Statistical Validation of Machine Learning and Optimized MLP Models for Reliable Heart Attack Prediction

Siripurapu Sridhar [1,*], Gudla Sateesh [2], Bonula Ramarao [3], Bokka Sridhar [4], Dasari Nataraj [5], and Eppili Jaya [3]

[1] Department of Electronics and Communication Engineering, Nadimpalli Satynarayana Raju Institute of Technology, Visakhapatnam, India
[2] Department of Computer Science and Engineering, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, India
[3] Department of Electronics and Communication Engineering, Aditya Institute of Technology and Management, Tekkali, India
[4] Department of Electronics and Communication Engineering, Lendi Institute of Engineering & Technology, Vizianagaram, India
[5] Department of Electronics and Communication Engineering, Swarnandhra College of Engineering & Technology, Narsapur, India
Email: Sridhar.vskp@gmail.com (S.S.); sateesh.research@gmail.com (G.S.); drbrrao2015@gmail.com (B.R.); srib105@gmail.com (B.S.); dasarinataraj@gmail.com (D.N.); jaya.baratam@gmail.com (E.J.)
*Corresponding author

***Abstract*—Cardiovascular diseases remain the leading cause of mortality worldwide, emphasizing the need for accurate early prediction. With the increased accessibility to clinical datasets, Machine Learning (ML)-classification algorithms and Deep Learning (DL) approaches (Convolutional Neural Networks) have become integral to early prediction of cardiovascular diseases. This work proposes an effectual hyperparameter optimized heart disease prediction system using a Multilayered Perceptron (MLP) and compares its performance with other Machine Learning (ML) algorithms. The Cleveland, Statlog and Hungarian heart disease datasets sourced from the UC Irvine Machine Learning Repository (UCI ML) were used. Initially, data pre-processing includes correlation analysis between the relevant body values and disease-followed by applying Synthetic Minority Over-sampling Technique (SMOTE) on the training folds only to address the class imbalances without affecting the data integrity. A 10-Fold Cross-Validation approach was used to ensure the robust model evaluation. Hyperparameter optimization using GridsearchCV further enhanced the generalization. Optimized MLP attained the highest accuracy levels of 96.89 followed by the Random Forest (96.07%), Decision Tree (94.99%), K-Nearest Neighbor (KNN) (94.10%), Support Vector Machines (SVM) (93.48%) and Logistic Regression (LR) (93.55%). The Findings obtained illustrate the efficacy of MLP architecture in the prediction of cardio vascular diseases for clinical applications.**

***Keywords*—cardiovascular diseases, machine learning, hyperparameter optimization, random search, grid search, artificial neural networks, k-fold cross validation, Synthetic Minority Over-sampling Technique (SMOTE)**

I. Introduction

The escalating expenditures in the healthcare sector and increasing burden of Cardiovascular Diseases (CVDs) are greatly attributed to lifestyle factors like poor diet, inactivity, and smoking etc. Today CVDs are generating nearly 12 million deaths annually-emphasizing the serious need for early detection and effective prevention. While adopting healthier lifestyles is important, early identification of diseased individuals and access to high-quality diagnostic services are crucial in reducing heart-related issues. Today AI algorithms are under effectual deployment for prediction of cardiovascular diseases to ensure early discovery and timely intrusions. Though there are different technologies to predict heart diseases, their affordability and effectiveness remained as a challenge. ML enables faster and more acceptable diagnoses rather than the conventional CVD tests like blood tests, Electrocardiogram (ECG), echocardiograms and cardiac catheterization etc. However, ML algorithms can manage large, structured datasets to discover hidden patterns and ensure accurate, early identification of heart diseases. This work utilizes the benchmark CVD datasets from UC Irvine Machine Learning Repository (UCI ML) repository (comprised of 13 medical features and 1 target feature) in order to determine the presence of disease, and compare the effectiveness of various ML models in predicting the risk of CVDs.

Manuscript received April 23, 2025; revised December 26, 2025; accepted February 3, 2026; published June 10, 2026.

In literature numerous researchers investigated the application of ML techniques for prediction of CVDs [1, 2]. Earlier works have explored algorithms like Random Forest (RF), Logistic Regression (LR), Naïve Bayes, Support Vector Machines (SVM) and ensemble approaches to forecast heart diseases. Various other works employed K-Nearest Neighbor (KNN), SVM and Multi-linear Regression approaches for diagnosing heart diseases, achieving prominent accuracy levels [3–7]. Hybrid techniques combining Random Forest (RF) and Long-Short Term Memory (LSTM) networks were tested for performance evaluation [8]. Several reviews examined application of Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) architectures for prediction of CVDs [9, 10], additionally eXtreme Gradient Boosting (XGBoost) and Categorical Boosting (CatTBoost) (algorithms) were explored [11, 12], One-Dimensional Variant-CNNs (1D-CNNs) architecture-based analysis with ECG signals and deployment of Morlet type wavelet preprocessing methods were explored [13–15]. Other works investigated CNN frameworks for CVD prediction system using temporal data modeling and Bayesian approach [16, 17]. Hyperparameter optimization strategies are widely explored [18], emphasizing training-time and computational challenges.

Despite vast research efforts, several gaps remain unresolved in literature. Most of the previous studies performed investigations on single datasets only with no statistical validation at all, thereby leaving minimal scope for reproducibility and clinical reliability. Typically, traditional classifiers demonstrate unstable performance on imbalanced datasets due to inadequate focus on precision-recall dynamics and sensitivity-specificity trade-offs.

Although technological advancements have enhanced the performances of ensemble models by leveraging the combined strength of several algorithms, there remains a need for a comprehensive statistically authorized valuation framework to ensure the model generalizability reliability and clinical applicability.

A. Novelty and Motivation

This research work aims to develop a comprehensive evaluation framework using various benchmark CVD datasets through a systematic assessment of clinically relevant metrics. The datasets undergo standard preprocessing to mitigate dataset biases and improve generalizability. Statistical significance analysis are conducted to validate the robustness of observed results, thereby strengthening the clinical applicability of proposed framework.

B. Main Contributions

- Benchmark CVD datasets, Cleveland, Statlog, and Hungarian, are combined to build a fused clinical dataset.
- Methodical analysis of conventional ML models and optimized Multilayered Perceptron (MLP) under similar investigational environment.
- Performance is assessed through clinically relevant metrics, including precision-recall behavior, AUC-ROC analysis, and confusion matrix evaluation.
- Results are validated through statistical significance testing and confidence interval analysis to ascertain robustness and reproducibility.

The remainder of this manuscript is organized as follows. Section II describes the materials and overall workflow methodology. Section III presents experimental evaluation and results. Section IV Concludes the manuscript with key findings and directions for future research.

II. Materials and Methods

A. Datasets

This work utilized publicly available Cleveland (303), Statlog (573) and Hungarian (294) heart disease datasets of UCI-ML repository, which contained around 867 patient instances. The dataset is comprised of 13 clinical attributes and 1 target feature depicting the presence (1) and absence (0) of heart disease, as illustrated in Table I.

TABLE I. Patient Attributes Depicting Heart Diseases

Sl. No	Parameter (Description)	Value
1	Age (in Years)	Continuous
2	Sex (Both Genders)	1 = Male, 0 = Female
3	Cp (Nature of chest pain)	1 = Typical angina, 2 = Atypical angina, 3 = non-anginal pain, 4 = Asymptomatic
4	Trestbps (Blood pressure levels at rest)	Continuous (94–200) in mm hg
5	Chol (Cholesterol in serum)	Continuous (126–564) in mh/dl
6	Fbs (Sugar Levels When Fasting)	1 ≥ 120 mg/dl, 0 ≥ 120 mg/dl
7	Restecg (Electrocardiogram Readings Under Rest)	0 = Normal,1 = Having ST_T wave abnormality, 2 = left ventricular hypertrophy
8	Thalach (Maximum Heart Rate Attained)	Continuous (71–202)
9	Exang (Angina Induced After Exercised)	1 = Yes, 0 = No
10	Oldpeak (Exercise-induced ST depression compared to rest)	Continuous (0–6.2)
11	Slope (Angle of workout segment's peak)	1 = Up sloping, 2 = Flat, 3 = Down sloping
12	Ca (Major Coloured Fluorescence Number)	(1–4)
13	Thal (Scan for thorium)	3 = Normal, 6 = Fixed defect, 7 = reversible defect
14	Target (Cardiac Disease)	1 = Yes, 0 = No

Datasets Used: Cleveland, Statlog, and Hungarian.

B. Work-Flow Methodology

Fig. 1 is comprised of following operations.

(i) Data pre-processing. Ensures data cleaning, missing values imputation and transformation operations followed by feature selection to enhance the ML model's ability and interpret the training dataset effectively.

(ii) Data splitting. Splits the data into training data and test data (80:20 ratio).

(iii) K-Fold Cross-Validation & Synthetic Minority Over-sampling Technique (SMOTE). Augments minority class samples and eliminate the bias in dataset, thereby enabling models to capture patterns associated with the minority class.

(iv) Training ML Models (LR, SVM, KNN, DT, RF) and MLP.

(v) Weighted evaluation of ML models and MLP-Feed Forward Neural Network (FFNN) comparatively.

(vi) Hyperparameter Regularization.

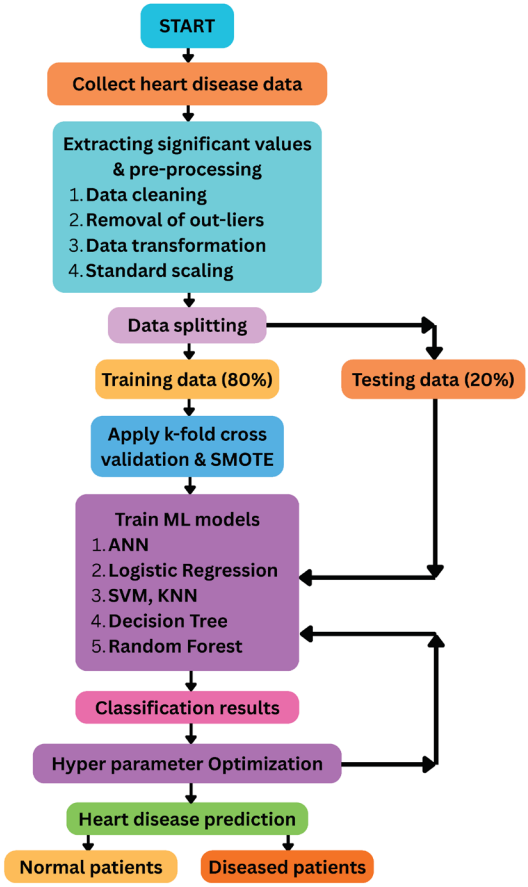

Fig. 1. Work-flow methodology.

1) *Data pre-processing*

Involves data cleaning, outliers' removal and data transformation operations (outlined below).

(i) Cleaning data. Involves correcting noise and inconsistencies, while handling the missing values via deletion or imputation. Deletion removes unprocessable cases but is often deemed unethical in medical datasets due to data loss. Imputation replaces missing values with estimates to preserve dataset integrity, ensuring accurate disease predictions and avoiding misclassifications.

ii) Removal of Outliers. Outliers are irregular data points that impact the model performance [19]. They are commonly identified using boxplots (that display the key values: min, Q1, median Q3, max). While the points outside this range are classified as outliers.

iii) Transformation of data. This process transforms data via aggregation, standardization, normalization and smoothing [20]. While normalization rescales numeric values within the 0–1 range.

2) *Dataset integration and preprocessing strategy*

The Cleveland, Statlog, and Hungarian heart disease datasets share identical clinical features and originate from the UCI Heart Disease repository, making them suitable for integration. Dataset integration was performed to increase sample size, improve class diversity, and enhance the robustness of model training, which is particularly important for data-driven approaches such as artificial neural networks.

To ensure feature compatibility and distributional alignment, only clinical attributes common to all datasets were retained. All datasets were first concatenated, after which global normalization was applied using statistics computed exclusively from the training data to prevent information leakage into the test set. The datasets originate from different institutions and represent independent patient populations, making duplicate patient records unlikely. This was further confirmed through feature-level comparisons, which revealed no identical records.

Although dataset integration may introduce dataset-specific bias due to institutional or demographic variations, such effects were mitigated through uniform preprocessing and stratified cross-validation. Nevertheless, potential impacts on generalization to unseen clinical populations are acknowledged and will be addressed in future work through external validation on independent cohorts.

3) *Feature selection method*

Correlation-based feature selection process was used to determine the most appropriate clinical prognosticators for heart disease classification. The Pearson correlation coefficients were calculated between every input feature and target variable to ascertain the intensity of linear association. A threshold of $|r| > 0.10$ was applied to retain features with meaningful predictive relevance in-line with established medical ML practices. Features with least correlation values were examined for clinical importance to avoid discarding of medically significant variables. Further, multicollinearity was assessed through a correlation heatmap and the highly correlated features ($|r| > 0.80$) were reviewed to mitigate redundancy while preserving complementary diagnostic information [21]. Only the clinical attributes common across the integrated datasets (Cleveland, Statlog, and Hungarian) were preserved, resulting in a feature set that is both statistically relevant and clinical interpretable for robust model training.

4) *K-Fold Cross-Validation (KCV)*

KCV technique [22] is deployed for model building processes to remove the dataset bias (with $k = 10$) to achieve realistic results (Fig. 2). Here the dataset was divided into 10 equal parts, out of which one partition is retained for validation (testing), and remaining 9 partitions are used for training the models.

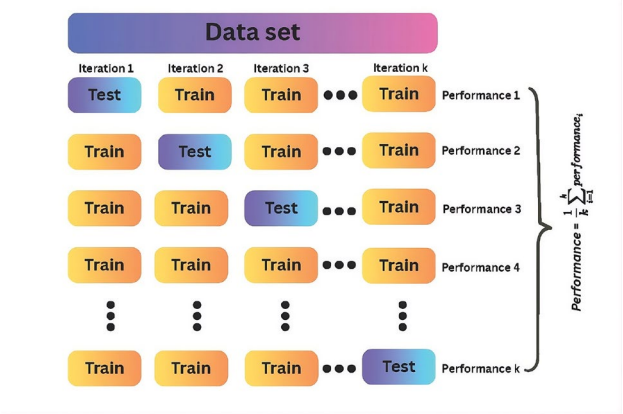

Fig. 2. K-fold cross-validation technique.

In every iteration one unique partition is used for validation while the remaining 9 are meant for training the models. The whole process is iterated 10 times and the results obtained in all iterations are aggregated by a function. The problem associated with overfitting and underfitting are diminished by dataset to meet the performance of training and validation datasets.

5) Synthetic Minority Over-sampling Technique (SMOTE)

Is employed for tackling data imbalances by producing new synthetic samples of minority class. SMOTE creates new samples through linear interpolation of minority class samples by utilizing information from neighboring data points (k = 5) rather than merely duplicating the existing instances [23]. SMOTE generates new synthetic samples that preserve the characteristics of existing instances while introducing variability. This approach augments the diversity of minority class samples, thereby boosting performances of ML models on imbalanced datasets. For a given minority class sample z_i, the k nearest neighbor z_j (in the feature space) of the minority class sample is computed by its Euclidean distance, as shown in Eq. (1).

$$d(z_i, z_j) = \sqrt{\sum_{i=1}^{p} \left(z_{i,l}, z_{j,l}\right)^2} \quad (1)$$

where i represents the first minority class sample of the dataset, j denotes the neighboring minority class sample employed, p depicts the total number of features in the dataset and l demonstrates the feature index ranging between 1 to p.

The newly generated synthetic samples ($Z_{synthetic}$) are generated through Eq. (2).

$$Z_{synthetic} = \delta z_j + (1 - \delta) z_i \quad (2)$$

where delta δ is a random interpolation factor (ranges between 0 and 1) that demonstrates the controlled variability. Here the number of nearest neighbors was set to k = 5, which is an empirically optimal value for medical datasets to strike balance between the sample diversity and boundary smoothness. The oversampling ratio was configured to prevent the biasing of learning towards majority class and ensure a balance of minority and majority class distribution in the training set.

C. Proposed Methods

1) MLP

The Multilayer Perceptron (MLP) employed in this work is a feedforward neural network designed to model nonlinear relationships between clinical features and heart disease outcomes (Fig. 3).

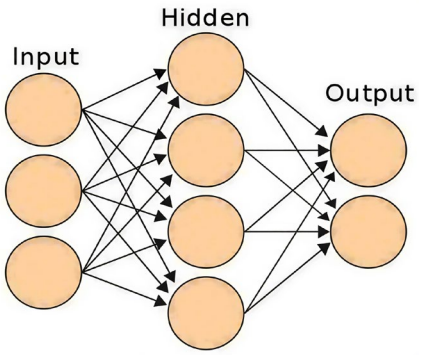

Fig. 3. MLP architecture.

The network contains an input layer, 2 hidden layers, and 1 output layer, where interconnected neurons equipped with activation functions enable the learning of complex patterns in the data. Pre-processed patient attributes are provided to the input layer, and the model is trained using backpropagation algorithm, which iteratively updates the inter-layer weights to minimize prediction error until convergence.

The final architecture and training parameters were determined through hyperparameter optimization using RandomSearchCV. Optimized MLP consisted of 2 hidden layers of 64 and 32 neurons, with Rectified Linear Unit (ReLU) activation function and a sigmoid activation function in output layer for classification. Network is trained via Adam optimizer (learning rate: 0.001) with binary cross-entropy loss function and dropout is applied after every hidden layer to reduce overfitting. Training runs for 200 epochs with batch size equal to 32.

The MLP effectually captures complex clinical patterns and demonstrates strong potential as a clinical decision-support tool for CVD assessment. However, ensuring robustness through appropriate preprocessing, bias mitigation and interpretability is essential for reliable deployment in real-world healthcare settings.

2) Logistic Regression (LR) model

LR delivers probabilities for categorical outcomes (e.g., True/False or 0/1) and its effectiveness largely lies in pre-processing operations: including data cleaning, managing missing values, and selecting relevant features. All the patient attributes were analyzed for correlation with target variable to determine the key predictors.

3) Support Vector Machine (SVM) model

SVM is a nonlinear classification technique used to categorize data, it uses a hyperplane to separate the positive and negative samples. In case of linearly separable

data, SVM identifies support vectors to determine the optimal separating line. Typically, SVM maps input into a wider-dimensional space deploying kernel functions to determine optimal hyperplane. This transformation creates a separate boundary to capture the distance or margin between 2 classes. SVM efficiently classifies heart disease patients from non-patients with an optimal hyperplane.

4) K-Nearest Neighbor (KNN) model

KNN is a nonparametric classification method that groups data based on similarity using distance metrics like Euclidean or Manhattan. It classifies new data points by comparing them to the closest stored instances. KNN doesn't assume any data distribution, as a "lazy learning" algorithm, it doesn't train but stores the entire dataset. While primarily used for classification, KNN also applies to regression and other tasks.

5) Decision Tree (DT) model

DT creates a tree structure, where each node represents an attribute, each branch a decision rule, and each leaf an outcome. Decision trees are applied in healthcare applications.

6) Random Forest (RF) model

RF constructs several decision trees with arbitrary samples of data and input features. However, the final forecast is made via majority voting among the trees. It handles missing values well but may be prone to overfitting, though proper parameter tuning can address this issue. Random Forest mitigates variance, leading to more accurate predictions relevant to single decision trees. While it is ideal for classification tasks and can handle large, incomplete datasets, it is less suited for regression. The algorithm works by generating decision trees from different data subsets and averaging their results. Although slower than a single decision tree, Random Forest improves prediction accuracy and mitigates overfitting by using a higher number of trees.

D. Hyperparameter Tuning

Typically, Hyperparameters influence the model's learning process. While poorly tuned hyperparameters leads to underperformance and inaccurate predictions. Hyperparameter tuning ensures better learning of data (by the models) to avoid overfitting or underfitting and ultimately make more accurate predictions of heart diseases. However, tuning parameters like regularization strength and kernel type can enhance accuracy for adopted ML models. Default hyperparameters values for MLP and ML models are listed in Tables II and III.

TABLE II. DEFAULT HYPERPARAMETER VALUES OF MLP

Classifier	Hyperparameter	Default Values
ANN	No of Hidden Layers	2
	No of Neurons	32, 64, 128
	Activation Function	Hidden: ReLU, Output: Sigmoid
	Optimizer	Adam (Learning Rate = 0.001)
	Epochs	200
	Regularization	Dropout (0.2–0.5)

TABLE III. DEFAULT HYPERPARAMETER VALUES OF ML MODELS

Classifier	Hyperparameter	Definition	Default Value
LR	l2 C Optimizer	Regularization Penalty Regularization Strength lbfgs	**1.0**
SVM	C	Regularization parameter [1–10]	10
	Gamma	It is the kernel coefficient for kernels rbf, poly and sigmoid	rbf
KNN	k Metric	No. of nearest neighbors Distance Metric	5 Euclidean
Decision tree	Max-depth	max depth of tree	none
	criterion	measure the quality of split in a decision tree	gini
	min_samples_split	A node must have at-least 5 samples to be considered for splitting	5
	Random State	shuffles the data for bias	run
Random Forest	n_estimators	No. of trees in forest	100
	max_depth	Max depth of trees	None
	min_samples_leaf	Min samples per leaf node	1
	max_features	Feature subset size	Auto, sq, log2, none
	random_state	Random state	none

E. Performance Evaluation Metrics

Performance evaluation depends on metrics such as Accuracy, Precision, Recall, F1-score and Specificity, as outlined below.

Accuracy: Accuracy is termed as proportion of correct predictions to total number of predictions (Eq. (3)).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

where True Positives (TP) indicate rightly predicted positive instances, True Negatives (TN) indicate rightly predicted negative instances, False Positives (FP) indicate negative instances wrongly classified as positive and False Negatives (FN) indicate the positive instances wrongly classified as negative.

Precision: Precision is termed as proportion of rightly predicted positive cases. It is particularly handy in scenarios where False Positives are of greater concern than False Negatives (Eq. (4)).

$$Precision = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{TP}{TP+FP} \tag{4}$$

Recall (Sensitivity or True Positive Rate): Recall is termed as proportion of rightly identified actual positive cases. It is critical in applications where the cost of false negatives outweighs that of false positives (Eq. (5)).

$$Recall = \frac{CorrectPredictions}{TotalGroundTruth} = \frac{TP}{TP+FN} \tag{5}$$

F1-score: Depicts Harmonic mean among precision and recall values, integrating them into a single value (Eq. (6)).

$$F1-score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \tag{6}$$

Specificity: Specificity determines the model's ability to appropriately compute the actual negative cases (Eq. (7)).

$$Specificity = \frac{Actual\ Negatives}{TotalGroundTruth} = \frac{TN}{TN+FP} \quad (7)$$

III. Experimental Investigations and Results

A. Experimental Setup

All the experiments were carried out in Python using Jupyter Notebook with Scikit-Learn, NumPy, Pandas, and Matplotlib for preprocessing, training, and evaluation. The Cleveland, Statlog, and Hungarian datasets were combined and subjected to various preprocessing operations. The data was divided into training and testing sets using an 80:20 split. Further, KCV (10-fold) and SMOTE is performed to address class level imbalances, Computed accuracy values are illustrated in Table IV. All models (LR, SVM, KNN, DT, RF, and MLP-FFNN) were trained using hyperparameters optimized via GridSearchCV. Performance metrics like Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating Curve (AU-ROC), and Precision–Recall curves are evaluated.

TABLE IV. Performance Values across 3 Different Datasets

Model	Accuracy (%)			
	Dataset			Average Accuracy
	Cleveland	Statlog	Hungarian	
LR	90.50	91.55	92.60	91.55
SVM	90.67	91.68	92.69	91.68
KNN	90.28	92.30	94.32	92.30
DT	93.19	91.17	95.21	93.19
RF	92.86	93.87	94.88	93.87
MLP	96.41	94.38	92.39	94.39

B. Evaluations before Hyperparameter Tuning

Initially experiments were conducted with the default configuration value.

TABLE V. Performance Metrics before Optimization

Model	Precision (%)	Recall (%)	F1-score (%)	AUC
LR	91.20	91.00	91.10	0.94
SVM	91.50	91.30	91.40	0.95
KNN	92.10	91.90	92.00	0.95
DT	92.80	92.50	92.65	0.96
RF	93.60	93.40	93.48	0.96
MLP	94.10	94.00	94.05	0.98

Comparative analysis (Table IV) revealed that models like DT and KNN are highly sensitive to class imbalances and noise in the data, whereas Logistic Regression demonstrated stable behavior but comparatively lower discriminative capability. While Random Forest and MLP outperform all the ML models across all the 3 datasets. MLP achieved competitive performance (followed by Random Forest), however, its predictive accuracy improved substantially after the hyperparameter optimization. These results demonstrate (Table V) that ensemble and neural network-based models, particularly Random Forest and MLP, effectively capture nonlinear relationships and adapt well to structured clinical datasets, thereby confirming their suitability for CVD. It was observed that MLP model outperform the remaining models, demonstrating the ability to learn complex nonlinear relationships. RF model ranked next with a precision of 93.60%, recall of 93.40% and F1-score of 93.48%.

Meanwhile DT model demonstrated a reasonable performance maintaining balanced precision and recall values above 92.5%. Overall, the superior discriminative ability and balanced sensitivity of MLP and RF models confirm their suitability for effectual heart disease prediction and clinical decision-support applications. These values are further optimized to boost the generalization ability of ML models.

C. Hyperparameter Optimization and Training

Common techniques for Hyperparameter optimization [24] include GridsearchCV, RandomSearchCV, and Bayesian optimization.

TABLE VI. Optimized Hyperparameter Values

Model	Hyperparameters Tuned	Optimized Values
LR	Regularization Penalty, Regularization Strength	Penalty = L2, C = 1.2
KNN	No. of neighbors (k), Distance Metric, Weights	K = 7, Metric = Euclidean, Weights = Distance
DT	Max depth, criterion, Min samples split	Max depth = 6, Criterion = gini, Min samples split = 4
RF	n_estimators, max_depth, Max features, Min samples	n_estimators = 300, max_depth = 10, Max features =sqrt, Min samples = 2
MLP	Hidden layers, Neurons, Learning rate, Activation, Epochs = 200	Hidden layers = 2, Neurons = 64–32, Learning rate = 0.001, Activation = ReLU, Alpha (L2) = 0.0001, Epochs = 200

Gridsearch performs an exhaustive search strategy to explore various combinations of hyper parameters and their default values. However, this approach is time-consuming and resource-intensive, particularly when the number of hyper parameters is large. While Random Search Cross-Validation (CV) is more efficient in randomly selecting a set of hyperparameter values for evaluation, rather than exhaustively testing all possible combinations.

Therefore, this approach identifies the optimal hyper parameters (shown in Table VI) more quickly.

D. Performance Evaluations after Hyperparameter Tuning

The performances of each model were evaluated after hyperparameter tuning followed by computation of metrics on the test set and aggregated over cross-validation. Initially learning rate was set to be 0.01 and escalated to 0.001 for the last epoch, counting to a maximum of 200 epochs.

Model performances after hyperparameter tuning (Table VII and Fig. 4) demonstrates a clear enhancement in all evaluation metrics, proving the effectiveness of

optimization in improving predictive reliability. ANN (MLP) model obtained largest accuracy value: 96.89%, precision: 94.90%, recall: 94.50%, and AUC nearing 0.98, indicating excellent sensitivity and strong discriminative capability. While RF model ranked next with accuracy of 96.07%, precision of 94.20%, recall of 93.75%, indicating better robustness and diminished variance due to optimized ensemble parameters.

TABLE VII. PERFORMANCE VALUES BEFORE AND AFTER TUNING

Model	Metric (%)	Before Tuning	After Tuning	Improvement
LR	Accuracy	91.55	93.35	+1.8
	Precision	91.20	91.32	+0.12
	Recall	91.00	91.4	+0.04
	F1-Score	91.10	91.12	+0.02
	AUC	0.94	0.97	+0.03
SVM	Accuracy	91.68	93.48	+1.8
	Precision	91.50	92.50	+1.0
	Recall	91.30	91.41	+0.11
	F1-Score	91.40	91.42	+0.02
	AUC	0.95	0.96	+0.01
KNN	Accuracy	92.30	94.10	+1.8
	Precision	92.10	93.10	+0.1
	Recall	91.90	92.15	0.25
	F1-Score	92.00	92.03	+0.03
	AUC	0.95	0.97	+0.02
DT	Accuracy	93.19	94.99	+1.8
	Precision	92.80	92.95	+0.15
	Recall	92.50	92.60	+0.01
	F1-Score	92.65	92.68	+0.03
	AUC	0.96	0.98	0.02
RF	Accuracy	93.87	96.07	+2.2
	Precision	93.60	94.20	+0.6
	Recall	93.40	93.75	+0.30
	F1-Score	93.48	93.50	+0.02
	AUC	0.96	0.98	+0.02
ANN	Accuracy	94.39	96.89	+2.5
	Precision	94.10	94.90	0.80
	Recall	94.00	94.50	0.5
	F1-Score	94.05	94.07	+0.02
	AUC	0.98	0.98	+0.01

The DT model showcased significance with accuracy ranging 94.99%, exercising control in overfitting (after tuning). The performance updates after hyperparameter optimization portray progressed sensitivity and reliability values, results obtained clarify the appropriateness of MLP and RF models for clinical support decisions.

ROC analysis (Fig. 5) demonstrates large AUC values for MLP and RF models, depicting strong class discrimination. Further, the Precision-Recall (PR) curves showcased MLP model's robustness to imbalanced class distributions, with high precision across the entire recall range than other models.

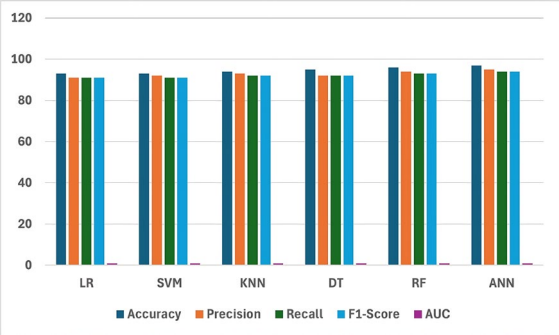

Fig. 4. Performance metrics after hyperparameter tuning.

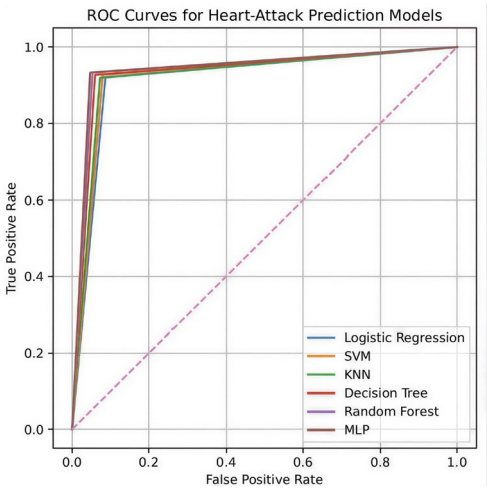

Fig. 5. ROC curves of all models.

E. Statistical Significance Analysis (SSA) and Confidence Intervals (CI)

Statistical Significance Analysis demonstrates that performance variations obtained across the models are statistically appropriate and not due to random nature. Statistical Significance is verified through paired 2-tailed t-test (2T) on the fold-wise results obtained of the 10-fold CV. Here fold-wise results are manifested as mean ± standard deviation, to exhibit accuracy and stability amongst the data-partitions.

A paired statistical testing framework is used with identical cross-validation folds and significance set at $\alpha = 0.05$. The MLP model consistently surpassed the conventional classifiers (LR, SVM, KNN, DT, and RF) by obtaining p-values <0.05 for the key metrics. These low p-values signify stable fold-wise gains with low variances (instead of data-leakage) thereby proving the robustness of MLP model. The mean accuracy improvements range from 0.82% to 3.41% (Table VIII).

TABLE VIII. PAIRED T-TEST RESULTS (10-F CROSS-VALIDATION)

Model Comparison	Accuracy (Mean ± SD) (%)	Mean Difference (%)	p-value	Significance
MLP vs LR	96.89 ± 0.21 vs 93.55 ± 0.22	+3.34	2.1×10^{-13}	Significant
MLP vs SVM	96.89 ± 0.21 vs 93.48 ± 0.23	+3.41	1.4×10^{-13}	Significant
MLP vs KNN	96.89 ± 0.21 vs 94.10 ± 0.22	+2.79	6.3×10^{-12}	Significant
MLP vs DT	96.89 ± 0.21 vs 94.99 ± 0.20	+1.90	8.7×10^{-10}	Significant
MLP vs RF	96.89 ± 0.21 vs 96.07 ± 0.21	+0.82	4.5×10^{-7}	Significant

Confidence Interval (CI) Analysis is employed to assess model reliability. Following optimization, MLP and RF secured accuracies of 96.89% and 96.07% with corresponding 95% CIs of 96.17%–97.61% and 95.35%–96.79%, indicating minimal variability and steady performance across validation folds.

Together SSA and CI analysis confirm that the observed performance gains are statistically robust, reproducible, and not attributable to chance, thereby strengthening the validity of the proposed CVD evaluation framework.

F. Confusion Matrix (CM) Analysis

CM (TP = 140, TN = 143, FP = 7, FN = 10) of MLP depicted a trade-off between precision and recall. The smallest False-Positive (FP) rate lead to large precision illustrating better prediction of CVDs. Simultaneously, the reduced number of False-Negative (FN) count yielded elevated recall (sensitivity) values, thereby demonstrating the strong potential of MLP to correctly determine the heart disease patients. This balanced performance is clinically needful to curtail the number of missed diagnoses and unnecessary clinical alarms. RF closely follows MLP, achieving TP = 140 and TN = 142 with slightly higher FP (8) but an equally low FN (10). Hence, RF maintains recall comparable to MLP, with a marginal reduction in precision, indicating effective disease detection with minimal compromise.

DT exhibits moderate precision and recall (TP = 139, TN = 141, FP = 9, FN = 11). Increase in FNs compared to MLP and RF resulted in reduced sensitivity, though precision remains reasonable due to controlled FPs. SVM and KNN show comparable performance with higher FP and FN counts than ensemble and neural models, giving moderate precision and recall, indicating reduced sensitivity in determining all diseased cases.

LR recorded lowest precision-recall performance, with highest FP (13) and FN (12) rates. Overall, the analysis confirms that MLP obtained best precision-recall trade-off, followed by Random Forest, making them the most amicable models for reliable heart attack prediction.

G. Performance Evaluations before and after Hyperparameter Tuning

Comparative analysis of performance metrics before and after hyperparameter tuning interprets a significant improvement in all the metrics. Before tuning, performances were constrained by the default parameter settings producing moderate accuracy and discriminative capability. However, after hyperparameter tuning, the accuracy improvements of approximately 1.8–2.5% were observed, with the most explicit gains observed in ANN and RF models. In addition to the accuracy, significant improvements in precision and recall values indicate a reduction in both the FP and FN predictions, which is essentially vital for medical diagnosis.

Further, the growth in AUC values (after tuning) confirms increased class separability and diagnostic confidence. This analysis clearly proved that hyperparameter tuning boosts the model's robustness, sensitivity, and clinical reliability, thereby converting the ML classifiers into efficient tools for prediction of heart diseases.

A comparative analysis of proposed MLP and other ML models (Table IX) with other investigative works clearly signified that MLP and RF models achieved accuracy of 96.89% and 96.07% respectively, demonstrating competitive and robust performance relevant to other works.

TABLE IX. COMPARATIVE ANALYSIS OF OTHER WORKS

Ref	Year	Method	Dataset	Accuracy (%)
[16]	2021	CNN	UCI	86
[20]	2020	DT, KNN, RF	A private dataset + public dataset	DT: 80.2, KNN: 90.7, RF: 84.2
[25]	2020	DT	UCI	92.41
[26]	2020	CNN	UCI with 303 instances	94.78
[27]	2022	XGB	UCI	91.80
[28]	2023	ANFIS, M5Tree	UCI with 1028 instances	ANN-LM Accuracy = 96.2
[29]	2023	RF	UCI	98
Our Model		**MLP, RF, DT**	**UCI**	**MLP: 96.89, RF: 96.07, DT: 94.99**

H. Model Interpretability and Clinical Feature Importance Analysis

Interpretability analysis is conducted to ensure transparency and clinical relevance of the model predictions. Using interpretability techniques like correlation analysis, Random Forest (RF) feature importance, permutation importance for MLP, and SHapley Additive exPlanations (SHAP), key predictors of CVD like chest pain type (cp), ST depression (oldpeak), maximum heart rate (thalach), major vessels (ca) and thalassemia (thal) are identified. Typically, correlation analysis demonstrates the association between target variable and key influential predictors of CVD. Similarly, RF importance emphasized the same variables as dominant predictors. Given that neural networks do not provide inherent feature attribution, Permutation importance analysis was applied for the MLP model, reaffirming same key features essential for accurate CVD prediction.

SHAP analysis provides deeper interpretability, identifying the same key features as influential in driving the MLP predictions, with positive SHAP values indicating elevated risk for CVDs. The analysis also revealed clinically meaningful interactions like thal × ca and cp × oldpeak, demonstrating the joint impact of combined risk factors on the model predictions. Overall, the interpretability findings confirm that classical models and optimized MLP rely on clinically justified features, thereby strengthening the model's suitability as a decision-support device [30].

I. Deployment and Computational Efficiency

Practical feasibility of optimized MLP was evaluated on standard hardware-achieving a per-sample inference latency of ≤50 ms on a standard CPU (Intel i5, 16 GB

Random Access Memory (RAM), confirming the real-time capability without the need for any GPU accelerators.

The model can be exported in TensorFlow Lite formats for deployment on mobile, embedded and clinical decision-support platforms. Its minimal computational requirements and fast inference times enable smooth incorporation into medical-workflows and resource-reserved environments.

J. Clinical Workflow Compatibility

The proposed predictive framework is developed to strengthen the clinical workflows by providing robust diagnostic support rather than replacing the physician's expertise. Its outputs—risk scores, class predictions and interpretability visualizations can be integrated into Electronic Medical Record (EMR) systems to improve clinician-led evaluations. Advanced interpretability methods outlined (above) allows clinicians to clearly appraise the influence of critical features like CP, ST depression, Thalach, CA, and Thal, ensuring alignment with clinical reasoning and reinforce the clinician-led decision making. Operationally, the framework facilitates early screening, triage and preventive care planning without compromising clinical expertise.

IV. Discussion and Conclusion with Future Scope

This research conducted a methodical analysis of conventional ML models: LR, SVM, KNN, DT, RF and MLP for prediction of CVDs with help of benchmark datasets. Experimental investigations clearly signify that MLP model obtained best performance among various evaluation metrics-with an accuracy of 96.89% followed by the RF model with an accuracy of 96.07% outscoring other classical models. Statistical significance testing and confidence interval analysis authorized that the performance improvements obtained are real and not attributable to random variation.

Despite these promising results, this work possessed certain limitations. The experiments were conducted using benchmark datasets (from UCI repository), which does not signify the real-world clinical populations data. Furthermore, external validation with standalone hospital datasets was not implemented. In addition, the proposed models utilized only structure tabular features rather than integrating clinically relevant data sources like ECG signals, medical imaging and longitudinal patient records.

Future works will focus on validating the framework using diverse real-world clinical datasets to improve the generalizability. Additionally, exploring advanced Deep Learning (DL) architectures such as CNNs, Deep Neural Networks (DNNs), Long-Short Term Memory (LSTM) networks, Transformer-based models and Graph Neural Networks (GNNs) may further escalate the accuracy of CVD predictions. The development of an integrated mobile-enabled clinical decision-support system could significantly improve the timeliness of interventions and streamline the patient management.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

SS led the conceptualization, methodology, supervision, project administration, writing, review, and editing; GS involved in methodology formal analysis, investigation, validation; BR contributed to software, formal analysis, visualization, data curation, writing, review, and editing; while BS supported investigation, validation, resources, DN is involved in supervision/resources, visualization; and EJ assisted with validation. All authors had approved the final version.

Acknowledgment

The authors gratefully acknowledge the managements of Nadimpalli Satyanarayana Raju Institute of Technology for providing the infrastructural and laboratory support essential for the experimental investigations.

References

[1] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart diseases machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, 23277, 2024.

[2] R. Kumar, S. Garg, R. Kaur *et al.*, "A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations and future directions," *Frontiers in Artificial Intelligence*, vol. 8, 1583459, 2025.

[3] Y. Sandhya, "Prediction of heart diseases using support vector machine," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 8, pp. 126–135, 2020.

[4] M. E. Farooqui and J. Ahmad, "Disease prediction system using support vector machine and multilinear regression," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 8, no. 4, pp. 331–336, 2020.

[5] B. M. Kumar and P. S. Priyadarsini, "Efficient prediction of heart disease using SVM classification algorithm and compare its performance with linear regression in terms of accuracy," *Journal of Pharmaceutical Negative Results*, vol. 3, pp. 1430–1437, 2022.

[6] S. Dange, P. Gaikwad, R. Sheral *et al.*, "Heart disease prediction system using SVM," *International Journal of Innovative Research in Technology (IJIRT)*, vol. 8, no. 12, pp 574–579, 2022.

[7] T. J. H. Mim, M. M. Hassan, B. K. Paul *et al.*, "Machine learning approaches for cardiovascular disease prediction: A comparative study," *Biomedical Materials & Devices*, November 2025. https://doi.org/10.1007/s44174-025-00564-2

[8] Y. Liu, M. Zhang, Z. Fan *et al.*, "Heart disease prediction based on random forest and LSTM," in *Proc. 2020 2nd International Conf. on Information Technology and Computer Application (ITCA)*, 2020, pp. 630–635.

[9] A. H. She, K. Pawan, P. Chaurasia *et al.*, "A review on heart disease prediction using machine learning techniques," *Int. J. Management, IT & Engineering*, vol. 9, no. 4, pp. 208–224, 2019.

[10] C. Zhou, P. Dai, A. Hou *et al.*, "A comprehensive review of deep learning-based models for heart disease prediction," *Artificial Intelligence Review*, vol. 57, 263, 2024.

[11] P. Rahman, A. Rifat, M. I. A. Chy *et al.*, "Machine learning and artificial neural network for predicting heart failure risk," *Computer Systems Science & Engineering*, vol. 44, no. 1, pp. 757–775, 2023.

[12] M. M. R. K. Mamun and A. Alouani, "Diagnosis of STEMI and Non-STEMI heart attack using nature-inspired swarm intelligence and deep learning techniques," *Journal of Biomedical Engineering and Biosciences (JBEB)*, vol. 7, pp. 1–12, 2020

[13] A. Kumar, S. Dhanka, A. Sharma *et al.*, "A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques," *Sci. Rep.*, vol. 15, 25040, 2025.

[14] A. W Sugiyarto, A. M. Abadi, and S. Sumarna, "Classification of heart disease based on PCG signal using CNN," *TELKOMNIKA*, vol. 19, no. 5, pp. 1697–1706, 2021.
[15] M. Fradi, L. Khriji, and M. Machhout, "Real-time arrhythmia heart disease detection system using CNN architecture based various optimizers-networks," *Multimedia Tools and Applications,* vol. 81, pp. 41711–41732, 2022.
[16] A. Mehmood, M. Iqbal, Z. Mehmood *et al.*, "Prediction of heart disease using deep convolutional neural networks," *Arab. J. Sci. Eng.*, vol. 46, pp. 3409–3422, 2021.
[17] G. Atteia, N. A. Samee, E. M. El-Kenawy *et al.*, "CNN-hyperparameter optimization for diabetic maculopathy diagnosis in optical coherence tomography and fundus retinography," *Mathematics*, vol. 10, no. 18,. 3274, 2022.
[18] R. Andonie, "Hyperparameter optimization in learning systems," *J. of Membrane Computing*, vol. 1, pp. 279–291, 2019.
[19] M. D. A. Pranatha, N. Pramaita, M. Sudarma *et al.*, "Filtering outlier data using box whisker plot method for fuzzy time series rainfall forecasting," in *Proc. 2018 4th Int. Conf. on Wireless and Telematics (ICWT)*, 2018, pp. 1–4.
[20] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Comput. Sci.*, vol. 1, 345, 2020.
[21] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, 23277, 2024.
[22] I. K. Nti, O. N. Boateng, and J. Aning, "Performance of machine learning algorithms with different k values in k-fold cross-validation," *I.J. Information Technology and Comp. Sc*., vol. 6, pp. 61–71, 2021.
[23] A. Fernandez, S. Garcia, F. Herrera *et al.*, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. of AI Research*, vol. 61, pp. 863–905, 2018.
[24] J. S. Bergstra, R. Bardenet, Y. Bengio *et al.*, "Algorithms for hyper-parameter optimization," in *Proc. 25th International Conf. on Neural Information Processing Systems*, 2011, pp. 2546–2554.
[25] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, 105400, 2020.
[26] T. K. Sajja and H. K. Kalluri, "A deep learning method for prediction of cardiovascular disease using convolutional neural network," *Revue d'Intelligence Artificielle*, vol. 34, no. 5, pp. 601–606, 2020.
[27] A. Saboor, M. Usman, S. Ali *et al.*, "A method for improving prediction of human heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2022, 1410169, 2022.
[28] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa *et al.*, "Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods," *Biology*, vol. 12, no. 1, 117, 2023.
[29] S. Pandey, "Cardiovascular disease prediction using machine learning," *Buana Inf. Technol. Comput. Sci.*, vol. 4, no. 1, pp. 24–27, 2023.
[30] T. Vu, Y. Kokubo, M. Inoue *et al.*, "Machine learning model for predicting coronary heart disease risk: Development and validation using insights from a Japanese population-based study," *JMIR Cardio*, vol. 9, e68066, 2025.