International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

An Integrated Machine Learning Framework for Cardiovascular Disease Risk Stratification and Early Detection An Integrated ML Framework for CVD Risk Stratification and Early Detection

Md. Shafiqul Islam*, Md. Ruhul Amin, Md. Mushfikur Islam, Khadiza Khatun

Department of Computer Science and Engineering

Dhaka University of Engineering & Technology (DUET), Gazipur

Gazipur-1707, Bangladesh

e-mail: shafiqul@duet.ac.bd, ruhul.cse.duet@gmail.com, mushfikur40@gmail.com, khadijakhatun581@gmail.com

Abstract— Heart and blood vessel disorders—collectively termed cardiovascular disease (CVD)—represent one of the foremost contributors to mortality across the globe. This study presents a predictive framework developed to facilitate early identification of CVD risk, leveraging the 2023 Behavioral Risk Factor Surveillance System (BRFSS) dataset, which comprises roughly 308,854 survey respondents. To enhance data quality, the Interquartile Range (IQR) technique was applied to eliminate extreme observations lying beyond the central 50% of the feature distributions. Class imbalance—where disease-positive samples are substantially outnumbered—was addressed using SMOTE, applied strictly to training partitions to prevent information leakage. The study benchmarked four learning algorithms: K-Nearest Neighbors, Decision Tree, Random Forest, and LightGBM, each subjected to cross-validation-based hyperparameter search. LightGBM emerged as the top performer, recording 95.7% classification accuracy alongside an AUC of 0.958. A soft-voting ensemble yielded 93.9% accuracy with an F1 score of 0.939, surpassing previously reported accuracies of 79%–91%.

Keywords- Cardiovascular Disease (CVD); Ensemble Learning; LightGBM; Early Risk Prediction; IQR Outlier Removal

I. INTRODUCTION

Cardiovascular disease (CVD) encompasses a broad range of conditions affecting cardiac and vascular function, persisting as one of the most pressing challenges in global public health, claiming millions of lives annually despite significant strides in medical treatment. Its continued prevalence is driven by a convergence of modifiable risk factors, including physical inactivity, poor nutritional habits, elevated body mass, tobacco use, and comorbidities such as diabetes and hypertension. According to the World Health Organization (WHO), CVD was responsible for nearly 19.8 million fatalities worldwide in 2022 [1], accounting for approximately 32% of total global deaths. Traditional clinical evaluation methods depend heavily on physician-administered examinations—a process that is both time-intensive and susceptible to human error, particularly in resource-limited healthcare settings. Machine learning (ML) has emerged as a promising computational approach for identifying cardiovascular risk at earlier stages, capitalizing on recent advances in data-driven science. By analyzing clinical and lifestyle variables, ML algorithms can uncover complex non-linear relationships that facilitate accurate forecasting based on indicators such as cholesterol levels, blood pressure, BMI, smoking habits, and physical activity patterns.

Machine learning is a subdivision of both computer science and artificial intelligence (AI) that develops algorithms capable of learning from data and iteratively refining their outputs. Deep learning extends these principles, enabling the modeling of highly intricate data representations. Unlike conventional statistical approaches that rely on confidence intervals and linear assumptions, machine learning uses flexible algorithms to approximate complex functions [2].

II. METHOD

This investigation adopts a supervised machine learning paradigm to estimate the likelihood of heart disease occurrence based on individual health profiles and lifestyle indicators. The overall pipeline encompasses five core stages: data acquisition, preprocessing, exploratory analysis, model development, and performance assessment.

A. Dataset Description

The study utilizes the Heart Disease 2023 BRFSS dataset, sourced from the Behavioral Risk Factor Surveillance System—a large-scale health survey administered by the Centers for Disease Control and Prevention (CDC) in the United States. This publicly accessible dataset, retrieved from the Kaggle repository, contains 308,854 individual records and 22 feature variables, each capturing a distinct demographic, behavioral, or clinical characteristic relevant to heart disease risk stratification [4].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

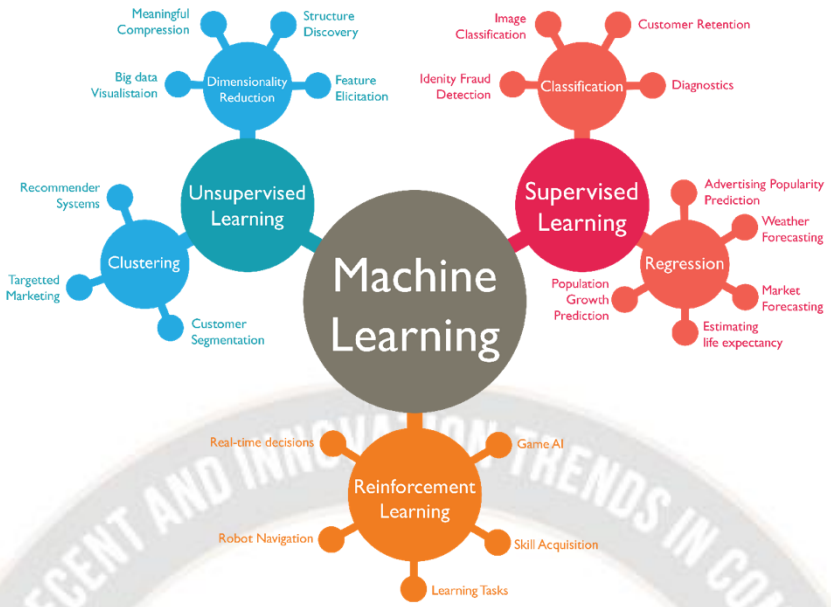

Fig 1 Classification of machine learning [3]

Table 1. Categorization of Features Utilized in This Study

Feature Type	Examples	Description
Categorical (14)	General_Health, Exercise, Checkup, Depression, Heart Disease, Skin Cancer, Other Cancer, Arthritis, Diabetes, Age Category, Sex, Smoking History, High BP, Cholesterol	Qualitative variables represented as grouped categories or binary (Yes/No) responses
Numerical (8)	BMI, Height (cm), Weight (kg), Fruit Consumption, Alcohol Consumption, Green-Vegetables Consumption, Maximum Oxygen Uptake, FriedPotato Consumption	Continuous quantitative variables capturing physiological and behavioral measurements

B. Data Preprocessing

Effective preprocessing is essential to ensure that training data is clean, well-structured, and appropriately formatted for ML algorithms. The key preparation steps are described below.

- *Outlier Detection and Removal*

  Outliers represent observations that deviate markedly from the general data distribution and may stem from measurement errors, recording inconsistencies, or rare events. Their presence can distort distributional assumptions, destabilize model optimization, and reduce predictive reliability. The Interquartile Range (IQR) technique was applied to detect and exclude such anomalous data points, defining outliers as those falling outside the range bounded by the first quartile (Q1) and third quartile (Q3):

  $$IQR = Q3 - Q1 \quad (1)$$

  Any observation below the lower threshold (Q1 - 1.5 × IQR) or above the upper threshold (Q3 + 1.5 × IQR) was excluded before model training [5].

- *Label Encoding*

  To enable ML algorithms to process categorical features, label encoding was employed to convert textual categories into numeric representations. Each distinct category is mapped to an integer value (e.g., Male = 0, Female = 1). In this study, Gender and Smoking Status were numerically encoded; ordinal variables were assigned rank-based values (e.g., Excellent = 5 down to Poor = 1), and binary variables were encoded as 0 (No) or 1 (Yes).

- *Feature Normalization via StandardScaler*

  Numerical feature normalization was performed using StandardScaler to ensure all variables share a common scale with a mean of zero and a standard deviation of one. This prevents features with broader numerical

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: 05 July 2026 Revised: 22 July 2026 Accepted: 20 August 2026

ranges from exerting disproportionate influence during model fitting. The transformation for any feature x is defined as:

$$x_{scaled} = \frac{x - \bar{x}}{s} \quad (2)$$

where $\bar{x}$ represents the feature mean and s denotes its standard deviation. After transformation, values are distributed around zero and generally fall within [-3, 3].

- *Addressing Class Imbalance via SMOTE*

  A pronounced class imbalance was present in the original dataset, with heart disease-positive instances constituting only 8-10% of total samples. This skew caused naive classifiers to predominantly predict the majority class. SMOTE was applied exclusively to the training subset to generate synthetic minority-class instances through feature-space interpolation, thereby restoring distributional balance, limiting data leakage, and improving sensitivity toward positive cases.

C. Classification Algorithms

Four supervised ML classifiers were selected for training and evaluation: Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and LightGBM.

- *K-Nearest Neighbors (KNN)*

  KNN is a non-parametric, instance-based algorithm that classifies an unseen data point by examining the dominant class label among its k closest neighbors in feature space. The primary tuning parameter, n_neighbors, governs the number of proximal samples influencing each decision. In this work, k = 6 was optimal for the unmodified dataset, while k = 2 yielded superior results on the class-balanced version. The Euclidean distance between two samples x_i and x_j is computed as:

$$d(x_i, x_j = \sqrt{\sum_{n=1}^{N} \left(x_{(i,n)} - x_{(j,n)}\right)^2} \quad (3)$$

where N is the total number of features, $x_{(i,n)}$ is the n-th feature of sample i, and $x_{(j,n)}$ is the n-th feature of sample j. The predicted class $\hat{y}$ corresponds to the most frequently occurring class among the k nearest neighbors [6]:

$$\hat{y} = mod(y_i \mid x_i \in k \text{ nearest neighbor of } x)$$

- *Decision Tree Classifier*

  The Decision Tree algorithm recursively partitions the feature space into homogeneous subsets by selecting the most discriminative attribute at each internal node. Though highly interpretable, this model is susceptible to overfitting without proper depth constraints. The primary hyperparameters tuned were max_depth, min_samples_split, and min_samples_leaf. Depth values in the range 15-25 achieved the most favorable trade-off between accuracy and generalization. Splitting decisions are guided by the Gini Index, defined as:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \quad (4)$$

where p_i denotes the proportion of class i instances at node t [7].

- *Random Forest Classifier*

  Random Forest is a bagging-based ensemble method that builds a collection of decision trees, each trained on a randomly sampled subset of training data and features. Final predictions are obtained by aggregating individual tree outputs, reducing variance and mitigating overfitting. Deeper tree configurations (e.g., max_depth = 30) yielded improved results after class balancing. The ensemble prediction is obtained through majority voting across all B trees:

$$\hat{y} = mod\{h_1(x), h_2(x), \ldots, h_B(x)\} \quad (5)$$

where B is the total number of trees in the forest and $\hat{y}$ is the predicted class [8].

- *Light Gradient Boosting Machine (LightGBM)*

  LightGBM is a fast, efficient gradient boosting framework that uses decision-tree learners. It grows trees leaf-wise and performs well with large datasets. Increasing the number of leaves (e.g., to 65) and the tree depth (max_depth = 15) improved results on balanced training data. In each boosting round t, a new tree $f_t(x)$ is added to minimize the objective function:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

where l denotes the loss function, $\hat{y}_i^{(t-1)}$ is the cumulative prediction at the previous iteration, and $\Omega(f_t)$ represents the regularization term penalizing tree complexity [9].

D. Hyperparameter Tuning

Hyperparameter optimization was conducted for all four algorithms using a randomized search strategy over predefined parameter grids, performed separately under four preprocessing configurations: raw dataset, after outlier

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

removal, after class balancing, and combined outlier removal plus SMOTE. Classification accuracy served as the optimization criterion.

Notable patterns observed: KNN converged toward fewer neighbors after balancing; Decision Tree and Random Forest preferred minimal min_samples_split and min_samples_leaf values; and LightGBM required elevated num_leaves post-balancing, reflecting the need for greater model expressiveness.

Table 2. Optimal Hyperparameter Configurations after Tuning

Model	Full Dataset	Outlier Removed	Balanced Data	Balanced + Outlier Removed
KNN	n_neighbors=6	n_neighbors=6	n_neighbors=2	n_neighbors=2
RF	min_samples_split=20, min_samples_leaf=3, max_depth=20	min_samples_split=10, min_samples_leaf=1, max_depth=20	min_samples_split=10, min_samples_leaf=1, max_depth=30	min_samples_split=10, min_samples_leaf=1, max_depth=30
DT	min_samples_split=15, min_samples_leaf=3, max_depth=15	min_samples_split=15, min_samples_leaf=3, max_depth=15	min_samples_split=2, min_samples_leaf=1, max_depth=25	min_samples_split=2, min_samples_leaf=1, max_depth=25
LGB	num_leaves=15, max_depth=10	num_leaves=15, max_depth=15	num_leaves=65, max_depth=15	num_leaves=65, max_depth=15

E. Stratified K-Fold Cross-Validation

To obtain robust performance estimates, stratified K-Fold cross-validation was employed. This strategy preserves the class proportion across each fold, mitigating evaluation bias toward the majority class. The procedure partitions the data into K folds, using each fold in turn as the validation set while training on the remaining K-1 folds. Performance metrics are subsequently averaged across all folds to yield a stable estimate of generalization capability. Final models were trained with identified optimal hyperparameters and evaluated on a held-out test set.

F. Voting Ensemble Classifier

To further bolster predictive robustness, a Voting Classifier ensemble was implemented, integrating forecasts from multiple heterogeneous base learners to construct a more reliable final prediction. Two voting modes exist:

- Hard Voting: Each base classifier casts a single class vote; the majority class is selected as the final prediction.
- Soft Voting: Each classifier contributes class probability estimates, which are averaged to determine the final class label.

This work employed soft voting, combining the outputs of all four classifiers—Decision Tree, Random Forest, KNN, and LightGBM—by averaging their per-class probability estimates [10][11]. This design exploits the complementary strengths of each model: interpretability of Decision Trees, variance reduction of Random Forests, local adaptability of KNN, and sequential optimization power of LightGBM.

G. Performance Metrics

Accuracy quantifies the fraction of all predictions correctly assigned:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (7)$$

Precision measures the proportion of positive predictions corresponding to actual positive instances:

$$\text{Precision} = TP / (TP + FP) \quad (8)$$

Recall captures the proportion of true positive cases the model correctly identifies:

$$\text{Recall} = TP / (TP + FN) \quad (9)$$

In medical applications, recall holds particular importance, as missing a genuine heart disease case may have severe clinical consequences.

F1-Score is the harmonic mean of precision and recall:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (10)$$

ROC-AUC quantifies overall discriminative ability across all decision thresholds. A value approaching 1.0 signifies strong class separability; near 0.5 implies random prediction.

III. Results and Discussion

The proposed ML pipeline delivered strong predictive results, with LightGBM attaining the best individual performance in terms of both accuracy and AUC. The voting ensemble further consolidated these gains, yielding superior overall F1 scores. The findings collectively affirm that careful preprocessing, effective handling of class imbalance, and ensemble combination techniques meaningfully advance cardiovascular disease risk prediction.

A. Baseline Model Performance

As an initial benchmark, each classifier was evaluated on the raw, unprocessed dataset exhibiting structural inconsistencies and severe class imbalance. Performance outcomes are reported in Table 3.

Table 3. Classifier Performance on the Unprocessed Full Dataset

Classifier	Precision	Accuracy	Recall	F1-Score	ROC-AUC
KNN	0.3273	0.9168	0.0302	0.0553	0.5123
RF	0.5753	0.9195	0.0077	0.0152	0.5036
DT	0.2875	0.9087	0.0899	0.1370	0.5352
LGBM	0.5496	0.9198	0.0326	0.0616	0.5151

In the final preprocessing configuration, IQR-based outlier removal and SMOTE were applied jointly, addressing both dataset noise and class distribution skew to maximize model stability and generalizability. As shown in Table 4, this dual preprocessing approach yielded consistent improvements across all classifiers. LightGBM achieved the highest results, attaining 95.16% accuracy and an ROC-AUC of 0.9517, highlighting that comprehensive preprocessing is a decisive factor in healthcare classification tasks.

Table 4. Classifier Performance following Combined Outlier Removal and SMOTE

Classifier	Precision	Accuracy	Recall	F1-Score	ROC-AUC
KNN	0.8487	0.8555	0.8665	0.8575	0.8555
RF	0.9370	0.9256	0.9131	0.9249	0.9257
DT	0.8955	0.8861	0.8751	0.8852	0.8862
LGBM	0.9927	0.9516	0.9226	0.9564	0.9517

B. Voting Ensemble Results

A soft-voting ensemble was constructed by combining the probability outputs of all four classifiers, evaluated across four dataset conditions: raw full dataset, outlier-removed dataset, SMOTE-balanced dataset, and fully preprocessed dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

Table 5 Voting Ensemble Performance across Preprocessing Conditions

Condition	Precision	Accuracy	Recall	F1-Score	ROC-AUC
Full Dataset	0.0000	0.9189	0.0000	0.0000	0.4997
After Outlier Removal	0.5869	0.9242	0.0142	0.0278	0.5067
After SMOTE	0.9289	0.9333	0.9268	0.9328	0.9333
Outlier Removal + SMOTE	0.9404	0.9289	0.9375	0.9390	0.9389

C. Confusion Matrix Analysis

Confusion matrices were constructed for all four classifiers using the fully preprocessed dataset. Each matrix records four outcome categories:

- True Positives (TP): Heart disease cases correctly identified by the model.
- False Positives (FP): Healthy subjects mistakenly classified as having heart disease.
- True Negatives (TN): Non-disease cases correctly classified as healthy.
- False Negatives (FN): Actual heart disease patients incorrectly labeled as healthy.

Table 6. Confusion Matrix Metrics and Error Rates for All Classifiers

Classifier	TP	TN	FP	FN	TPR	FNR
KNN	24000	23000	4300	3700	0.86	0.13
RF	25000	26000	1700	2400	0.91	0.08
DT	24000	25000	2800	3500	0.87	0.13
LGBM	26000	27000	1900	2100	0.92	0.07

As shown in Table 6, LightGBM recorded the highest True Positive Rate (TPR) at 0.92 and the lowest False Negative Rate (FNR) at 0.07, underscoring its exceptional capacity to correctly identify heart disease-positive patients. Random Forest also delivered strong results with a TPR of 0.91 and FNR of 0.08. These outcomes confirm that the two ensemble-based approaches outperform KNN and Decision Tree in this evaluation framework.

D. Accuracy Trends across Preprocessing Stages

Figure 2 displays comparative accuracy trajectories for all four classifiers across successive preprocessing stages, capturing performance variation under progressively refined data preparation. Measurable improvements were consistently observed following outlier exclusion and SMOTE oversampling. LightGBM sustained the highest accuracy across all stages, achieving a peak of approximately 95.6% under the fully preprocessed configuration.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: 05 July 2026 *Revised:* 22 July 2026 *Accepted:* 20 August 2026

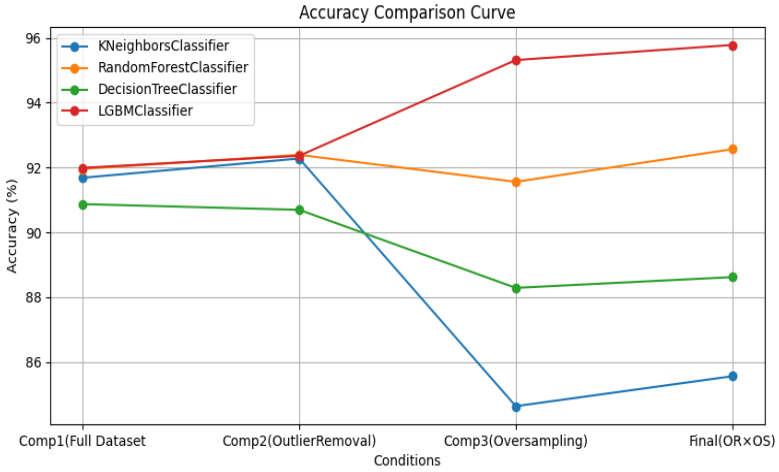

Fig 2 Comparative accuracy trajectories for all classifiers across preprocessing stages

E. Benchmarking against Prior Studies

To contextualize the present findings, this study reviewed previously published works reporting CVD prediction accuracies spanning 79% to 91%. Marcus et al. [12] attained 79.18% accuracy (AUC 0.83) through logistic regression, while Kumar et al. [13] reached 87% with KNN-based approaches. Jaya et al. [14] reported 89% accuracy using a hybrid language model and random forest approach. Krishnan et al. [15] achieved 91% with decision trees. Bhatt et al. [16] obtained 87.28% accuracy (AUC 0.95) using a multilayer perceptron, and Saha et al. [17] also reached 91% (AUC 0.91) via random forest modeling. The optimized pipeline in this study achieved 95.16% accuracy with LightGBM (AUC 0.951) and 93.2% with the voting ensemble (F1 score 0.933), attributable to balanced class distributions, structured outlier removal, and systematic hyperparameter search.

Table 7. Accuracy Comparison between the Proposed Approach and Prior Studies

Author	Technique Used	Best Accuracy
Jaya et al. [14]	Hybrid approach combining language models and Random Forest	89%
Singh et al. [13]	Four supervised ML algorithms	87% (KNN)
Krishnan et al. [15]	Naive Bayes and Decision Tree classifiers	91% (DT)
Bhatt et al. [16]	RF, DT, MLP, XGBoost with k-modes preprocessing	87.28% (MLP), 0.95 AUC
Saha et al. [17]	LR, RF, KNN, DT, XGBoost	91% (RF), 0.91 AUC
This Study	KNN, DT, RF, LightGBM, Voting Ensemble	95.73% (LGBM), 0.9579 AUC; 93.89% (Voting), 0.9390 F1

IV. Conclusion and Future Directions

This study introduced a systematic ML pipeline applied to the BRFSS-2023 dataset [4], integrating fold-aware SMOTE, IQR-based outlier filtering, and stratified k-fold hyperparameter optimization. LightGBM attained the best individual performance at 95.73% accuracy, while the heterogeneous soft-voting ensemble reached 93.89%, both outperforming established baselines. The combination of class imbalance correction and rigorous cross-validation markedly enhanced sensitivity toward minority-class instances without sacrificing overall accuracy. Systematic feature engineering and threshold calibration contributed to deployment-ready robustness, affirming that targeted engineering decisions can substantially improve real-world CVD prediction.

Looking ahead, training on multi-year BRFSS datasets and geographically diverse populations will be pursued to evaluate generalizability. Incorporation of cost-sensitive learning is expected to further refine screening utility. Subsequent efforts will also emphasize improving model interpretability through explainability frameworks and ensuring equitable performance across demographic subgroups. The intended deployment setting is a clinical environment, with planned iterative refinement driven by practitioner feedback.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO, Jul. 31, 2025. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] IBM, "What Is Machine Learning?" IBM, Sep. 22, 2021. https://www.ibm.com/think/topics/machine-learning

[3] J. Smith, J. Doe, R. Brown, and E. White, "Artificial Intelligence in Public Administration: A Comprehensive Review," AI, vol. 6, p. 44, 2025.

[4] Centers for Disease Control and Prevention (CDC), "CDC - 2021 BRFSS Survey Data and Documentation," www.cdc.gov, Aug. 30, 2022.

[5] D. K, S. Nandini, and R. Sanjjushri Varshini, "Comparative Study of ML Algorithms in Detecting Cardiovascular Diseases," arXiv.org, 2024.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Information Theory, vol. 13, no. 1, pp. 21-27, Jan. 1967.

[7] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81-106, Mar. 1986.

[8] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, Oct. 2001.

[9] G. Ke et al., 'LightGBM: A highly efficient gradient boosting decision tree', in Advances in NeurIPS, 2017, vol. 30.

[10] Bane, L. (2025). Detection of false position attacks in VANETs through bagging ensemble learning. PLoS One, 20(8), e0328829.

[11] L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, no. 1-2, pp. 1-39, Nov. 2009.

[12] R. Marcus et al., "Integrated ML Model for Heart Disease Risk Assessment," vol. 11, no. 3, pp. 44-58, Mar. 2023.

[13] R. Kumar et al., "A comprehensive review of ML for heart disease prediction," Frontiers in Artificial Intelligence, vol. 8, May 2025.

[14] T. Jaya, M. Mohan, and M. S. Alam, "Effective Heart Disease Prediction Using ML—Modified KNN," Advances in Intelligent Systems and Computing, pp. 479-489, Sep. 2022.

[15] O. Terrada et al., "Classification and Prediction of atherosclerosis using ML algorithms," ICOA 2019.

[16] C. M. Bhatt et al., "Effective Heart Disease Prediction Using ML Techniques," Algorithms, vol. 16, no. 2, p. 88, Feb. 2023.

[17] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using ML Techniques," SN Computer Science, vol. 1, no. 6, Oct. 2020.