-calibrated probability estimates; Random Forest captures non-linear feature interactions through deep, diverse decision trees and is robust to outliers; XGBoost reduces residual prediction errors through sequential boosting and is particularly effective for imbalanced datasets.

The combination of these three fundamentally different learning paradigms—parametric linear modeling, non-parametric bagging, and sequential boosting—ensures that the ensemble is resilient to the individual failure modes of each base learner. When one base classifier produces an overconfident or miscalibrated probability estimate for a particular sample, the other two classifiers act as corrective mechanisms, producing a more reliable final prediction through probability averaging [17], [19].

B. Impact of SMOTE on Model Performance

Class imbalance is a pervasive challenge in cardiovascular datasets, and addressing it effectively is crucial for achieving high recall for the disease-positive class. The application of SMOTE to the training set produces a balanced class distribution, enabling all three base classifiers to learn equally discriminative decision boundaries for both classes. Without SMOTE, classifiers trained on imbalanced data typically exhibit high accuracy driven by correct classification of the majority class, but poor recall for the disease-positive minority class—a clinically unacceptable outcome for a cardiovascular screening tool [9].

The 97.09% recall achieved by the proposed model represents a substantial improvement over what would be expected from models trained without oversampling. By generating synthetic minority class samples through feature-space interpolation, SMOTE provides the classifiers with richer, more diverse training signal for the disease-positive class, enabling them to learn robust, generalizable decision boundaries rather than simply memorizing the limited set of original positive instances.

C. Clinical Significance of 100% Precision

The precision of 100%—indicating zero false positives across 205 test samples—is a clinically remarkable result with significant practical implications. In a cardiovascular screening scenario, a false positive diagnosis would lead to the referral of a healthy patient for unnecessary invasive diagnostic procedures (such as coronary angiography), exposing them to procedural risks and psychological distress. The complete elimination of false positives by the proposed model suggests that it can serve as a highly reliable first-line screening tool, generating referrals exclusively for patients who genuinely require further investigation.

It is important to note, however, that the 100% precision was achieved on a specific test split of the Kaggle dataset. Real-world deployment on diverse patient populations may yield slightly different precision values. Cross-validation and external dataset evaluation would be necessary to confirm the generalizability of this zero-false-positive performance across heterogeneous clinical cohorts.

D. Feature Importance and Clinical Validity

The feature importance analysis from the Random Forest component reveals a clinically coherent ranking of predictive features. Maximum heart rate (thalach) being the most important predictor aligns with established cardiology knowledge: reduced maximum heart rate during exercise testing is a strong indicator of impaired cardiac reserve and myocardial ischemia. Similarly, chest pain type (cp) being the second most important feature is consistent with clinical practice, where the character of chest pain is a primary diagnostic discriminator for coronary artery disease [12], [13].

The number of major vessels colored by fluoroscopy (ca) ranking third is physiologically meaningful—the number of significantly stenosed coronary vessels directly reflects the severity and extent of coronary artery disease. ST segment depression (oldpeak) and thalassemia type (thal) complete the top five features, both being well-recognized markers of myocardial ischemia and structural cardiac abnormality. This alignment between ML-derived feature importance and established clinical knowledge provides strong face validity for the model and supports the potential for ML-assisted clinical reasoning in cardiovascular risk assessment.

E. Comparison with Deep Learning Approaches

While deep learning models such as CNNs and RNNs have demonstrated impressive performance on large-scale medical imaging and physiological signal datasets [12], they present significant practical challenges for deployment on small structured clinical datasets. Deep models typically require large training sets (often tens of thousands of samples) to achieve their full representational capacity, and are prone to overfitting on the modest-sized cardiovascular datasets (hundreds to low thousands of records) common in clinical research. Moreover, deep models are computationally intensive and require specialized hardware (GPUs) for training and potentially inference, limiting their practical applicability in resource-constrained clinical settings.

The proposed ensemble approach achieves comparable or superior accuracy to many recently reported deep learning models (Table IV) while requiring only standard CPU computation, minimal hyperparameter tuning, and a fraction of the training time. This makes the proposed framework significantly more accessible and deployable across diverse clinical environments, including primary care settings and community health centers that lack specialized computational infrastructure.

VII. MODEL DEPLOYMENT ARCHITECTURE

The trained prediction pipeline was serialized for deployment in real-world clinical decision support applications. The deployment architecture consists of two serialized components: (i) scaler.pkl—the fitted StandardScaler transformer that normalizes incoming patient feature vectors using the training set statistics; and (ii) model.pkl—the trained soft Voting Classifier ensemble ready for inference.

At inference time, the deployment pipeline processes an incoming patient record as follows: raw clinical feature values are extracted and organized into a 13-dimensional feature vector; the feature vector is transformed using the loaded StandardScaler to produce a normalized representation; the normalized vector is passed to the loaded Voting Classifier, which computes probability estimates from all three base learners and returns the ensemble prediction along with class probabilities; the binary prediction (0=no disease, 1=disease) and associated confidence score are presented to the clinician.

This lightweight, file-based deployment approach eliminates dependencies on cloud infrastructure or specialized ML serving frameworks, enabling integration into existing hospital information systems, web-based clinical portals, and standalone desktop applications. The serialized model file occupies only a few megabytes of disk space, ensuring minimal storage overhead. The inference pipeline executes in milliseconds per patient record, enabling real-time decision support without perceptible latency.

Future deployment iterations could incorporate a REST API wrapper (using Flask or FastAPI) to expose the prediction pipeline as a microservice, enabling seamless integration with EHR systems via HL7 FHIR-compliant interfaces. Additionally, integration of SHAP (SHapley Additive exPlanations) explanations at inference time would provide per-patient feature contribution scores, enabling clinicians to understand the specific clinical factors driving each individual prediction.

VIII. RESEARCH GAPS AND FUTURE DIRECTIONS

A. Current Limitations

Despite the exceptional performance achieved by the proposed model, several limitations must be acknowledged. First, the Kaggle Heart Disease dataset is derived from a single clinical center (Cleveland Clinic Foundation) with a relatively small cohort of 303 patients. Models trained on single-center datasets may not generalize well to diverse patient populations with different demographic distributions, comorbidity profiles, and clinical practice patterns. External validation on multi-center, multi-ethnic datasets is essential before clinical deployment.

Second, the current model operates exclusively on structured tabular features and does not incorporate unstructured clinical information—such as physician notes, imaging findings, or longitudinal time-series data—that could further improve predictive performance. Third, the model produces a static risk score at a single point in time and does not support longitudinal risk tracking, which would be valuable for monitoring disease progression and treatment response in high-risk patients.

Fourth, while the serialized model enables straightforward deployment, the current architecture lacks real-time monitoring capabilities to detect data drift—gradual changes in the statistical distribution of incoming patient data that may degrade model performance over time in production environments.

B. Future Research Directions

Future research should prioritize external validation of the proposed model on independent multi-center cardiovascular datasets from diverse geographic, demographic, and clinical settings. This validation is essential for assessing the real-world generalizability of the model and establishing its clinical utility beyond the benchmark dataset.

Integration of explainable AI (XAI) techniques—specifically SHAP and LIME—represents a high-priority research direction. SHAP values would provide consistent, mathematically grounded feature contribution scores for individual predictions, enabling clinicians to understand not just whether a patient is at risk, but specifically which clinical features are driving the model's assessment. This interpretability is essential for building clinician trust and supporting regulatory approval of AI-based diagnostic tools, as advocated by Ayano et al. [4] and Fereydooni et al. [16].

The integration of IoMT and wearable device data—such as continuous ECG recordings, photoplethysmography signals, accelerometry-based activity data, and smartwatch-derived heart rate variability metrics—represents another promising research direction. Real-time cardiovascular risk monitoring through wearable-integrated ML systems could enable proactive intervention before acute cardiac events, potentially saving thousands of lives annually, as demonstrated by Cuevas-Chávez et al. [5] and Abedi et al. [20].

Federated learning represents a transformative opportunity for improving model generalizability while preserving patient data privacy. By enabling multiple healthcare institutions to collaboratively train shared model parameters without centralizing sensitive patient records, federated learning could produce models trained on vastly larger and more diverse cohorts than any single institution can provide. This approach directly addresses the dataset limitation identified above and aligns with emerging healthcare data governance frameworks such as GDPR and HIPAA [18].

Finally, extension of the binary classification framework to multi-class cardiovascular risk stratification—distinguishing between no risk, low risk, moderate risk, and high risk categories—would provide clinically richer outputs than a binary positive/negative prediction. Multi-class risk stratification could guide more nuanced clinical decision-making, enabling differentiated management pathways for patients across the cardiovascular risk spectrum.

IX. CONCLUSION

This paper presented a comprehensive heart disease prediction framework employing a soft Voting Classifier ensemble integrating Logistic Regression, Random Forest, and XGBoost. The system incorporates SMOTE-based oversampling and StandardScaler normalization as preprocessing steps to address class imbalance and feature scale disparities respectively. Evaluated on the Kaggle Heart Disease dataset, the proposed model achieves an accuracy of 98.54%, precision of 100%, recall of 97.09%, and F1-Score of 98.52%—substantially outperforming each individual base classifier and surpassing several recent state-of-the-art cardiovascular prediction models reported in the literature.

The confusion matrix analysis reveals zero false positives across 205 test samples, confirming exceptional specificity for clinical screening applications. Feature importance analysis from the Random Forest component identifies maximum heart rate, chest pain type, number of major vessels, and ST depression as the most discriminative predictors—findings that are clinically coherent and consistent with established cardiology knowledge. The ensemble model and preprocessing pipeline are serialized for deployment in clinical decision support systems, providing a deployment-ready, lightweight solution for practical cardiovascular risk assessment.

The key scientific contributions of this work are: (i) demonstration that soft voting over diverse base learners—spanning linear, bagging, and boosting paradigms—yields exceptional performance on structured cardiovascular data; (ii) validation of SMOTE oversampling as an effective strategy for improving minority class sensitivity in cardiovascular screening; (iii) a comprehensive comparative evaluation demonstrating consistent superiority of the ensemble over individual classifiers and recent benchmark models; (iv) feature importance analysis providing clinically validated insights into the most informative cardiovascular risk markers; and (v) a deployment-ready, serialized prediction pipeline suitable for clinical decision support integration.

Future work will focus on external multi-center validation, integration of explainable AI (SHAP/LIME) for clinician-facing interpretability, extension to multi-class risk stratification, incorporation of wearable IoMT data streams for real-time monitoring, and federated learning for privacy-preserving multi-institutional model training. These advances will move the proposed framework toward robust, interpretable, and clinically deployable cardiovascular AI systems that can meaningfully reduce the global burden of cardiovascular disease.

REFERENCES

[1] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," Sensors, vol. 22, no. 3, pp. 1–18, 2022.

[2] C. A. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers," Sensors, vol. 22, pp. 1–17, 2022.

[3] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," Computational Intelligence and Neuroscience, pp. 1–10, 2023.

[4] Y. M. Ayano, F. Schwenker, B. D. Dupha, and T. G. Debelee, "Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review," Diagnostics, vol. 13, pp. 1–37, 2023.

[5] A. Cuevas-Chávez, Y. Hernández, J. Ortiz-Hernandez, E. Sánchez-Jiménez, G. Ochoa-Ruiz, J. Pérez, and G. González-Serna, "A Systematic Review of Machine Learning and IoT Applied to the Prediction and Monitoring of Cardiovascular Diseases," Healthcare, vol. 11, pp. 1–50, 2023.

[6] A. Altaf, H. Mahdin, A. M. Alive, M. I. H. Ninggal, A. Altaf, and I. Javid, "Systematic Review for Phonocardiography Classification Based on Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 14, no. 8, pp. 806–815, 2023.

[7] J. Zhou, D. You, J. Bai, X. Chen, Y. Wu, Z. Wang, Y. Tang, Y. Zhao, and G. Feng, "Machine Learning Methods in Real-World Studies of Cardiovascular Disease," Cardiovascular Innovations and Applications, vol. 7, pp. 25–36, 2023.

[8] A. Das and P. Dhillon, "Application of Machine Learning in Measurement of Ageing and Geriatric Diseases: A Systematic Review," BMC Geriatrics, vol. 23, pp. 1–35, 2023.

[9] H. A. Al-Alshaikh, P. Prabu, R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, and A. A. AlSanad, "Comprehensive Evaluation and Performance Analysis of Machine Learning in Heart Disease Prediction," Scientific Reports, vol. 14, pp. 1–15, 2024.

[10] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges," Algorithms, vol. 17, pp. 1–33, 2024.

[11] M. Ahmed and I. Husien, "Heart Disease Prediction Using Hybrid Machine Learning: A Brief Review," Journal of Robotics and Control, vol. 5, no. 3, pp. 884–892, 2024.

[12] C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, A. Li, and F. Wang, "A Comprehensive Review of Deep Learning-Based Models for Heart Disease Prediction," Artificial Intelligence Review, vol. 57, pp. 1–50, 2024.

[13] M. Singh et al., "Artificial Intelligence for Cardiovascular Disease Risk Assessment in Personalized Framework: A Scoping Review," eClinicalMedicine, pp. 1–15, 2024.

[14] Q. A. Hidayaturrohman and E. Hanada, "Predictive Analytics in Heart Failure Prediction: A Systematic Review," Preprints, pp. 1–20, 2024.

[15] F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, "The Most Efficient Machine Learning Algorithms in Stroke Prediction: A Systematic Review," Health Science Reports, vol. 7, pp. 1–14, 2024.

[16] I. Fereydooni, M. Vosoughi, A. Alighadr, N. Taghipour, M. Javaherinasab, M. A. Borujeni, and A. Mostafavinia, "A Systematic Review of Machine-Learning Models for Cardiovascular Risk Prediction," International Scientific Trends Journal, pp. 1–18, 2025.

[17] T. Banerjee and I. Paçal, "A Systematic Review of Machine Learning in Heart Disease Prediction," Turkish Journal of Biology, vol. 49, no. 5, pp. 600–634, 2025.

[18] N. H. Alhumaidi, D. Dermawan, H. F. Kamaruzaman, and N. Alotaiq, "The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review," Journal of Medical Internet Research, vol. 26, pp. 1–22, 2024.

[19] G. Gul, I. A. Korejo, D. N. Hakro, H. Alqahtani, A. Abbasi, M. Babar, O. Al Rahbi, and N. I. Ali, "Machine Learning and Ensemble Methods for Cardiovascular Disease Prediction: A Systematic Review of Approaches, Performance Trends, and Research Challenges," Computers, vol. 15, no. 25, pp. 1–30, 2026.

[20] A. Abedi, A. Verma, D. Jain, J. Kaetheeswaran, C. Chui, M. Lankarany, and S. S. Khan, "AI-Driven Real-Time Monitoring of Cardiovascular Conditions with Wearable Devices: Scoping Review," JMIR, pp. 1–17, 2024.