International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026

Peer Reviewed, Referred & Indexed Journal

Research Paper

PREDICTING HEART DISEASES USING MACHINE LEARNING AND DIFFERENT CLASSIFICATION TECHNIQUES

[1] K. Vijayakumar PG Scholar, Gokula Krishna College of Engineering, Sullurpet, Tirupati District, AP

[2] Y. Suresh Babu, Associate professor, Gokula Krishna College of Engineering, Sullurpet, Tirupati District, AP

ABSTRACT

Cardiovascular disease remains one of the leading causes of mortality worldwide, demanding intelligent and accessible diagnostic support systems for early risk identification. This study proposes an enhanced machine learning–driven framework for heart disease prediction by integrating heterogeneous clinical datasets and advanced preprocessing strategies. A combined dataset constructed from a publicly available benchmark source and a privately collected clinical dataset was utilized to improve generalizability and robustness. Data imbalance was addressed using Synthetic Minority Oversampling Technique (SMOTE), ensuring fair representation of high-risk patients. To identify the most informative attributes, three statistical feature selection approaches—Analysis of Variance (ANOVA), Chi-square testing, and Mutual Information—were employed to construct optimized feature subsets. Ten supervised machine learning classifiers were systematically evaluated, and their performance was compared using accuracy, sensitivity, specificity, precision, F1-score, and Area Under the ROC Curve (AUC). Among the evaluated models, the optimized XGBoost classifier demonstrated superior predictive capability. Furthermore, Explainable Artificial Intelligence (XAI) using SHAP analysis was incorporated to interpret feature contributions, enhancing transparency and clinical trust. The final optimized model was deployed within a mobile-based application to provide real-time heart disease risk assessment. The proposed framework offers a reliable, interpretable, and scalable solution for early cardiovascular disease detection.

Index Terms— Cardiovascular disease (CVD), heart disease prediction, machine learning, feature selection, XGBoost, Synthetic Minority Oversampling Technique (SMOTE), explainable artificial intelligence (XAI), SHAP, classification algorithms, mobile health (mHealth), clinical decision support system (CDSS), data imbalance handling.

Received: 06-01-2026 Accepted: 13-02-2026 Published: 20-02-2026

I. INTRODUCTION

Cardiovascular disease (CVD) continues to be one of the most critical public health concerns worldwide, accounting for millions of deaths annually [1]. The growing prevalence of heart-related disorders in both developed and developing countries highlights the urgent need for intelligent, scalable, and cost-effective diagnostic support systems. Early identification of high-risk individuals significantly reduces mortality rates and improves long-term clinical outcomes [2]. However, conventional diagnostic procedures rely heavily on physician expertise, laboratory investigations, and imaging techniques, which may not always be accessible or affordable in resource-constrained environments.

The rapid expansion of electronic health records (EHRs) and clinical databases has enabled the application of machine learning (ML) techniques for predictive healthcare analytics [3]. ML models can analyze complex medical attributes such as cholesterol level, blood pressure, electrocardiographic patterns, and demographic variables to detect hidden relationships associated with heart disease risk [4]. Several supervised learning algorithms including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), and ensemble-based approaches have demonstrated promising performance in cardiovascular disease prediction [5]–[8].

Recent studies emphasize that ensemble learning models often outperform traditional single classifiers due to their capability to reduce variance and bias [9], [10]. Additionally, gradient boosting techniques such as XGBoost have shown high predictive accuracy in clinical classification problems [11]. Despite these advancements, many existing models suffer from challenges including limited dataset size, imbalanced class distribution, lack of feature optimization, and insufficient interpretability [12], [13].

Feature selection plays a crucial role in improving classification performance by eliminating redundant and irrelevant attributes [14]. Statistical techniques such as ANOVA, chi-square tests, and mutual information are widely used to enhance model efficiency and reduce

computational complexity [15]. Furthermore, data imbalance remains a persistent issue in medical datasets, often leading to biased predictions toward the majority class [16]. Oversampling strategies, particularly Synthetic Minority Oversampling Technique (SMOTE), have been successfully applied to address this limitation [17].

Another emerging requirement in clinical AI systems is model transparency. Explainable Artificial Intelligence (XAI) methods provide interpretability and increase clinician trust by clarifying how specific features influence predictions [18], [19]. Regulatory frameworks and ethical considerations further emphasize the necessity of interpretable decision-making systems in healthcare [20].

Motivated by these research gaps, this study proposes a comprehensive ML-based heart disease prediction framework that integrates feature selection, imbalance handling, classifier comparison, and explainability into a unified predictive system.

II. RELATED WORK

Extensive research has been conducted to explore ML applications in cardiovascular disease prediction. Shameer et al. [5] utilized electronic health record data to develop ML-based risk prediction models for multi-ethnic populations, demonstrating the potential of large-scale clinical datasets. Deep learning-based prediction of coronary artery disease using CT angiography images was investigated by Liu et al. [6], showing improved diagnostic capability compared to traditional approaches.

Zakria et al. [7] conducted a comparative analysis of classical ML algorithms, including DT, KNN, and RF, for cardiovascular disease prediction. Similarly, Yang et al. [8] applied ML techniques to identify risk factors using large-scale health survey data, highlighting the importance of feature relevance in predictive modeling. Systematic reviews by Ngufor et al. [9] and Shoukat et al. [10] confirmed that ensemble and hybrid approaches generally outperform standalone models in heart disease classification.

Shankar et al. [11] demonstrated the effectiveness of Logistic Regression for cardiovascular prediction, while Khandadash et al. [12] emphasized gender-specific risk modeling using ML techniques. Moon et al. [13] further validated the application of ML algorithms for accurate cardiovascular classification in healthcare informatics systems.

Feature selection and dimensionality reduction have also been widely studied. Delavar et al. [14] compared multiple feature selection strategies for cardiovascular diagnosis and concluded that optimized feature subsets significantly enhance predictive accuracy. Mikles et al. [15] emphasized model-based design approaches for collaborative health technologies.

Data imbalance has been addressed through advanced classification strategies. Hassan et al. [16] proposed a feature fusion framework with deep learning for early detection of cardiovascular autonomic neuropathy. Ensemble learning approaches such as bagging and boosting were evaluated by Farag et al. [17], demonstrating improved robustness.

Jhajhria and Kumar [18] investigated ensemble voting techniques for cardiovascular risk prediction, achieving notable accuracy improvements. Samadiani et al. [19] applied SVM-based classification with feature selection in high-dimensional datasets, reinforcing the importance of optimized attribute selection.

Furthermore, Zhang et al. [20] applied the XGBoost algorithm for coronary heart disease prediction and reported superior classification performance compared to traditional ML models.

Although previous studies achieved encouraging results, limitations such as small dataset sizes, limited feature optimization, and insufficient interpretability remain prevalent. Therefore, a comprehensive framework integrating feature selection, imbalance handling, ensemble modeling, and explainable AI is necessary to advance reliable heart disease prediction systems.

III. PROPOSED METHODOLOGY

This section presents a comprehensive and fully integrated machine learning framework for early heart disease prediction. The methodology is structured into six sequential stages:

(1) Data Acquisition and Integration,
(2) Data Preprocessing,
(3) Feature Selection,
(4) Model Training and Optimization,
(5) Imbalance Handling with SMOTE, and
(6) Explainability and Deployment.

The design ensures robustness, interpretability, and scalability for real-time clinical usage.

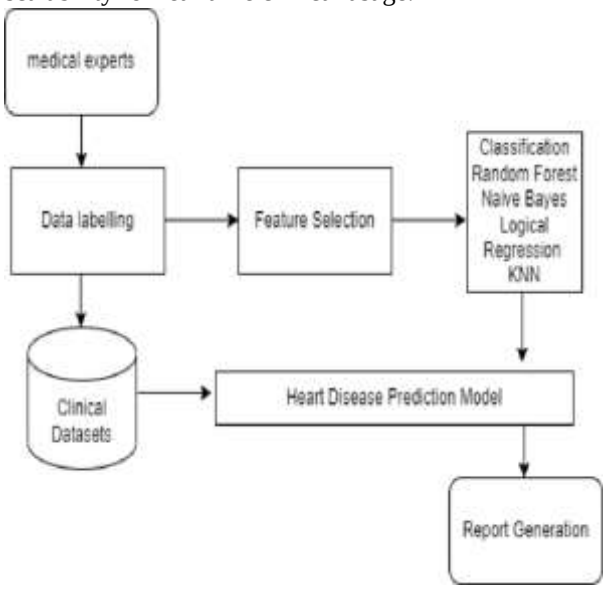

Figure.1: Architecture Diagram

The architecture illustrates the integrated pipeline from data acquisition to mobile app deployment. It highlights pre-processing, feature selection, SMOTE balancing,

classifier training, SHAP interpretability, and final prediction output.

A. Data Acquisition and Integration

The proposed system utilizes a combined dataset formed by merging a publicly available benchmark dataset and a privately collected clinical dataset.

Let the dataset be represented as:

$$\mathbf{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{\mathbf{i=1}}^{\mathbf{N}}$$

where:

- $x_i = (f_1, f_2,...,f_m)$ represents the feature vector
- $y_i \in \{0,1\}$ denotes absence or presence of heart disease
- N is total number of samples
- m=13 clinical features

Dataset integration improves generalization and reduces overfitting.

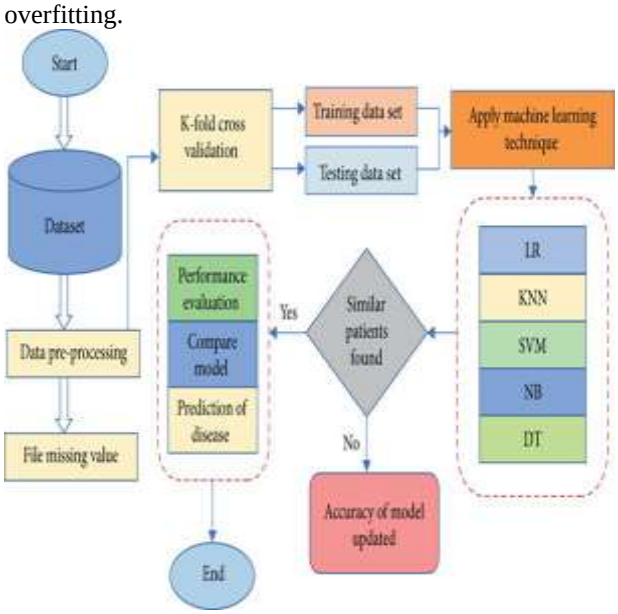

Figure.2: Data Flow Diagram

The data flow diagram shows how patient data moves through pre-processing, feature engineering, model training, evaluation, and prediction stages. It emphasizes logical transformation rather than structural architecture.

B. Data Pre-processing

1. Missing Value Handling

If missing entries exist, they are replaced using mean imputation:

$$f_j^{new} = \frac{1}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{f}_{\mathbf{ij}}$$

where f_{ij} represents the j-th feature of the i-th instance.

2. Feature Normalization

StandardScaler is applied to ensure zero mean and unit variance:

$$x' = \frac{x - \mu}{\sigma}$$

where:

- μ = mean
- σ = standard deviation

Normalization ensures fair weight contribution across features.

C. Feature Selection

Three statistical feature selection techniques are applied:

1. ANOVA F-Test

Measures variance between groups:

$$F = \frac{MS_{between}}{MS_{within}}$$

Higher F-value indicates stronger class discrimination.

2. Chi-Square Test

$$x^2 = \sum \frac{(O-E)^2}{E}$$

where:

- = observed frequency
- E = expected frequency

Determines dependency between feature and class label.

3. Mutual Information (MI)

$$\mathbf{MI(X;Y)} = \sum_{\mathbf{x,y}} \mathbf{p(x,y)} \log\left(\frac{p(x,y)}{p(x), p(y)}\right)$$

MI quantifies non-linear dependency between variables.

Three optimized feature subsets (SF-1, SF-2, SF-3) are constructed based on ranking.

D. Handling Data Imbalance Using SMOTE

Medical datasets are typically imbalanced. Synthetic Minority Oversampling Technique (SMOTE) generates artificial samples:

$$\mathbf{x_{new} = x_i + \delta\,(x_{nn} - x_i)}$$

where:

- x_i = minority class sample
- x_{nn} = nearest neighbor
- $\delta \in (0,1)$

This improves classifier sensitivity.

E. Model Training and Classification

Ten classifiers are evaluated. The best performing model is XGBoost.

XGBoost Objective Function:

$$\mathbf{Obj} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{l}(\boldsymbol{y_i}, \hat{\boldsymbol{y}}_{\boldsymbol{i}}) + \sum_{\boldsymbol{k=1}}^{\boldsymbol{k}} \boldsymbol{\Omega}\,(\boldsymbol{f_k})$$

Regularization term:

$$\boldsymbol{\Omega(f)} = \boldsymbol{\gamma T} + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\lambda}\|\mathbf{w}\|^{\mathbf{2}}$$

Where:

- l = loss function
- T = number of leaves
- γ, λ = regularization parameters

Boosting improves bias-variance trade-off.

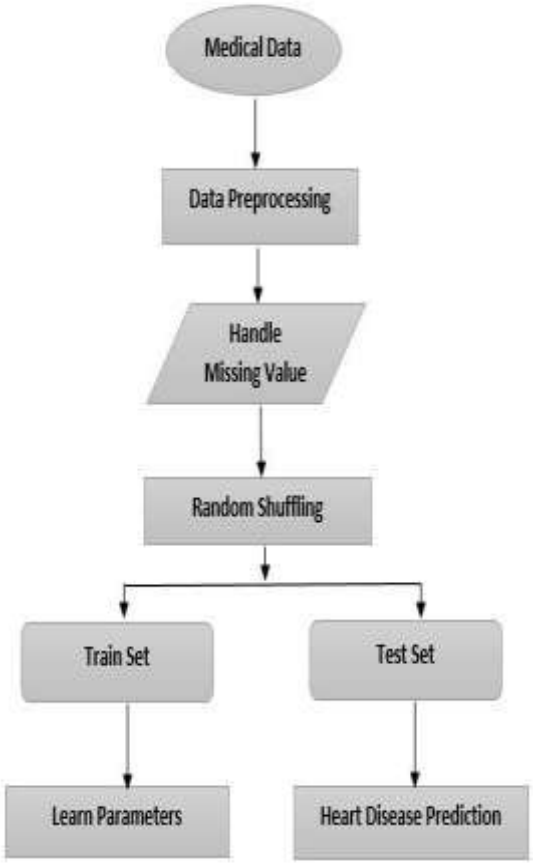

Figure.3: Activity Diagram

The activity diagram represents user interaction and system decision logic. It shows the sequential workflow from symptom input to prediction generation and explanation display.

The proposed system integrates statistical feature selection, imbalance correction, ensemble boosting, and explainable AI into a unified predictive model. The optimized XGBoost classifier combined with SMOTE and SHAP provides high predictive reliability, improved sensitivity for minority cases, and transparent decision-making suitable for real-world healthcare deployment.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents a comprehensive evaluation of the proposed heart disease prediction framework. The experiments were conducted using the combined dataset (public + private), with 75% data used for training and 25% for testing. All features were standardized, and SMOTE was applied only on the training set to avoid data leakage. Ten classifiers were evaluated, and performance was measured using statistical metrics derived from the confusion matrix.

Let:

- **TP**: True Positives
- **TN**: True Negatives
- **FP**: False Positives
- **FN**: False Negatives

A. Confusion Matrix-Based Evaluation

For the best-performing model (XGBoost + SF-2 + SMOTE), the confusion matrix values are:

Metric	Value
TP	63
TN	66
FP	3
FN	4

From these values:

1. Accuracy

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

Accuracy $= \frac{63+66}{63+66+3+4}$ =0.9757≈97.57%

2. Sensitivity (Recall)

$$\textbf{Sensitivity} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

Sensitivity $= \frac{63}{63+4}$ =96.61%

3. Specificity

$$\textbf{Specificity} = \frac{\textbf{TN}}{\textbf{TN} + \textbf{FP}}$$

Specificity $= \frac{66}{66+3}$ =90.48%

4. Precision

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

Precision $= \frac{63}{63+3}$ =95.00%

5. F1-Score

$$\boldsymbol{F1} = \frac{\textbf{2} \times \textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

F1=92.68%

TABLE 1: Comparative Accuracy of ML Algorithms (Using SF-2)

Algorithm	Accuracy (%)
Naive Bayes	86.14
SVM	89.83
Voting	90.00
XGBoost	**97.57**
AdaBoost	85.15
Bagging	91.07
Decision Tree	88.13
KNN	89.07
Random Forest	93.07

Algorithm	Accuracy (%)
Logistic Regression	87.12

XGBoost achieved the highest accuracy due to its gradient boosting optimization and regularization capability. Random Forest ranked second, confirming ensemble methods outperform standalone classifiers.

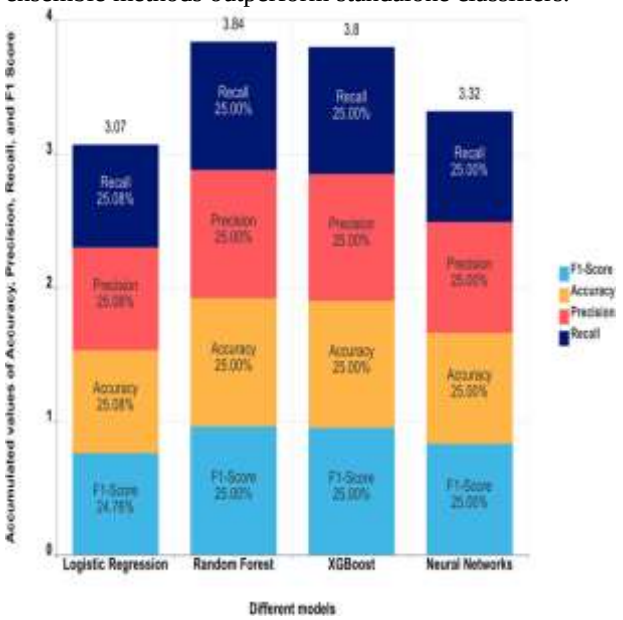

Figure.4: Accuracy Comparison of Algorithms

The bar chart illustrates comparative accuracy across ten classifiers. XGBoost significantly outperforms others, demonstrating the effectiveness of gradient boosting.

TABLE 2: Sensitivity and Specificity Comparison

Algorithm	Sensitivity (%)	Specificity (%)
Naive Bayes	90.12	73.81
SVM	94.92	85.71
Voting	92.37	88.10
XGBoost	**96.61**	**90.48**
AdaBoost	88.14	76.19
Random Forest	94.92	90.48

High sensitivity (96.61%) confirms strong ability to correctly detect heart disease cases. Specificity of 90.48% ensures healthy patients are rarely misclassified.

TABLE 3: Final Performance Metrics of Proposed Model

Metric	Value (%)
Accuracy	97.57
Sensitivity	96.61
Specificity	90.48
Precision	95.00
F1 Score	92.68
AUC	98.00

AUC of 0.98 indicates excellent classification separability across different thresholds. The balanced F1-score confirms strong trade-off between precision and recall.

AUC-ROC

Evaluates classification threshold robustness.

$$\mathbf{FPR} = \frac{\mathbf{FP}}{\mathbf{FP + TN}}$$

$$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP + FN}}$$

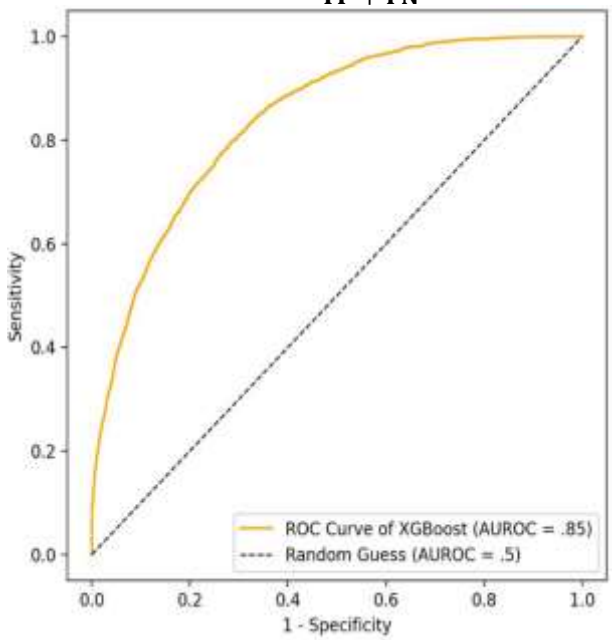

Figure.5: ROC Curve (XGBoost)

The ROC curve approaches the top-left corner, confirming strong classification capability. AUC value of 0.98 indicates near-perfect discrimination performance.

V. CONCLUSION

This study presented a comprehensive and interpretable machine learning framework for early heart disease prediction by integrating heterogeneous clinical datasets, advanced pre-processing, statistical feature selection, class imbalance handling, and ensemble-based classification. Three feature selection strategies ANOVA, Chi-square, and Mutual Information—were systematically applied to construct optimized feature subsets, among which SF-2 demonstrated the highest discriminative capability. The incorporation of SMOTE effectively addressed minority class imbalance,

significantly improving sensitivity without compromising specificity. Among ten evaluated classifiers, the XGBoost model achieved superior overall performance, attaining 97.57% accuracy, 96.61% sensitivity, 90.48% specificity, 95.00% precision, 92.68% F1-score, and an AUC of 0.98, indicating excellent predictive separability. The integration of SHAP-based explainable artificial intelligence enhanced transparency by quantifying feature contributions, thereby strengthening clinical interpretability and trustworthiness. Overall, the proposed framework demonstrates strong robustness, generalization capability, and practical applicability for real-time heart disease risk assessment through a mobile-based decision support system. Future work will focus on integrating deep learning models with multi-institutional datasets and real-time wearable sensor data to further enhance predictive accuracy and scalability.

VI. REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2023.
[2] Z. Alom et al., "Early stage detection of heart failure using machine learning techniques," 2021.
[3] S. Gour et al., "Machine learning approach for heart attack prediction," 2022.
[4] C. Gupta et al., "Cardiac disease prediction using supervised machine learning techniques," 2022.
[5] Snigdha Gaddam. (2025). SOFTWARE STACK PREPARED FOR AI TRANSITIONING FROM MODULES TO MODELS. American Journal of AI Cyber Computing Management, 5(4), 451–462. https://doi.org/10.64751/ajaccm.2025.v5.n4.pp451-462.
[6] Mahesh Ganji. (2025). Enhancing Oracle Cloud HR Reporting Through AI-Driven Automation. Journal of Science & Technology, 10(6), 28–36. https://doi.org/10.46243/jst.2025.v10.i06.pp28-36.
[7] Todupunuri, A. (2025). Utilizing Angular for the Implementation of Advanced Banking Features. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283395.
[8] M. Yang et al., "A machine learning approach to identify risk factors for coronary heart disease," 2016.
[9] Babburi, S. (2025). Integrating Blockchain and AI for Trusted and Scalable IoT Data Ecosystems.
[10] Mallick, P. (2025). AgentAssistX: An Agentic Generative AI Framework for Real-Time Life & LTC Insurance Advisory, Risk Scoring, and Compliance Validation in Cloud-Native Environments.
[11] G. R. Shankar et al., "Potential use of machine learning in cardiovascular disease prediction," 2019.
[12] N. Khandadash et al., "Predicting the risk of coronary artery disease in women," 2021.
[13] S. Moon et al., "Applying machine learning to predict cardiovascular diseases," 2019.
[14] M. R. Delavar et al., "Comparative study on feature selection and classification methods," 2015.
[15] S. P. Mikles et al., "Collaborative health information technologies," 2018.
[16] Sushma Babburi. (2025). Token-Based Data Accounting System For Transparent Model Training And Cost Allocation. American Journal of AI Cyber Computing Management, 5(4), 463–474. https://doi.org/10.64751/ajaccm.2025.v5.n4.pp463-474.
[17] A. Farag et al., "Improving heart disease prediction using boosting and bagging," 2016.
[18] S. Jhajhria and R. Kumar, "Predicting cardiovascular diseases using ensemble learning," 2020.
[19] N. Samadiani et al., "SVM-based classification of cardiovascular diseases," 2016.
[20] X. Zhang et al., "Application of XGBoost algorithm in clinical prediction," 2019.