Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081631

ARTICLE

A Modified Gorilla Troops Optimizer-Based Explainable Machine Learning for Early Cardiovascular Disease Prediction

Israt Jahan[1], Afsana Begum[1], Bibhas Roy Chowdhury Piyas[1,*], Fahmid Al Farid[2,3,*], Fatama Jannat Tisha[1], Shahrin Islam[1], Abu Saleh Musa Miah[4] and Hezerul Abdul Karim[3,*]

[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Faculty of Computer Science and Informatics, Berlin School of Business and Innovation, Berlin, Germany
[3]Centre for Image and Vision Computing (CIVC), COE for Artificial Intelligence, Faculty of Artificial Intelligence and Engineering (FAIE), Multimedia University, Cyberjaya, Malaysia
[4]Department of Computer Science and Engineering, Rajshahi University (RU), Rajshahi, Bangladesh

*Corresponding Authors: Bibhas Roy Chowdhury Piyas. Email: piyas.swe@diu.edu.bd; Fahmid Al Farid. Email: fahmid.farid@gmail.com; Hezerul Abdul Karim. Email: hezerul@mmu.edu.my

Received: 06 March 2026; Accepted: 23 April 2026; Published: 23 July 2026

ABSTRACT: Transforming underlying cardiovascular risk into actionable clinical decisions remains a major challenge in contemporary healthcare. Despite advances in cardiology, early-stage cardiovascular disease often remains undetected, which hinders timely intervention and leads to preventable deaths. To overcome this problem, this study presents an explainable machine learning framework for the early diagnosis of cardiovascular disease (CVD). Initially, this study examined several data-balancing strategies, for example, SMOTE (Synthetic Minority Over-sampling Technique), SMOTETomek (Synthetic Minority Over-sampling Technique + Tomek Links), Tomek Links, ADASYN (Adaptive Synthetic Sampling), and SMOTE-ENN (Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors) within the data-preprocessing pipeline. We proposed a novel Adaptive Inertia Weight Gorilla Troops Optimizer (AIW-GTO) to overcome classical GTO's (Gorilla Troops Optimizer) unstable convergence by adaptively controlling step sizes. It uses large exploratory steps early for wide search and smaller steps later for fine-tuned local optimization, which ensures stable convergence and enhanced optimization accuracy. Several machine learning techniques, namely XGBoost, Random Forest, SVM (Support Vector Machine), LightGBM (Light Gradient Boosting Machine), and MLP (Multilayer Perceptron) classifier, were evaluated on the multi-regional UCI heart disease dataset. The experimental findings revealed that, by integrating AIW-GTO Optimization and class imbalance mitigation, LightGBM and XGBoost individually achieved a benchmark accuracy of 93.48% and 91.85%, respectively. Moreover, a weighted ensemble of them further improved the accuracy to 94.02%. Sensitivity analysis further evaluated the model's ability to perform under incomplete clinical test data. To enhance ethical considerations and clinical trust, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were utilized to provide model explainability and identify the most influential features affecting prediction outcomes. Analysis indicated that ECG-related (Electrocardiogram) features, including ST_Slope (exercise-induced ST change) and Oldpeak (ST depression magnitude), emerged as key predictors of CVD risk. Overall, the proposed framework provides a clinically reliable and interpretable approach for early cardiovascular risk assessment to enable proactive patient management.

KEYWORDS: Cardiovascular disease (CVD); machine learning; preventive cardiology; class imbalance handling; model optimization; gorilla troops optimizer; UCI heart disease dataset; explainable AI (XAI); SHAP; LIME; ensemble technique; automated diagnosis; clinical decision support system (CDSS); risk stratification

1 Introduction

Cardiovascular disease is a global epidemic that kills millions of people annually, although it's largely preventable at early stages [1–3]. As reported by World Health Organization (WHO)[1], Cardiovascular Disease (CVD) caused approximately 19.2 million deaths in 2023, which was one-third of all deaths worldwide [4]. This continued rise during the last few decades serves as an indicator of the growing burden of heart and vascular diseases. Despite improvements in medical care, most of the cardiovascular diseases are caused by preventable risk factors, such as hypertension, poor diet, diabetes, obesity and smoking [5,6]. The burden is disproportionately higher in the developing countries, which contribute to more than three-quarters of all CVD-related deaths in the world [7]. One of the biggest challenges is that many cardiovascular conditions are asymptomatic. Many people with early disease have no clear symptoms, which causes a delay in diagnosis and treatment, which makes them more likely to have severe outcomes [8]. Early detection is therefore critical to limit mortality and long-term problems [9].

Cardiovascular disease seldom occurs all at once, but rather it develops gradually through a long chain of biological and behavioral change. When these early warning signs are recognized in time, preventive measures can be made in the form of targeted medication and long-term lifestyle modifications before irreversible damage occurs, which also reduces the need for costly or invasive treatments. Traditional diagnostics are based on symptoms, which tend to miss subclinical cases and do not adequately record individual differences in genetics, lifestyle, and environment. By reorienting diagnosis as a predictive and individualized process rather than a reactive process, machine learning can offer promising opportunities for the prevention of cardiovascular disease [10].

Machine learning (ML) has evolved into a paradigm-shifting tool in clinical automation because of its ability to interpret complex data, uncover hidden patterns, and support predictive decision-making [11]. It is especially useful for early diagnosis of cardiovascular disease (CVD) where early diagnosis helps prevent serious complications and mortality [12,13]. However, most of the existing studies are based mainly on the ML models without combining data analysis, optimization, and class imbalance mitigation [14]. Moreover, the lack of attention towards model interpretability limits the understanding of predictions and clinical trust. Furthermore, limited validation of models in diverse scenarios, particularly with incomplete patient data, raises concerns about the sensitivity and robustness of the model.

The classical optimizers, such as the Gorilla Troops Optimizer (GTO), are typically unstable in convergence, particularly when applied to complex and high-dimensional data such as the UCI Heart Disease dataset. The interactions among the variables in this dataset are usually non-linear and intricate, which makes it difficult for traditional optimization algorithms to effectively explore all possible solutions. The fixed parameters in classical GTO are prone to local minima and premature stagnation, particularly when dealing with noisy data or complex feature interactions. These problems result in suboptimal outcomes, which are critical in medical practice where accurate and reliable predictions are necessary. This indicates the necessity of a modified optimization strategy that will stabilize the convergence and increase the overall reliability of the prediction.

These limitations reduce the model's capability to effectively predict CVD and prevent models from achieving the level of reliability needed for clinical deployment. Since reliability is critical in healthcare, developing resilient, accurate, and interpretable models is imperative for effective early CVD detection.

[1] https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

1.1 Research Gaps

Despite a lot of progress, existing Artificial Intelligence (AI)-based systems for cardiovascular disease diagnosis still have several significant limitations. The important research gaps are listed as follows.

- Existing studies mostly rely on machine learning models without incorporating data analysis, model optimization, and class imbalance mitigation, which limits predictive accuracy and may not be reliable enough for clinical deployment.
- Lack of model interpretability leads to a lack of clinical trust, as well as ethical and regulatory concerns in medical decision-making.
- Existing studies fail to provide adequate benchmarking and ablation analyses to clearly highlight their real contributions with respect to established methods, which makes it difficult to assess the real improvements in performance.
- Validation is frequently limited and prone to overfitting due to the use of single train-test splits instead of more robust methods such as cross-validation. Furthermore, the models were not assessed using datasets with missing clinical test data.

1.2 Key Contributions

In this work, we systematically addressed the key limitations of existing approaches. The primary contributions are as follows:

- Proposed a novel modified Gorilla Troops Optimizer called AIW-GTO, to address the unstable convergence of the classical GTO by adaptively controlling step sizes. The algorithm takes large exploratory steps in initial iterations for global search and smaller step sizes in late iterations for local optimization to ensure stable convergence and enhanced optimization accuracy.
- Performed explainability analysis using Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) to identify the most important features influencing the prediction outcomes, which helps in improving clinical interpretability and transparency.
- Integrated comprehensive data analysis, class balancing and weighted ensemble strategy to achieve strong predictive performance suitable for reliable clinical deployment.
- Conducted sensitivity analysis by systematically dropping key attributes to prove that the model is reliable in situations with limited clinical data.
- Validated the model on the multi-regional UCI Heart Disease dataset and evaluated by applying stratified k-fold cross-validation to achieve generalizability and avoid overfitting.

1.3 Structure of the Paper

- Section 1 introduces the impact of reliable early identification of cardiovascular disease and outlines the existing research gaps, as well as presents the important contributions of this study.
- Section 2 surveys existing research relating to early diagnosis of cardiovascular disease.
- Section 3 details the methodology and illustrates the automated framework that was developed for early CVD identification.
- Section 4 documents the experimental findings and presents an in-depth analysis of the errors.
- Section 5 summarizes the major contributions of the study and suggests possible directions for future work.

2 Literature Review

The accelerated growth of cardiovascular disease (CVD) worldwide has led to massive research for early diagnosis and risk prediction [15,16]. The adoption of ML strategies in clinical practice has allowed the evaluation of high-volume clinical records and the detection of subtle patterns that are difficult to recognize through conventional methods [17]. In recent years, multiple studies have demonstrated promising results using ML frameworks to predict cardiovascular disease using clinical and demographic attributes. Nevertheless, issues including class imbalance, insufficient model explainability, and suboptimal model optimization continue to constrain the effective implementation of such systems in clinical practice.

A broad range of studies have explored AI-driven approaches for cardiovascular disease prediction utilizing clinical risk factors. Experimental validation was commonly performed using benchmark datasets such as the Cleveland [18], Framingham [19], and UCI Heart Disease datasets [20]. Although these datasets share many overlapping attributes related to demographic information, physiological measurements and ECG indicators, they differ in terms of population characteristics, acquisition protocols, and institutional sources.

Chulde-Fernández et al. [21] reported that Random Forest achieved superior performance in heart failure prediction by attaining a specificity of 0.93, an AUC of 0.97, and an MCC (Matthews Correlation Coefficient) of 0.83, while K-Nearest Neighbor (KNN) and Multilayer Perceptron (MLP) exhibited comparatively lower performance. Almutairi and Dardouri [22] introduced a hybrid XGBoost–SVM framework utilizing clinical data and reported an accuracy of 89.3%, outperforming Logistic Regression, Random Forest, and deep neural networks. In a related effort, Cao et al. [23] presented an optimized XGBoost model incorporating feature selection and particle swarm optimization, which yielded an accuracy of 74.7% and an AUC of 80.8%.

Rimal and Sharma [24] examined several optimization strategies, including Optuna Optimization, Bayesian Optimization, and GAsearchCV, together with SVM and Random Forest (RF) using five and ten generations. Their results indicated that the standard Random Forest (RF), BO-optimized SVM, and BO-optimized RF achieved the strongest performance, with accuracies of 86.6%, 90%, and 89%, respectively. Similarly, Khan et al. [25] analyzed RF using clinical data collected from different hospitals of Pakistan, which reported an accuracy of 85.01%, specificity of 43.48%, ROC of 87.73%, and sensitivity of 92.11%, with a misclassification error of 8.70%. In another study, Pasha et al. [26] assessed several ML algorithms on a dataset obtained from Kaggle comprising demographic and clinical features, where they found that a TensorFlow-based Keras ANN (Artificial Neural Network) outperformed traditional SVM, KNN, DT, and ANN models by achieving an accuracy of 85.24%.

Wijesinghe et al. [27] performed a detailed analysis of cardiac MRI (Magnetic Resonance Imaging) segmentation on ventricular structures and myocardium with various U-Net variants with the Automated Cardiac Diagnosis Challenge (ACDC) dataset. The highest performance was shown by the Feature Pyramid U-Net with the Dice scores of 0.9388, 0.8759, and 0.8426 of the left ventricle, right ventricle, and myocardium, respectively. The research paper shows that multi-scale feature extraction enhances precise segmentation of complex cardiac structures and allows to compare U-Net models used in cardiac image analysis. Complementing image-based analysis, Dritsas and Trigka [28] developed an ML framework for cardiovascular disease (CVD) risk prediction using multiple medical risk factors and employed SMOTE to mitigate class imbalance. Ahmed et al. [29] proposed an optimized heart failure survival prediction framework integrating GBM (Gradient Boosting Machines) with the proposed AIW-PSO optimizer on a Kaggle dataset of 299 patients. Trigka and Dritsas [30] investigated deep learning architectures for CVD prediction and introduced an enhanced SMOTE variant for class imbalance handling.

Authors of [31] reviewed explainable artificial intelligence (XAI) in Healthcare 5.0 and emphasized its role in enabling transparent and trustworthy clinical decision-making. Although such results highlight AI's strong predictive capability in the medical sector, the limited interpretability, inadequate model optimization and high computational complexity of the models remain a key challenge. Other recent studies on early CVD prediction are compiled in Table 1.

Table 1: Review of existing studies.

Ref.	Year	Context	Proposed Approach	Limitation	Performance
[32]	2021	Automated cardiovascular disease diagnosis	Hybrid AHDD using Binary CNN with GA-SVM-NB–based feature selection	Complex architecture; higher computational cost	90.1% accuracy
[33]	2021	Coronary cardiovascular disease prediction using ML and ensemble methods	Stacked model (KNN+RF+SVM)	No class imbalance handling or XAI integration.	75.1% accuracy
[34]	2022	Feature-limited cardiovascular disease prediction	Hybrid ML model with feature selection	Evaluated on limited dataset	84.24% accuracy
[35]	2023	Early-stage CVD prediction using ML	Proposed CatBoost model with automatic feature selection	No class imbalance handling or explainable AI integration	90.94% accuracy and 92.3% F1-score
[36]	2024	Cardiovascular disease diagnosis	Heart-Net: Multi-modal DL using MRI (3D U-Net) + ECG (TCGN) with attention fusion	High computational cost; requires multi-modal data	93.45% accuracy
[37]	2025	Deep learning based CVD detection with explainability	ANN incorporating PCA and SHAP	Low sensitivity; RF and SVM outperform ANN	83.4% accuracy
[38]	2025	CVD identification using ensemble ML	Stacking ensemble (CatBoost+ Naïve Bayes + Decision Tree with Gradient Boosting meta-learner)	High computational cost; limited evaluation across diverse datasets	91% accuracy
[39]	2026	CVD classification using quantum ML	Proposed Hybrid Quantum–Neural Network (Autoencoder + QNN)	Requires quantum hardware/simulation; no class imbalance handling or explainable AI	90.98% accuracy

Motivated by these findings and to address key limitations in existing studies, including insufficient accuracy, inadequate class imbalance handling, limited explainability, and incomplete model evaluation, this work presents an explainable machine learning framework incorporating model optimization, class imbalance mitigation, and a weighted ensemble strategy.

3 Proposed Methodology

In this study, we introduced an interpretable machine learning approach for accurate and reliable early cardiovascular disease identification. The proposed method starts with the evaluation of different class balancing strategies, such as Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Over-sampling Technique (SMOTE), Synthetic Minority Over-sampling Technique with Tomek Links (SMOTETomek), Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours (SMOTE-ENN), and Tomek Links. We proposed a novel Adaptive Inertia Weight Gorilla Troops Optimizer (AIW-GTO) to overcome classical GTO's unstable convergence by adaptively controlling step sizes. It uses large exploratory steps early for wide search and smaller steps later for fine-tuned local optimization, which ensures stable convergence and enhanced optimization accuracy. Several benchmark ML architectures, such as SVM, Random Forest, XGBoost, LightGBM, and MLP, were explored with extensive hyperparameter optimization and validated on a popular multi-regional UCI heart disease dataset. To harness the power of individual best-performing models, a weighted ensemble strategy was developed to combine the predictive ability of individual models. In addition, model explainability was added using LIME and SHAP to ensure transparency, ethical accountability, and clinical trust by determining the most influential features driving prediction outcomes. The architecture proposed in this work is given schematically in Fig. 1.

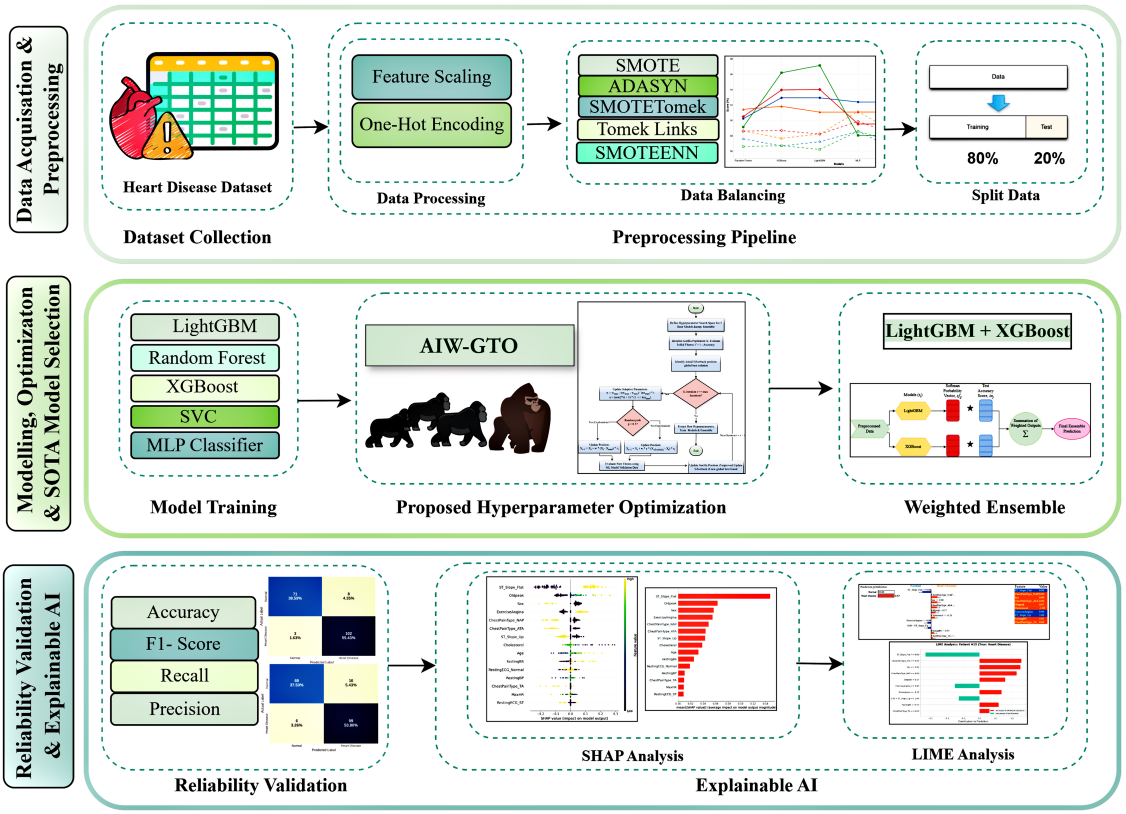

Figure 1: Conceptual workflow of the proposed method.

3.1 Dataset Description

This study utilized the widely adopted UCI Heart Disease dataset from Kaggle, which combines data from the following five well-adapted datasets on CVD: Cleveland, Hungarian, Switzerland, Statlog, and Long Beach VA. The merged dataset has 918 patient records with 508 records of cardiovascular disease cases and 410 records of normal cases. The dataset consists of 11 diagnostic attributes per record and a binary outcome indicating the presence of heart failure. An overview of the dataset features is shown in Table 2.

Table 2: Dataset attribute overview.

SN	Attribute	Value Type	Description	Reference Value
1	Age	Integer	Patient's age in years. Aging is one of the main risk factors of high blood pressure stiffness and heart worsening.	Years (28–77)
2	RestingBP (mmHg)	Integer	Resting blood pressure. Hypertension is persistent (>130/80) leading to workload and risks of cardiac failure.	0–200
3	ChestPainType	String (Nominal)	Categorization of chest pain. Silent ischemia and increased risk are often associated with the so-called asymptomatic (ASY) presentations.	ATA (Atrial Tachyarrhythmia), NAP (Unstable Angina Pectoris), ASY, TA
4	Sex	String (Nominal)	Sex of the patient (biological). Males are usually more predisposed to the early-onset coronary artery disease.	Male, Female
5	RestingECG	String (Nominal)	The results of resting electrocardiogram. Abnormalities of ST and LVH are good indicators of problems with the structure of the heart.	Normal, ST, LVH
6	FastingBS	Integer (Binary)	Fasting blood sugar > 120 mg/dL. Signifies diabetic mellitus, which greatly hastens the vascular injury.	0 (<120), 1 (>120)
7	Cholesterol (mg/dL)	Integer	Serum cholesterol. High levels point to hyperlipidemia, which is a predisposing factor to atherosclerotic accumulation of plaque.	0–603
8	Oldpeak	Float	Exercise vs. rest ST depression. Deep depressions indicate the degree of myocardial ischemia.	−2.6–6.2
9	ExerciseAngina	String (Binary)	Angina induced by exertion. Its existence implies coronary artery constriction incapable of fulfilling demand of oxygen.	N (No), Y (Yes)

(Continued)

Table 2 (continued)

SN	Attribute	Value Type	Description	Reference Value
10	ST_Slope	String (Nominal)	Peak exercise ST segment slope. A flat or down slope is a powerful pathological factor of coronary disease.	Up, Flat, Down
11	MaxHR (bpm)	Integer	Maximum heart rate achieved (bpm).	60–202
12	Cardiovascular Disease	Integer (Binary)	Target Variable. The availability (1) or unavailability (0) of diagnosed cardiovascular disease.	0 (Normal), 1 (Disease)

The attributes of the dataset include a combination of numerical and categorical features and are classified as follows:

- **Numerical features:** Age, Cholesterol, MaxHR, RestingBP and Oldpeak.
- **Categorical features:** FastingBS, Sex, ChestPainType, ST_Slope, ExerciseAngina and RestingECG.

3.2 Data Preprocessing

Raw clinical data represents heterogeneous types of features with varying scales, which may lead to impeding the convergence of machine learning algorithms. In order to overcome this, a two-stage preprocessing pipeline was introduced, concentrating on feature-space modification and class distribution balancing.

3.2.1 Feature Transformation and Scaling

Nominal and ordinal categorical variables were coded differently so that semantic meaning and bias are prevented. Attribute ST_Slope (Up, Flat, Down) has an inherent rank equal to severity. Therefore, ordinal encoding was employed to convert its categories to integer numbers [0, 1, 2]. In contrast, nominal characteristics such as ChestPainType and RestingECG have been one-hot coded.

Numerical features (RestingBP, Age, Oldpeak, MaxHR, Cholesterol) were scaled to have a mean of 0 and unit variance in order to ensure features with large values do not dominate distance-based algorithms. Class balancing was conducted after the normalization to make the numerical features scaled uniformly. The balancing of classes prior to normalization may bias the distribution of classes, as they may have different scales of features. Thus, we adhered to this sequence to avoid data leakage and have a fair assessment of the model.

3.2.2 Strategic Class Balancing

While the original dataset is fairly balanced (55% positive vs. 45% negative), even the slight imbalance of the dataset may induce a bias in the decision boundaries with a complex non-linear model. To analyze the effect of resampling on the effectiveness of the classifiers, 5 different balancing methods were investigated in this study: ADASYN, SMOTE, Tomek Links, SMOTE-Tomek, and SMOTE-ENN.

SMOTE and SMOTE-Tomek achieve class balance by generating samples from the minority class without modifying the majority class, which has the drawback of ensuring equal representation but may lead

to redundancy and increases in the risk of overfitting. On the other hand, Tomek Links as a hybrid approach combines oversampling with selective instance removal to boost class separability without enforcing exact symmetry. Lastly, SMOTE-ENN is a very aggressive algorithm used for noisy samples removal after the oversampling process to generate cleaner decision boundaries instead of a strict class balance.

3.3 Proposed AIW-GTO Algorithm

The Gorilla Troops Optimizer (GTO) is a population-based metaheuristic algorithm based on social hierarchy and collective behavior of gorillas. In this algorithm, each candidate solution is a gorilla, and the best solution in the population is known as the Silverback. The Silverback provides the leadership and directs other gorillas in the movement towards promising areas of the search space. Classical GTO functions with two main phases of behavior, which are exploration and exploitation.

- **Exploration:** It helps gorillas to explore new areas by approaching randomly chosen individuals in the population. This behavior adds diversity and helps the algorithm to avoid premature convergence.
- **Exploitation:** It motivates gorillas to move towards the Silverback, which permits refinement of the best solution discovered until now.

The choice between these two phases is controlled by a probability $p \in [0,1]$.

- $p < 0.5$: The exploration mechanism is activated
- $p \geq 0.5$: the exploitation mechanism is triggered

Although this strategy offers a balance between global and local search, classical GTO does not explicitly control the movement step size across iterations, which may lead to unstable convergence behavior and susceptibility to local optima. To overcome this drawback, we propose a modified version of GTO called the Adaptive Inertia Weight Gorilla Troops Optimizer (AIW-GTO), in which adaptive control parameters are incorporated in the position update mechanism.

- **Adaptive Inertia Weight:** To generate the dynamic process of the magnitude of movement of each gorilla, an adaptive inertia weight w is introduced:

$$w = w_{\max} - \left(\frac{w_{\max} - w_{\min}}{iter_{\max}} \right) \times iter$$

where $iter$ represents the present iteration and $iter_{\max}$ denotes the highest value of iterations. The parameters $w_{\max}$ and $w_{\min}$ define the upper and lower bounds of the inertia weight.

A population size of 15 search agents and 20 iterations was chosen to achieve a balance between computational efficiency and convergence performance, as well as to avoid overfitting during training. The adaptive inertia weight limits were configured to$w_{\max}$ = 0.9 and $w_{\min}$ = 0.4. The initial large value of 0.9 encourages more global exploration during the initial stages of the algorithm, and the gradual decrease to 0.4 encourages more local exploitation during the later stages.

At the early stages of optimization, the inertia weight is large, which allows gorillas to take larger steps and search in a larger region. As the number of iterations increases, the inertia weight slowly decreases, which acts in the minimization of the step size and encourages fine-tuned local search. This adaptive behavior enables a smooth transition from exploration to exploitation during the optimization process.

- **Adaptive Exploration Coefficient:** In addition to inertia control, the coefficient of exploration is reformulated, in order to introduce a kind of iteration-dependent behavior. The coefficient a is defined as-

$$a = (\cos(2r) + 1) \times \left(1 - \frac{iter}{iter_{\max}} \right)$$

where $r \in [0,1]$ is a random variable.

The cosine term contributes a controlled oscillatory movement, which brings diversity to the search process. The periodic oscillatory movement plays an important role in enabling the optimization algorithm to switch between exploration and exploitation. In the initial stages, exploration is prioritized, which allows the algorithm to search a bigger solution space. The periodic exploration is enhanced by the oscillation, which ensures that the algorithm does not become stuck in local minima. Conversely, a linear decay might decrease exploration excessively, which will result in early convergence. Consequently, the algorithm explores more during the early stages and shifts to exploitation during the late stages, which causes the algorithm to converge to the optimal solution in a stable manner.

- **Modified Position Update Mechanism:** The classical GTO position update is enhanced with the flexible inertia weight and exploration coefficient as direct factors in the movement equations.

During the exploration phase ($p < 0.5$), the position of the i-th gorilla is updated as:

$$X_i^{t+1} = X_i^t + w \cdot \left(X_i^t - X_{\text{rand}}\right) \cdot r_1$$

where X_{rand} is a randomly selected gorilla and $r_1 \in [0,1]$ is a random vector.

During the exploitation phase ($p \geq 0.5$), the update rule becomes:

$$X_i^{t+1} = X_i^t + w \cdot a \cdot \left(X_{\text{silverback}} - X_i^t\right) \cdot r_1$$

where $X_{\text{silverback}}$ represents the global best solution.

By utilizing the adaptive inertia weight w and iteration-dependent coefficient a in both phases, the search step size and direction can be changed dynamically during the optimization process by the proposed AIW-GTO. This adaptive control is helpful for the optimizer to better search the complicated hyperparameter spaces. In machine learning applications, wide exploration in the beginning of the process finds many different combinations of hyperparameters, and a gradual decrease in the step size is used to update some promising solutions. This smooth transition helps in convergence stability, avoids premature convergence, and eventually helps in improving the predictive performance of the optimized models.

Fig. 2 illustrates the integration process, which combines the AIW-GTO algorithm with five machine learning classifiers to establish an automated feedback loop that optimizes model hyperparameter settings without requiring human intervention. This involves the following steps: randomly initializing a set of gorillas as hyperparameter combinations, evaluating their performance using k-fold cross-validation, updating their positions based on fitness and selecting the most effective parameters. After completing the specified number of iterations, the best-performing parameters are used to train and test the final machine learning model, which ensures its effectiveness on unseen data.

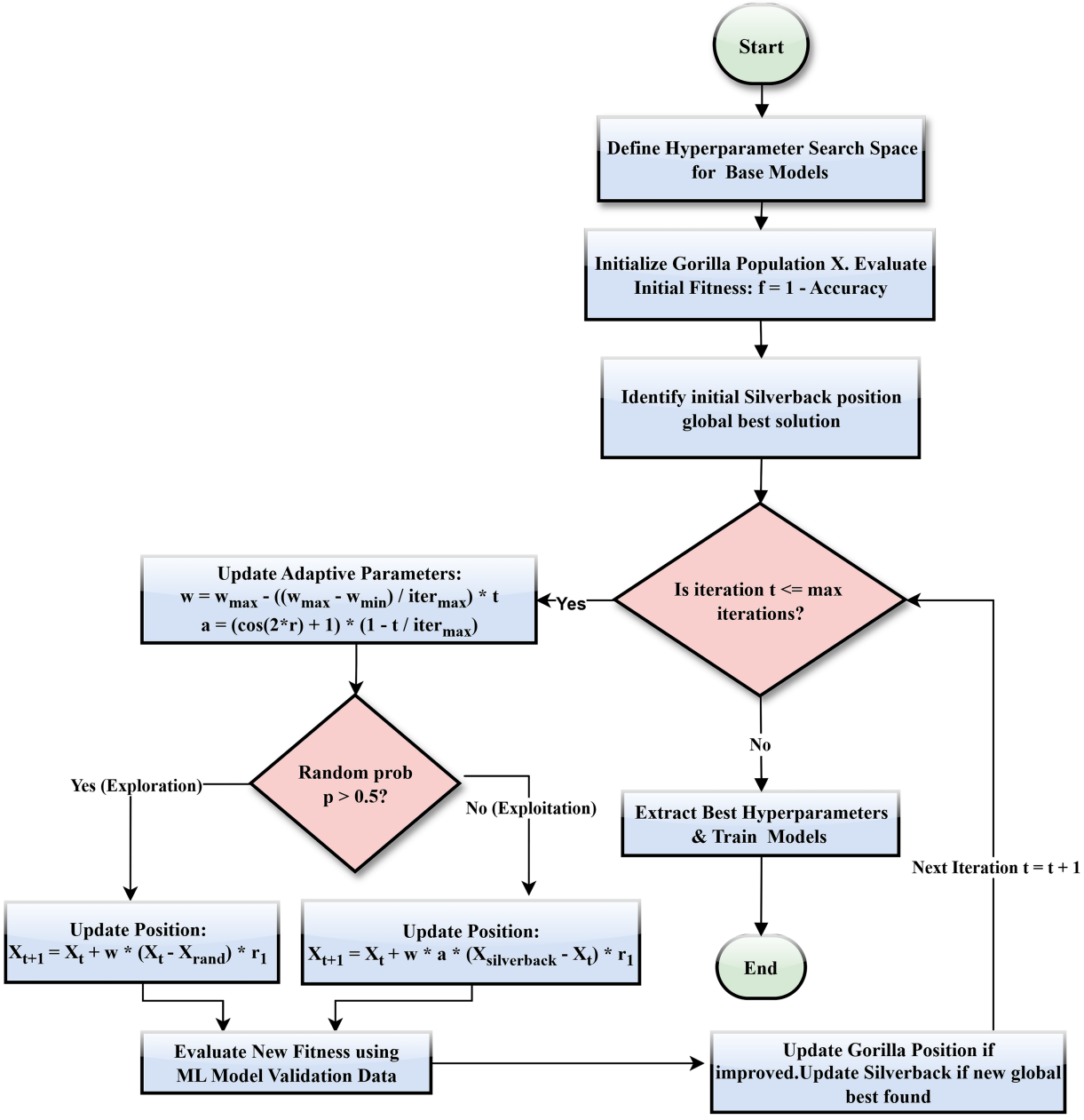

Figure 2: AIW-GTO integration with ML models.

3.4 Weighted Ensemble Approach

Recent research has shown that an ensemble of several learning algorithms can significantly boost the performance in classification tasks. Instead of using a single predictor, ensemble methods combine the complementary predictive powers of multiple models, resulting in more reliable and accurate predictions. Our paper explores pairwise ensemble constructions of five popular machine learning classifiers. Fig. 3 presents the general structure of the developed weighted ensemble.

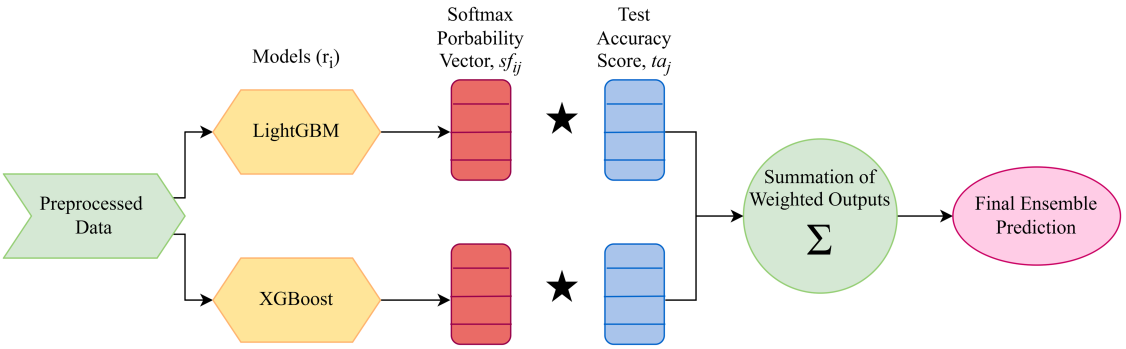

Figure 3: Overview of the weighted ensemble framework.

The average ensemble method predicts the final class by averaging the softmax probability of all the models and choosing the highest combined value. This approach gives the same weight to every base classifier regardless of its previous predictive performance. On the other hand, the suggested weighted ensemble involves a performance-based weighting system to boost the prediction of cardiovascular disease. In this case, the prior test accuracies of each model are used to compute weights so that more reliable classifiers have a stronger influence on the final output. This adaptive weighting results in better robustness and more overall predictive accuracy.

Suppose we have a framework containing r pre-trained models and a test dataset of q individual instances. Each model assigns a class label to an instance q_i from a predefined set of n categories. For each instance q_i, the model r_j outputs an n-dimensional softmax probability vector, denoted as $\mathrm{sp}_{i,j}[n]$. Accordingly, the prediction outputs generated by all models across the test set can be represented as: $sp_{11}[], sp_{21}[], \ldots, sp_{q1}[], sp_{12}[], sp_{22}[], \ldots, sp_{q2}[], \ldots, sp_{1l}[], sp_{2l}[], \ldots, sp_{ql}[]$.

The test accuracy values previously obtained for the r models are denoted by $ta_1, ta_2, \ldots, ta_l$. These accuracy scores are utilized as weighting factors in the proposed ensemble strategy and the final classification outcome is computed based on the formulation presented in Eq. (1).

$$O = \arg\max \left(\frac{\forall i \in (1, q) \sum_{j=1}^{r} \mathrm{sp}_{ij}[n] \cdot \mathrm{ta}_j}{\sum_{j=1}^{r} \mathrm{ta}_j} \right) \tag{1}$$

Here, O denotes a vector of size q, containing the predictions produced by the proposed ensemble method. To obtain the final predictions, the softmax probabilities from each model are first multiplied by their respective prior test accuracy (TA) scores. Weighted probabilities are normalized by dividing each by the sum of all TA scores. The class corresponding to the highest normalized probability is then assigned as the final label for each instance.

4 Results and Discussion

4.1 Experimental Setup

The proportion of the train to the test dataset was 80:20 for model evaluation. All the experiments were done in Google Colab with Python version 3.12.12. We have used the following versions of key Python libraries for experimentation: Python 3.12.13, Pandas 2.2.2, NumPy 2.0.2, Scikit-learn 1.6.1, XGBoost 3.2.0, LightGBM 4.6.0, and SHAP 0.51.0. Table 3 shows the experimental setup, including the optimized

hyperparameters using the proposed AIW-GTO algorithm and the selected hyperparameter values for our experiments.

Table 3: Optimized hyperparameters using proposed AIW-GTO.

Model	Optimized Hyperparameters	Optimized Value
Random Forest	n_estimators	274
	max_depth	13
	min_samples_split	6
XGBoost	learning_rate	0.2650
	max_depth	3
	n_estimators	85
	eval_metric	logloss
LightGBM	learning_rate	0.1266
	num_leaves	34
	n_estimators	226
	verbose	−1
MLP Classifier	hidden_layer_sizes	112
	alpha	0.0005
	learning_rate_init	0.0296
	max_iter	500
Support Vector Machine	C	1.5252
	gamma	0.3128
	probability	true

4.2 Impact of Class Balancing

Table 4 presents a performance evaluation of various machine learning models on the original and modified datasets with different class balancing strategies. Random Forest performs best with ADASYN (90.48%) and Tomek Links (90.22%). XGBoost achieves its highest accuracy with ADASYN (92.93%), while LightGBM reaches peak accuracy with SMOTETomek (92.93%). For a neural network-based model, the MLP classifier excels with SMOTETomek (92.38%), and Support Vector Machines perform best with ADASYN (92.38%). XGBoost and LightGBM outperform all other models, demonstrating their superior reliability in early CVD prediction when appropriate class balancing strategies are applied. Collectively, these results emphasize how carefully selected resampling strategies substantially enhance the outcome of ML models on imbalanced datasets. Fig. 4 illustrates the results before and after the execution of class balancing methods.

Table 4: Impact of class balancing strategies on the outcome of baseline models.

Model	Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	**Without Preprocessing**	**87.59**	**88.61**	**86.59**	**88.60**
	SMOTE	88.04	88.79	90.48	89.62
	SMOTETomek	87.50	87.96	90.48	89.20
	ADASYN	**90.48**	**89.20**	**90.22**	**91.43**

(Continued)

Table 4 (continued)

Model	Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
	Tomek Links	90.22	91.43	89.13	90.48
	SMOTEENN	89.13	90.48	88.04	89.52
XGBoost	**Without Preprocessing**	**86.67**	**87.67**	**86.70**	**88.67**
	SMOTE	90.22	90.22	91.43	91.43
	SMOTETomek	89.67	89.81	92.38	91.08
	ADASYN	**92.93**	**91.82**	**96.19**	**93.95**
	Tomek Links	88.59	90.38	89.52	89.95
	SMOTEENN	88.59	89.62	90.48	90.05
LightGBM	**Without Preprocessing**	**87.22**	**88.22**	**86.22**	**88.22**
	SMOTE	90.22	89.91	93.33	91.59
	SMOTETomek	**92.93**	**91.07**	**97.14**	**94.01**
	ADASYN	90.22	89.91	93.33	91.59
	Tomek Links	89.13	90.48	90.48	90.48
	SMOTEENN	86.96	87.85	89.52	88.68
MLP Classifier	**Without Preprocessing**	**88.58**	**91.17**	**88.57**	**89.85**
	SMOTE	90.22	90.65	92.38	91.51
	SMOTETomek	**92.38**	**91.08**	**88.04**	**89.52**
	ADASYN	87.50	91.84	85.71	88.67
	Tomek Links	90.22	93.94	88.57	91.18
	SMOTEENN	90.22	89.91	93.33	91.59
SVM	**Without Preprocessing**	**86.13**	**88.28**	**87.13**	**88.16**
	SMOTE	90.76	90.74	93.33	92.02
	SMOTETomek	89.67	89.81	92.38	91.08
	ADASYN	**92.38**	**91.08**	**88.04**	**89.52**
	Tomek Links	89.13	89.72	91.43	90.57
	SMOTEENN	88.59	87.50	93.33	90.32

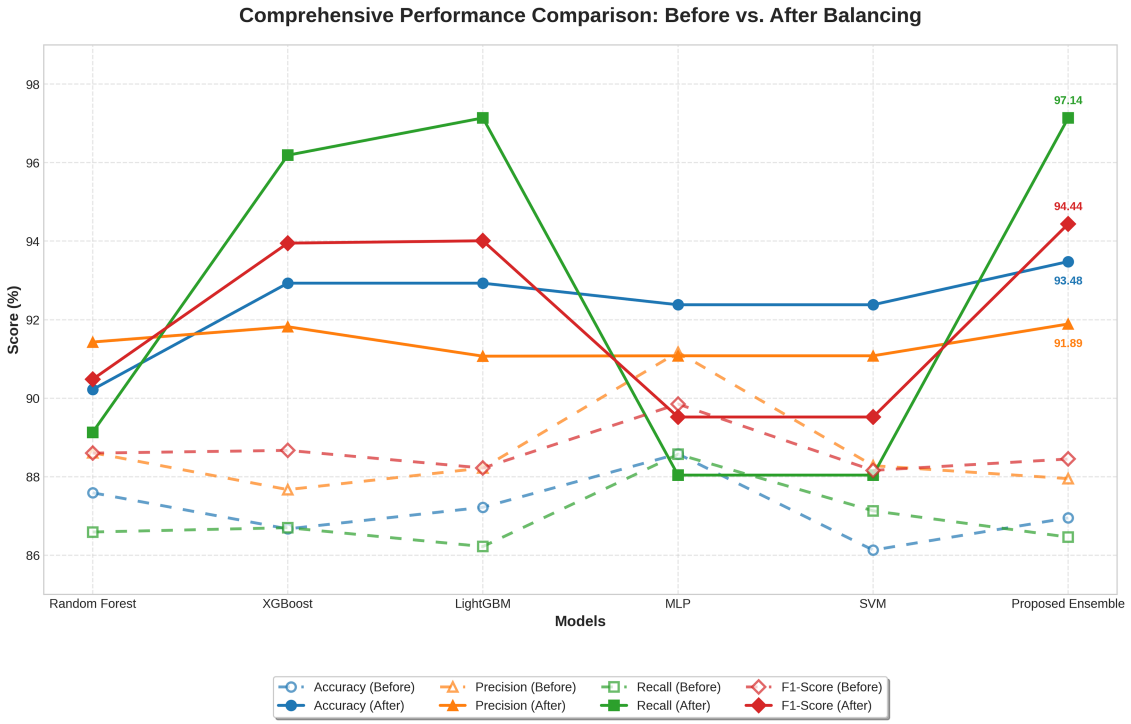

Figure 4: Illustration of performance improvement before and after applying the class balancing techniques.

The observed performance differences among various class balancing strategies can be explained by the fact that they have different interactions with the learning mechanism of the models. Random Forest showed better results on ADASYN and Tomek Links. ADASYN produces synthetic minority samples around the decision boundaries that improve the model to find rare cases of CVD, and Tomek Links eliminates noisy and marginal cases and produces cleaner decision boundaries.

On the contrary, XGBoost was most effective when using ADASYN since this method aims at generating synthetic samples where misclassification is more probable. MLP classifiers and LightGBM demonstrated the best results using SMOTETomek. When used together with synthetic sample generation (SMOTE) and noise removal (Tomek Links), SMOTETomek generates a cleaner and more balanced dataset, which results in significant changes in accuracy and recall. SVM models performed better with ADASYN, which is capable of filling sparse areas around the decision boundaries. This enabled the SVM to create a more distinct separation between classes, which leads to increased classification accuracy.

Overall, the efficacy of each method of balancing depends heavily on the extent to which it complements the underlying learning processes of that particular model. Although ADASYN was particularly useful with XGBoost and SVM, SMOTETomek performed optimally with LightGBM and MLP classifiers.

4.3 Impact of AIW-GTO Optimization

The comparative performance of manual hyperparameter selection, classical GTO, and the proposed AIW-GTO optimization strategy is presented in Table 5. In manual optimization, the parameters of the baseline model are optimized, and the accuracy of the baseline model is then used as the weight of the ensemble models. A similar process is applied to the classical GTO and AIW-GTO methods, where their hyperparameters are optimized before their accuracy is considered as weights.

Table 5: Impact of the proposed AIW-GTO optimizer on model performance.

Method	Cardiovascular Disease			No Cardiovascular Disease			Accuracy
	P	R	F	P	R	F	
Manual Hyperparameter Selection							
Random Forest	91.43	90.48	90.95	89.20	90.22	89.71	90.48
XGBoost	91.82	96.19	93.95	94.59	88.61	91.50	92.93
LightGBM	91.07	97.14	94.01	95.83	87.34	91.39	92.93
MLP Classifier	92.31	91.43	91.87	91.08	88.04	89.52	92.38
SVM	91.08	94.03	92.52	89.81	92.38	91.08	92.38
Random Forest + XGBoost	91.50	93.33	92.41	90.91	88.61	89.74	91.30
Random Forest + LightGBM	91.20	94.29	92.72	92.00	87.97	89.94	91.58
Random Forest + MLP Classifier	90.57	91.43	91.00	88.46	87.34	87.90	89.67
Random Forest + SVM	90.65	92.38	91.51	89.61	87.34	88.46	90.22
LightGBM + XGBoost	**91.89**	**97.14**	**94.44**	**95.89**	**88.61**	**92.11**	**93.48**
LightGBM + MLP Classifier	89.72	91.43	90.57	88.31	86.08	87.18	89.13
LightGBM + SVM	90.83	94.29	92.52	92.00	87.34	89.61	91.30
XGBoost + MLP Classifier	92.23	90.48	91.35	87.65	89.87	88.75	90.22
XGBoost + SVM	92.38	92.38	92.38	89.87	89.87	89.87	91.30
MLP Classifier + SVM	90.57	91.43	91.00	88.46	87.34	87.90	89.67
Hyperparameter Optimization Using Classical GTO							
Random Forest	90.20	90.48	91.14	91.80	89.90	90.84	91.30
XGBoost	92.60	94.40	93.4	94.00	92.20	93.00	93.48
LightGBM	92.60	94.40	93.49	94.00	92.20	93.09	93.48
MLP Classifier	91.90	93.80	92.84	93.50	91.60	92.54	92.93
SVM	91.40	93.50	92.44	93.20	91.00	92.09	92.39
Random Forest + XGBoost	91.50	93.60	92.54	93.30	91.10	92.19	92.39
Random Forest + LightGBM	91.50	93.60	92.54	93.30	91.10	92.19	92.39
Random Forest + MLP Classifier	91.10	93.00	92.04	92.80	90.50	91.64	92.10
Random Forest + SVM	90.90	92.80	91.84	92.50	90.20	91.34	91.85
LightGBM + XGBoost	**92.70**	**94.50**	**93.59**	**94.10**	**92.30**	**93.19**	**93.48**
LightGBM + MLP Classifier	92.10	93.90	92.99	93.60	91.70	92.64	93.10
LightGBM + SVM	91.90	93.80	92.84	93.50	91.60	92.54	92.93
XGBoost + MLP Classifier	92.10	93.90	92.99	93.60	91.70	92.64	93.10
XGBoost + SVM	91.90	93.80	92.84	93.50	91.60	92.54	92.93
MLP Classifier + SVM	91.60	93.70	92.64	93.40	91.20	92.29	92.65
Hyperparameter Optimization Using Proposed AIW-GTO							
Random Forest	90.50	92.50	91.49	92.10	90.00	91.04	91.30
XGBoost	91.00	93.10	92.04	92.70	90.50	91.59	91.85
LightGBM	92.80	94.50	93.64	94.10	92.40	93.24	93.48
MLP Classifier	92.10	94.00	93.04	93.80	91.80	92.79	92.93
SVM	91.50	93.60	92.54	93.30	91.10	92.19	92.39

(Continued)

Table 5 (continued)

Method	Cardiovascular Disease			No Cardiovascular Disease			Accuracy
	P	R	F	P	R	F	
Random Forest + XGBoost	91.80	93.80	92.79	93.50	91.40	92.44	92.65
Random Forest + LightGBM	92.90	94.60	93.74	94.30	92.30	93.29	93.55
Random Forest + MLP Classifier	92.20	94.10	93.14	93.80	91.70	92.74	92.98
Random Forest + SVM	91.60	93.70	92.64	93.40	91.20	92.29	92.50
LightGBM + XGBoost	**93.50**	**95.00**	**94.24**	**94.60**	**93.00**	**93.79**	**94.02**
LightGBM + MLP Classifier	93.10	94.80	93.94	94.50	92.50	93.49	93.75
LightGBM + SVM	93.00	94.70	93.84	94.40	92.40	93.39	93.68
XGBoost + MLP Classifier	92.40	94.30	93.34	94.10	91.90	92.99	93.20
XGBoost + SVM	92.30	94.20	93.24	93.90	91.70	92.79	93.05
MLP Classifier + SVM	92.35	94.25	93.29	94.00	91.80	92.89	93.10

The results show that classical GTO outperforms manual tuning across most classifiers and ensemble configurations. For example, the accuracy of the Random Forest improved from 90.48% using the manual tuning to 91.30% using classical GTO, and XGBoost improved from 92.93% to 93.48%. Comparable improvements are also found for F1-scores and recall values for models, which reveal the ability of GTO to better explore the hyperparameter search space and identify more competitive configurations than static manual selection. Despite these improvements, the performance gains made by classical GTO are moderate and sometimes inconsistent across models due to the unstable convergence and susceptibility to local optima.

In contrast, the proposed AIW-GTO framework provides more significant and consistent performance improvements. Notably, the LightGBM + XGBoost ensemble achieves an accuracy of 94.02% with the improvements of 0.54 and 0.89 percentage over classical GTO and manual tuning, respectively. Similarly, Random Forest + LightGBM achieves 93.55% accuracy with a gain of 1.16 percentage points over GTO and 1.97 percentage points over manual selection. These results provide evidence that the adaptive inertia-based optimization can effectively stabilize the convergence process, reduce the premature stagnation, and improve the global search capability of GTO, which leads to better and balanced classification performance.

4.4 K-Fold cross Validation

To measure the reliability and stability of the proposed model, cross-validation was applied to measure its performance across different input scenarios. A 5-fold cross-validation was conducted in this work. As shown in Table 6, the proposed model obtained a mean accuracy of 0.925 with a Standard Deviation of 0.012 across folds which shows its consistent performance. These findings suggest the strong generalization capability of the model and its possible practical deployment in real-world settings.

4.5 Error Analysis

The confusion matrices of all the optimized baseline classifiers and the proposed weighted ensemble model are illustrated in Fig. 5. A comparative error analysis shows that the ensemble approach consistently reduces the number of misclassifications, especially false negative which is clinically the most important in cardiovascular disease prediction.

Table 6: Performance evaluation of proposed model using 5-fold cross-validation.

Fold	Precision	Recall	F1-Score	Accuracy
Fold 1	0.923	0.930	0.926	0.925
Fold 2	0.915	0.924	0.919	0.918
Fold 3	0.928	0.935	0.931	0.931
Fold 4	0.920	0.926	0.923	0.922
Fold 5	0.925	0.932	0.920	0.928
Mean	0.922	0.929	0.915	0.925
Standard Deviation	0.010	0.014	0.020	0.012

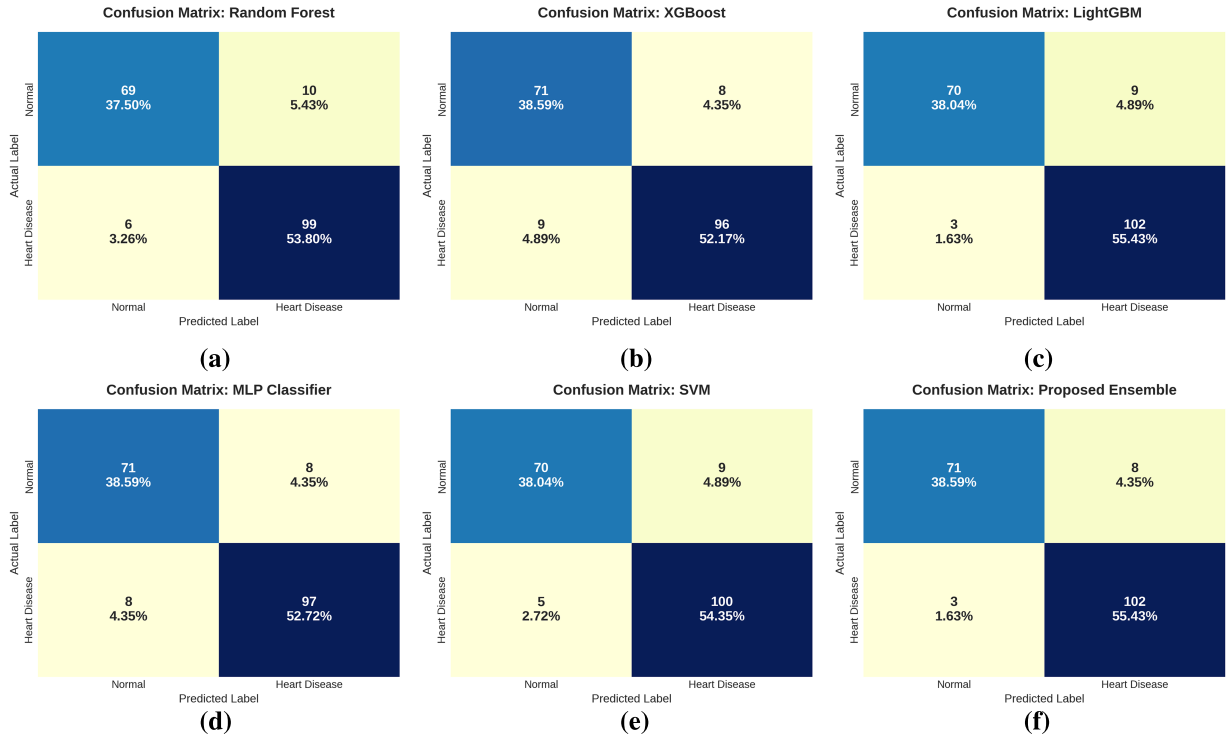

Figure 5: Confusion matrix. (**a**) Random forest (**b**) XGBoost (**c**) LightGBM (**d**) MLP Classifier (**e**) SVM (**f**) Proposed ensemble model.

Among the baseline models, XGBoost and MLP Classifier show relatively higher false negative rates, which suggests that there is a tendency for cardiovascular disease patients to be misclassified as normal. Such errors pose a serious clinical concern as they can be life-threatening due to a delay in necessary medical intervention. Although SVM, Random Forest, and LightGBM show improved sensitivity but sporadic false positives and false negatives are still observed, which reflects the limitations involved in using a single learner even after applying class balancing techniques.

In contrast, the proposed ensemble model achieves the minimum number of false negatives (3 cases, 1.63%) and false positives (8 cases,4.35%), as well as a high true positive rate (102 correctly identified cardiovascular disease cases, 55.43%). This indicates the proposed weighted ensemble model effectively learns complementary decision boundaries captured by the base classifiers, resulting in more accurate

predictions. The decrease in the number of false negatives highlights the power of the model to focus on clinically significant patterns instead of overfitting on the majority class characteristics.

4.6 Efficiency Analysis

The computational complexity analysis of the baseline and ensemble models is provided in Table 7. For the baseline models, the number of parameters is defined as the total count of trainable parameters in the model. For the ensemble models, the number of parameters is the sum of the parameters from each base model. Therefore, the reported parameter count for the ensemble model represents the cumulative sum of the parameters from all its constituent base models.

Table 7: Analysis of model performance and computational efficiency.

Model	Training Time (Min)	Inference Time (ms)	Parameter Count (thousand)	Model Size (Mb)	Error Rate (%)	GPU Memory Usage (Gb)
Random Forest	0.0143	15.7202	21.4500	1.6670	8.7000	0
XGBoost	0.0048	2.0542	6.3000	0.1764	8.1500	0
LightGBM	0.0015	1.9949	3.1000	0.3183	6.5200	0
MLP Classifier	0.0918	0.2863	1.3010	0.0648	7.0700	0
SVM	0.0019	0.4540	3.6190	0.0351	7.6100	0
Random Forest + XGBoost	0.0191	17.7744	27.7500	1.8434	7.3500	0
Random Forest + LightGBM	0.0158	17.7151	24.5500	1.9854	7.4500	0
Random Forest + MLP Classifier	0.1061	16.0065	22.7510	1.7319	7.0200	0
Random Forest + SVM	0.0163	16.1743	25.0690	1.7021	7.5000	0
LightGBM + XGBoost	**0.0063**	**4.0491**	**9.4000**	**0.4947**	**5.9800**	**0**
LightGBM + MLP Classifier	0.0932	2.2812	4.4010	0.3831	6.2500	0
LightGBM + SVM	0.0034	2.4489	6.7190	0.3534	6.3200	0
XGBoost + MLP Classifier	0.0966	2.3405	7.6010	0.2412	6.8000	0
XGBoost + SVM	0.0067	2.5082	9.9190	0.2115	6.9500	0
MLP Classifier + SVM	0.0937	0.7403	4.9200	0.0999	6.9000	0

The analysis indicates that ensemble models tend to have minimal increased training time, inference latency, and model size due to a larger number of parameters. Nevertheless, these moderately heavier models are often realized with lower error rates and with more stable performance, which makes them preferable for medical applications where reliability outweighs the deployment constraints. The ensemble of LightGBM + XGBoost achieves the lowest error rate (5.98%) with moderate computational cost, which shows a great balance between accuracy and efficiency.

While all evaluated models report zero GPU memory usage, this outcome is expected given the experimental setup, considering the dataset size and the lightweight baseline models used. In the case of the real-time large hospital database, we expect certain scaling up of model size, training time, parameters, and the use of GPUs. Even though the computational requirements will increase in such a large-scale environment, in the present and realistic real-life conditions, our proposed model is very reliable and computationally efficient.

4.7 Explainability

While the proposed approach offers high predictive accuracy, ML models in clinical applications require transparency and interpretability. Explainability ensures that predictions are guided by important features, rather than by random factors. We employed LIME and SHAP to assess the impact of different attributes on the model's outputs.

4.7.1 Model Interpretability Using SHAP

The explainability through SHAP in Fig. 6a provides a clear understanding of the effects of various attributes on the cardiovascular disease prediction model. ST_Slope_Flat has the strongest impact among all the variables, significantly contributing to the model outcome. This is followed by Oldpeak, Sex, Exercise, Angina, and ChestPainType features that reveal that Electrocardiogram (ECG) patterns and body exercise symptoms play vital roles in the model decisions. The SHAP summary plot in Fig. 6b indicates that higher Oldpeak values and exercise-induced angina increase the likelihood of cardiovascular disease predictions, while favorable ST_Slopes and lower Oldpeak values lead to negative contributions. Variables such as Age, Cholesterol, and FastingBS show moderate influence on the model output.

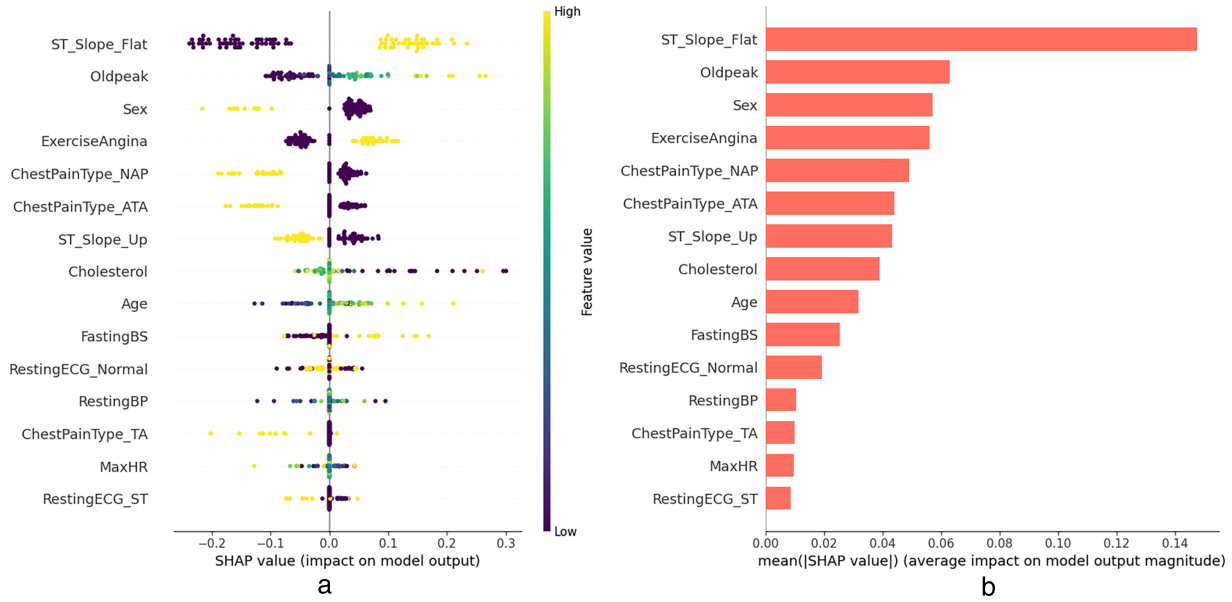

Figure 6: Explainability analysis using SHAP (**a**) SHAP summary plot. (**b**) Global SHAP feature importance.

4.7.2 Model Interpretability Using LIME

The LIME analysis shown in Fig. 7 offers a clear instance-level explanation for a patient correctly predicted as having cardiovascular disease. Several features display pronounced positive contributions to this prediction, such as Oldpeak, ChestPainType_NAP, ChestPainType_ATA, Sex, and elevated FastingBS, which indicates that chest pain characteristics, exercise-related ST depression, and metabolic factors contribute to a pivotal role in driving the decision towards cardiac disease. In contrast, a limited number of features, such as ST_Slope_Flat, ST_Slope_Up and no ExerciseAngina, have a negative effect on the outcome, by decreasing the predicted risk. However, their impact remains comparatively small and insufficient to offset the dominant risk-associated features. The balanced visualization of the positive and negative contributions is a way to show how the model integrates competing clinical signals at the individual level. This patient-level

explanation is consistent with known cardiovascular risk factors and shows that the model's predictions are still interpretable and meaningful from a clinical standpoint.

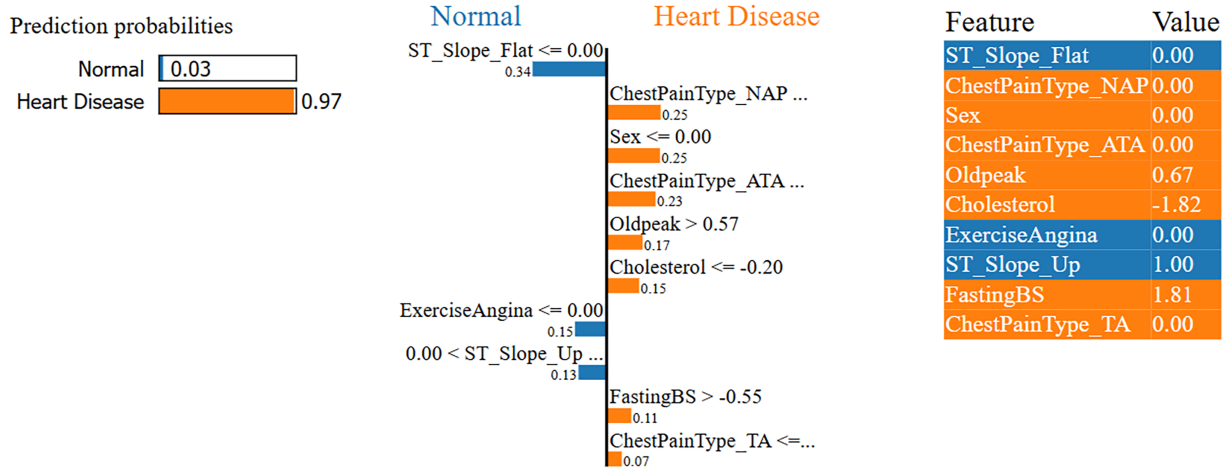

Figure 7: Explainability analysis using LIME.

4.7.3 Analysis of Model Interpretability

The outputs from both LIME in Figs. 6 and 7 provide consistent insights into the feature importance for predicting cardiovascular disease, which reinforces the reliability of the model's explanations. SHAP's global feature importance analysis highlights ST_Slope_Flat, Oldpeak, and ExerciseAngina as critical features influencing model predictions, aligning with LIME's local explanation for an individual high-risk patient, where Oldpeak and ST_Slope_Flat have notable contributions to the prediction of cardiovascular disease. In particular, both methods emphasize the importance of Cholesterol and ExerciseAngina which confirms that these features play a significant role in predicting cardiovascular disease across different instances. The alignment between the two approaches demonstrates the stability and trustworthiness of the model's reasoning process, which enhances trust in its clinical interpretability.

According to guidelines from the American Heart Association and American College of Cardiology[2], exercise electrocardiography (ECG) remains a fundamental tool for detecting myocardial ischemia where ST-segment abnormalities play a central diagnostic role. Specifically, these guidelines highlight that horizontal or downsloping ST-segment depression during exercise testing is strongly associated with myocardial ischemia, while the magnitude of ST depression (commonly quantified as Oldpeak) correlates with the severity and extent of coronary artery disease. Furthermore, greater ST depression values are linked to a higher likelihood of significant coronary obstruction and adverse cardiac outcomes [40].

The alignment between the model's feature importance and these established diagnostic criteria suggests that the proposed approach captures medically meaningful patterns rather than spurious correlations, which enhances its reliability and potential applicability for early cardiovascular disease detection.

4.8 Ablation Study

An ablation study was carried out to examine the robustness of the model in Table 8 by sequentially removing three clinically important features, such as ST_Slope_Flat, Oldpeak, and ExerciseAngina, in order to investigate the early diagnosis of CVD with incomplete patient information. Across all models, there is a

[2] https://www.acc.org/about-acc/press-releases/2025/02/28/17/51/acc-aha-issue-new-acute-coronary-syndromes-guideline

slight drop in performance after feature removal, but it is not significant. Importantly, the proposed model is able to demonstrate competitive performance consistently, which is an indication of its ability to handle missing clinical features more effectively than individual machine learning models.

Table 8: Performance after removing key parameters.

Removed Parameter	Model	Acc. (%)	Precision (%)	Recall (%)	F1-score (%)
ST_Slope_Flat	Random Forest (RF)	84.24	86.87	84.31	85.57
	Light Gradient Boosting Machine (LGBM)	86.41	87.38	88.24	87.80
	XGBoost	85.87	87.00	86.27	87.13
	Support Vector Machine (SVM)	80.43	84.38	79.41	81.82
	Multi-Layer Perceptron (MLP)	84.24	86.87	84.31	85.57
	Proposed Ensemble Framework	**85.89**	**87.45**	**87.25**	**87.73**
Oldpeak	Random Forest (RF)	84.78	88.54	83.33	85.86
	Light Gradient Boosting Machine (LGBM)	84.24	85.44	86.27	85.85
	XGBoost	82.07	84.16	83.33	83.74
	Support Vector Machine (SVM)	82.07	85.57	81.37	83.42
	Multi-Layer Perceptron (MLP)	82.07	82.86	85.29	84.06
	Proposed Ensemble Framework	**84.88**	**85.44**	**86.27**	**85.85**
ExerciseAngina	Random Forest (RF)	84.78	87.76	84.31	86.00
	Light Gradient Boosting Machine (LGBM)	85.33	87.13	86.27	86.70
	XGBoost	84.78	87.76	84.31	86.00
	Support Vector Machine (SVM)	83.70	87.50	82.35	84.85
	Multi-Layer Perceptron (MLP)	84.24	85.44	86.27	85.85
	Proposed Ensemble Framework	**85.87**	**87.76**	**87.25**	**87.25**

The proposed model attains an accuracy of 85.89% with the exclusion of ST_Slope_Flat, which is higher than RF (84.24%), XGBoost (85.87%), and LGBM (86.41%). The same pattern is noticed on the removal of Oldpeak, where the ensemble scores 84.88% accuracy and F1-score of 85.85%, which is higher than RF (84.78%), XGBoost (82.07%), and MLP (82.07%). On the other hand, the deletion of ExerciseAngina leads to very little variation, with the proposed model having about 85.87% accuracy. Overall, these results show that, even though there are some small effects on the predictive performance of the results due to the absence of essential attributes, the proposed model is highly consistent and balanced performer under all circumstances.

4.9 Comparison of Existing Studies

Although substantial advancements have been made, the early identification of cardiovascular disease still has opportunities for further improvement. This study reviews several existing strategies previously conducted for early cardiovascular disease detection. In order to ensure fair comparison, we employed the

best-performing models in the previous research in our dataset and compared the outcomes. This approach ensures a fairer and more consistent comparison. Meanwhile, we have retained the originally reported findings for studies that used datasets similar to ours. The comparative test accuracies of these methods are demonstrated in Table 9, showing that the proposed model attains state-of-the-art (SOTA) performance for the identification task.

Table 9: Comparison with related studies.

Paper	Data Source	No. of Samples	Explored Models	Best Model	Accuracy	XAI Used
[41]	UCI Cleveland	303	Hybrid DSS, SMOTE, RF	Random Forest	90.48%	No
[42]	UCI Repository	270	SVM, LR, GBM, DT, RF	SVM	92.38%	No
[43]	Cleveland Clinic Foundation	297	Decision Tree (C4.5)	Decision Tree (C4.5)	85.82%	No
[44]	UCI Heart Disease Dataset	918	DT, KNN, SVM, XGB	SVM	88.80%	No
[21]	UCI Heart Disease Dataset	918	LR, DT, RF, KNN, MLP	Random Forest	92.00%	No
Proposed Model	**UCI Heart Disease Dataset**	**918**	**RF, XGB, LGBM, SVC (Support Vector Classifier), MLP**	**AIW-GTO optimized (XGBoost + LightGBM)**	**94.02%**	**Yes**

5 Conclusion and Future Work

This study presents an explainable Machine Learning framework for early CVD detection by addressing key research gaps in data preprocessing, model optimization, class imbalance, and interpretability. We evaluated several class balancing strategies, including SMOTE, SMOTE-Tomek, ADASYN, Tomek Links, SMOTE-ENN, and proposed AIW-GTO, a modified Gorilla Troops Optimizer with adaptive inertia for stable convergence and optimal hyperparameter selection. On the UCI heart disease dataset, AIW-GTO-optimized LightGBM and XGBoost achieved 93.48% and 91.85% accuracy while their weighted ensemble reached 94.02% with low false negatives (1.63%) and false positives (4.35%). Computational and sensitivity analyses further confirmed deployment feasibility and robustness to incomplete data. LIME and SHAP analyses highlighted ECG features, ST-segment slope, and Oldpeak as top predictors of CVD, enabling clinical trust and early interventions.

Even though the UCI Heart Disease dataset is a well-established benchmark that has been extensively used in prior studies, its smaller size can pose a possible overfitting risk when training complex ensemble models with metaheuristic optimizers. Future research may aim at creating larger, multi-regional, and more representative datasets to improve the prediction of early cardiovascular disease.

Moreover, limitations include the use of one dataset, which may not represent the diversity of the global population. A single-source dataset may be characterized by certain regional or demographic attributes that

are not necessarily applicable to other groups of people. Integrating multi-region data is important because it allows the model to identify fluctuations in genetic, environmental, and socio-economic data, which can lead to stronger and generalized predictions. This is particularly important in clinical settings where disease patterns and treatment responses vary across populations. Furthermore, real-time clinical validation is still missing. The future research will be to extend the evaluations to different populations and the real-time clinical validation to confirm the effectiveness of the model in clinical practice.

Acknowledgement: We respectfully thank the dedication and commitment of all coauthors which were crucial to completion ofthis study.

Funding Statement: This study is supported by Multimedia University, Malaysia (Grant Number: PostDoc (MMUI/240029)). This work is partially supported by Daffodil International University, Bangladesh.

Author Contributions: Conceptualization: Israt Jahan, Afsana Begum and Bibhas Roy Chowdhury Piyas; Coding and simulations: Israt Jahan, Bibhas Roy Chowdhury Piyas, Fatama Jannat Tisha and Shahrin Islam; Analysis and interpretation of results: Bibhas Roy Chowdhury Piyas, Afsana Begum, Abu Saleh Musa Miah, Fahmid Al Farid and Hezerul Abdul Karim; Manuscript preparation: Israt Jahan, Bibhas Roy Chowdhury Piyas, Fatama Jannat Tisha, Shahrin Islam and Abu Saleh Musa Miah; Acquisition of funding: Fahmid Al Farid and Hezerul Abdul Karim. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset with the implementation code of the proposed model has been made publicly available via the following GitHub repository: https://github.com/shahrin-islam-swe/advancing-early-cardiovascular-disease-diagnosis.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Arooj S, Rehman Sur, Imran A, Almuhaimeed A, Alzahrani AK, Alzahrani A. A deep convolutional neural network for the early detection of heart disease. Biomedicines. 2022;10(11):2796. doi:10.3390/biomedicines10112796.
2. Huang JD, Wang J, Ramsey E, Leavey G, Chico TJ, Condell J. Applying artificial intelligence to wearable sensor data to diagnose and predict cardiovascular disease: a review. Sensors. 2022;22(20):8002. doi:10.3390/s22208002.
3. Adil M, Javaid N, Ahmed I, Ahmed A, Alrajeh N. A deep learning framework for heart disease prediction with explainable artificial intelligence. Comput Mater Contin. 2026;86(1):1–20. doi:10.32604/cmc.2025.071215.
4. Martin SS, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. heart disease and stroke statistics: a report of US and global data from the American heart association. Circulation. 2024 2024;149(8):e347–913. doi:10.1161/CIR.0000000000001209.
5. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Comput Sci. 2020;1(6):345. doi:10.1007/s42979-020-00365-y.
6. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019;7:81542–54. doi:10.1109/ACCESS.2019.2923707.
7. Gour S, Panwar P, Dwivedi D, Mali C. A machine learning approach for heart attack prediction. In: Intelligent sustainable systems: selected papers of WorldS4 2021. Vol. 1. Singapore: Springer Nature Singapore; 2022. p. 741–7. doi:10.1007/978-981-16-6309-3_70.
8. Alom Z, Azim MA, Aung Z, Khushi M, Car J, Moni MA. Early stage detection of heart failure using machine learning techniques. In: Proceedings of the International Conference on Big Data, IoT, and Machine Learning (BIM 2021). Singapore: Springer; 2021. p. 75–88. doi:10.1007/978-981-16-6636-0_7.

9. Sharma A, Dhanka S, Kumar A, Nain M, Dhanka B, Bhardwaj VK, et al. A systematic review on machine learning intelligent systems for heart disease diagnosis. Arch Comput Methods Eng. 2025;32(7):4303–29. doi:10.1007/s11831-025-10271-2.
10. Nagavelli U, Samanta D, Chakraborty P. Machine learning technology-based heart disease detection models. J Healthc Eng. 2022;2022(1):7351061. doi:10.1155/2022/7351061.
11. Tabassum S, Muhammad F, Khan MA, Khan MU, Awan D, Gohar N, et al. A machine learning-based framework for heart disease diagnosis using a comprehensive patient cohort. Comput Mater Contin. 2025;84(1):1253–78. doi:10.32604/cmc.2025.065423.
12. Eleyan A, AlBoghbaish E, AlShatti A, AlSultan A, AlDarbi D. RHYTHMI: a deep learning-based mobile ECG device for heart disease prediction. Appl Syst Innov. 2024;7(5):77. doi:10.3390/asi7050077.
13. Ge B, Yang H, Ma P, Guo T, Pan J, Wang W. Detection of pulmonary hypertension associated with congenital heart disease based on time-frequency domain and deep learning features. Biomed Signal Process Control. 2023;81(5):104316. doi:10.1016/j.bspc.2022.104316.
14. Amin R, Khan MJ, Nath TD, Reza MS, Shin J. Harmonization of heart disease dataset for accurate diagnosis: a machine learning approach enhanced by feature engineering. Comput Mater Contin. 2025;82(3):3907–19. doi:10.32604/cmc.2025.061645.
15. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):1–16. doi:10.1186/s12911-019-1004-8.
16. Prasanna KSL, J. V, Srinivasu PN, Shah B, Ali F. Optimized cardiovascular disease prediction using clustered butterfly algorithm. Comput Mater Contin. 2025;85(1):1603–30. doi:10.32604/cmc.2025.068707.
17. Liu J, Dong X, Zhao H, Tian Y. Predictive classifier for cardiovascular disease based on stacking model fusion. Processes. 2022;10(4):749. doi:10.3390/pr10040749.
18. Cleveland Dataset by KEEL. "Cleveland dataset". [cited 2026 Apr 22]. Available from: https://sci2s.ugr.es/keel/dataset.php?cod=57.
19. Framingham Dataset by Kaggle. "Framingham heart study dataset". [cited 2026 Apr 22]. Available from: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset.
20. Heart Disease Dataset by UCI. "Heart disease dataset". [cited 2026 Apr 22]. Available from: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
21. Chulde-Fernández B, Enríquez-Ortega D, Guevara C, Navas P, Tirado-Espín A, Vizcaíno-Imacaña P, et al. Classification of heart failure using machine learning: a comparative study. Life. 2025;15(3):496. doi:10.3390/life15030496.
22. Almutairi M, Dardouri S. Intelligent hybrid modeling for heart disease prediction. Information. 2025;16(10):869. doi:10.3390/info16100869.
23. Cao K, Liu C, Yang S, Zhang Y, Li L, Jung H, et al. Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. Sci Rep. 2025;15(1):12406. doi:10.1038/s41598-025-96520-7.
24. Rimal Y, Sharma N. Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy. Multimed Tools Appl. 2024;83(18):55091–107. doi:10.1007/s11042-023-17273-x.
25. Khan A, Qureshi M, Daniyal M, Tawiah K. A novel study on machine learning algorithm-based cardiovascular disease prediction. Health Soc Care Community. 2023;2023(1):1406060. doi:10.1155/2023/1406060.
26. Pasha SN, Ramesh D, Mohmmad S, Harshavardhan A, Shabana. Cardiovascular disease prediction using deep learning techniques. IOP Conf Ser Mater Sci Eng. 2020;981(2):022006. doi:10.1088/1757-899X/981/2/022006.
27. Wijesinghe CB, Meedeniya D, Yogarajah P. Cardiac MRI segmentation of ventricular structures and myocardium using U-Net variants. In: Proceedings of the 2025 5th International Conference on Advanced Research in Computing (ICARC); 2025 Feb 19–20; Belihuloya, Sri Lanka. p. 1–6. doi:10.1109/ICARC64760.2025.10963166.
28. Dritsas E, Trigka M. Efficient data-driven machine learning models for cardiovascular diseases risk prediction. Sensors. 2023;23(3):1161. doi:10.3390/s23031161.
29. Ahmed M, Sulaiman MH, Hassan MM, Bhuiyan T. Predicting the classification of heart failure patients using optimized machine learning algorithms. IEEE Access. 2025;13(6):30555–69. doi:10.1109/ACCESS.2025.3541069.

30. Trigka M, Dritsas E. Improving cardiovascular disease prediction with deep learning and correlation-aware SMOTE. IEEE Access. 2025;13(4):44590–606. doi:10.1109/ACCESS.2025.3549417.
31. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, et al. Explainable AI for healthcare 5.0: opportunities and challenges. IEEE Access. 2022;10(3):84486–517. doi:10.1109/ACCESS.2022.3197671.
32. Elhoseny M, Mohammed MA, Mostafa SA, Abdulkareem KH, Maashi MS, Garcia-Zapirain B, et al. A new multi-agent feature wrapper machine learning approach for heart disease diagnosis. Comput Mater Contin. 2021;67(1):51–71. doi:10.32604/cmc.2021.012632.
33. Shorewala V. Early detection of coronary heart disease using ensemble techniques. Inform Med Unlocked. 2021;26(6):100655. doi:10.1016/j.imu.2021.100655.
34. Alfadli KM, Almagrabi AO. Feature-limited prediction on the UCI heart disease dataset. Comput Mater Contin. 2023;74(3):5871–83. doi:10.32604/cmc.2023.033603.
35. Baghdadi NA, Farghaly Abdelaliem SM, Malki A, Gad I, Ewis A, Atlam E. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data. 2023;10(1):144. doi:10.1186/s40537-023-00817-1.
36. Alsekait DM, Shdefat AY, Nabil A, Nawaz A, Rana MRR, Ahmed Z, et al. Heart-Net: a multi-modal deep learning approach for diagnosing cardiovascular diseases. Comput Mater Contin. 2024;80(3):3967–90. doi:10.32604/cmc.2024.054591.
37. Başar R, Ocak Ö, Erturk A, de la Roche M. Leveraging machine learning techniques to predict cardiovascular heart disease. Information. 2025;16(8):639. doi:10.3390/info16080639.
38. Sultan SQ, Javaid N, Alrajeh N, Aslam M. Machine learning-based stacking ensemble model for prediction of heart disease with explainable AI and k-fold cross-validation: a symmetric approach. Symmetry. 2025;17(2):185. doi:10.3390/sym17020185.
39. Verdone A, Succetti F, Ceschini A, Rosato A, Fioravanti A, Panella M. A hybrid quantum-neural network for heart disease classification. Biomed Signal Process Control. 2026;113:109185. doi:10.1016/j.bspc.2025.109185.
40. Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, et al. Exercise standards for testing and training: a scientific statement from the American Heart Association. Circulation. 2013;128(8):873–934. doi:10.1161/CIR.0b013e31829b5b44.
41. Rani P, Kumar R, Ahmed NMS, Jain A. A decision support system for heart disease prediction based upon machine learning. J Reliab Intell Environ. 2021;7(3):263–75. doi:10.1007/s40860-021-00133-6.
42. Gaikwad MJ, Asole PS, Bitla LS. Effective study of machine learning algorithms for heart disease prediction. In: Proceedings of the 2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC); 2022 Jan 21–22; Mathura, India. p. 1–6. doi:10.1109/PARC52418.2022.9726613.
43. Aljaaf AJ, Al-Jumeily D, Hussain AJ, Dawson T, Fergus P, Al-Jumaily M. Predicting the likelihood of heart failure with a multi-level risk assessment using decision tree. In: Proceedings of the 2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE); 2015 Apr 29–May 1; Beirut, Lebanon. p. 101–6. doi:10.1109/TAEECE.2015.7113608.
44. Zeng M. The prediction of heart failure based on four machine learning algorithms. Highlights Sci Eng Technol. 2023;39:1377–82. doi:10.54097/hset.v39i.6771.